The Last Word: Books as a Statistical Metaphor for Microbial Communities

Patrick D. Schloss[1] and Jo Handelsman[1,2]

[1]Department of Plant Pathology and [2]Department of Bacteriology, University of Wisconsin, Madison, Wisconsin 53706; email: pschloss@microbio.umass.edu, joh@plantpath.wisc.edu

Annu. Rev. Microbiol. 2007. 61:23–34

First published online as a Review in Advance on April 17, 2007

The *Annual Review of Microbiology* is online at micro.annualreviews.org

This article's doi: 10.1146/annurev.micro.61.011507.151712

0066-4227/07/1013-0023$20.00

Key Words

richness, community structure, census, 16S rRNA genes, Shakespeare

Abstract

Microbial communities contain unparalleled complexity, making them difficult to describe and compare. Characterizing this complexity will contribute to understanding the ecological processes that drive microbe-host interactions, bioremediation, and biogeochemistry. Moreover, an estimate of species richness will provide an indication of the completeness of a community profile. Such estimates are difficult, however, because community structure rarely fits a well-defined distribution. We present a model based on the word usage in books to illustrate the power of statistical tools in describing microbial communities and suggesting biological hypotheses. The model also generates data to test these methods when there are insufficient data in the literature. For example, by simulating the word distribution in books, we can predict the number of words that must be read to estimate the size of the vocabulary used to write the book. Combined with other models that have been used to make inaccessible problems tractable, our book model offers a unique approach to the complex problem of describing microbial diversity.

Contents

INTRODUCTION

Just as the authors of the United States Constitution mandated regular censuses to aid in determining the boundaries of congressional districts and to assess the effects of social policies, microbial ecologists need censuses of the communities they study. Determining the number and comparing the types of bacteria of an environment would greatly aid attempts to assess the effects of environmental perturbations on community composition, diversity, evenness, and richness. Furthermore, an accurate census would estimate the size of the part of the microbial community that is not taken into account by current models of community structure and function. Conducting a census in a microbial community is difficult because of the large population sizes (e.g., up to 10^{12} cells per gram of feces), high richness (e.g., more than 5000 species per gram of soil), and the recalcitrance to the culturing of more than 0.1–10% of these organisms.

The resistance of microorganisms to growing in pure culture has been circumvented over the past 20 years through the development of culture-independent techniques (19, 22, 31), which have shown that the task of characterizing microbial biodiversity is considerably more complex than was previously thought. Microbiologists have relied upon ribosomal RNA (rRNA) genes, because they are found in every cell from *Escherichia coli* to *Homo sapiens*, appear to be less vulnerable to horizontal gene transfer than many other genes, evolve slowly, and contain a sufficient number of base pairs to provide a robust phylogenetic signature. Prior to the use of culture-independent sequencing of 16S rRNA genes, there were 15 known bacterial phyla and now there are well over 50 bacterial phyla, most of which lack any cultured representatives (32, 34). Although bacterial biodiversity can now be described without culturing, defining a bacterial species based entirely on the sequence of one rRNA gene sequence is problematic (41) and has propelled the adoption of arbitrary definitions of bacterial species, operational taxonomic units (OTUs), which represent a collection of 16S rRNA gene sequences that differ from each other by no more than 3% (21). Random sequences of 16S rRNA genes provide a snapshot of the members of a microbial community.

A controversy in the 2000 United States Census arose over whether the Constitution called for an exhaustive sampling of the country's population or permitted the use of sample-based techniques. Although a complete census of the United States is theoretically possible, it is currently impractical to survey every cell in a gram of soil, necessitating that we turn to a sample-based census to estimate the number of bacterial species in soil. The practical question becomes "How many bacteria must be sampled to estimate the number of bacterial species?" The purpose of the Decennial Census was to measure and account for demographic changes. By analogy we might ask how we can compare the census of multiple communities to describe changes in community membership and structure.

Table 1 Summary of word usage from seven books[a]

	On the Origin of Species	*The Descent of Man*	*The Voyage of the Beagle*	*The Adventures of Huckleberry Finn*	*The Adventures of Tom Sawyer*	*The Portrait of a Lady*	*Goodnight Moon*
n_1	The (10,196)	The (22,164)	The (16,924)	And (6228)	The (3681)	The (8449)	Goodnight (20)
n_2	Of (7396)	Of (12,304)	Of (9429)	The (4771)	And (3000)	To (7469)	And (17)
n_3	And (4387)	In (7547)	And (5765)	I (3210)	A (1795)	Of (6592)	A (10)
n_4	In (3920)	And (7342)	A (5325)	To (2914)	To (1705)	A (5485)	The (6)
n_5	To (3567)	To (5743)	In (4288)	A (2911)	Of (1446)	She (4772)	Little (4)
n_6	A (2473)	A (4412)	To (4091)	It (2279)	He (1181)	And (4504)	Of (3)
n_7	That (2059)	That (3529)	Is (2413)	Was (2063)	Was (1166)	Her (4480)	Moon (3)
n_8	Have (1760)	Is (3237)	It (1998)	He (1667)	It (1116)	I (4076)	20 words used twice
n_9	Be (1655)	As (3134)	That (1940)	Of (1641)	In (937)	That (3735)	
n_{10}	As (1579)	Are (2522)	On (1868)	In (1428)	That (890)	You (3735)	
S_T	7426	14,557	12,726	7263	8111	12,427	55
N_T	150,951	272,296	205,424	110,271	70,030	230,485	151

[a]The texts of all books, except *Goodnight Moon*, were obtained from Project Gutenberg (**http://www.gutenberg.org/wiki/Main_Page**) and parsed using a Perl script to count the number of times each word was used, which is shown in parentheses. S_T and N_T designate the total number of different words and the total number of words used in the book, respectively.

FREQUENCY DISTRIBUTIONS OF WORDS AND SPECIES

The seemingly simple question of determining the number of bacterial species in an environment (i.e., species richness) has yet to be answered in a convincing manner in most environments. Depending on how the data are analyzed, DNA-DNA hybridization estimated that 10–30 g of soil contained between 4000 and more than 10,000,000 genome equivalents (14, 29, 30, 38, 39). These values have been widely cited as a measure of species richness, but the conversion between a genome equivalent and species remains controversial (6, 40). Other richness estimates between 5000 and 10,000 species per gram of soil have been obtained using parametric models based on the assumption that the incidence of different species follows either a lognormal (13) or a uniform distribution (24) and that the rarest species has only one member in the community (13). However, there are insufficient data from any soil community to conclude that the species distribution follows a lognormal distribution and no evidence to support a uniform distribution (**Table 1**). To avoid these assumptions, estimates have been made with sample-based nonparametric richness estimators (7, 8, 11). Previous application of the nonparametric estimators to a collection of 16S rRNA gene sequences from a Scottish soil (28) estimated a richness of 590 species. We found, however, that an insufficient number of sequences had been sampled to obtain a reliable estimate using the nonparametric estimators (35). The inaccuracies introduced into the estimate by the assumptions used to conduct the DNA-DNA hybridization, parametric, and nonparametric analyses indicate that the problem of estimating bacterial species richness in soil needs more attention.

Other ecologically significant questions that present their own statistical challenges include comparisons of diversity (i.e., richness and evenness), membership, and structure of multiple communities without performing an exhaustive sampling of each

environment. Three primary problems surround these questions: defining a species or any other taxonomic unit, accessing the bacteria in that sample, and determining how many individuals need to be sampled before a reliable estimate of richness or the species distribution is obtained. We have circumvented the first two problems by assigning sequences to OTUs that are based on the genetic distance between 16S rRNA gene sequences obtained using culture-independent PCR and cloning methods (31, 35). The solution to the third problem, determining the necessary sampling effort, remains elusive.

SO MANY WORDS, SO MANY SPECIES: HOW MANY IS ENOUGH?

The fundamental problem in microbial ecology, estimating the richness in complex communities that cannot be exhaustively sampled, has been encountered in literature (15), linguistics (27), census taking (25), macroecology (9), computer science (11), archaeology (8), transportation (8), and numerous other fields. In microbial ecology, the problem is unique because of the extreme richness, large number of individuals in a community, and unknown frequency distributions of the members.

To circumvent these methodological problems, we sought a dataset of known composition and structure with which to develop statistical models appropriate to microbial communities. We found a number of such datasets in the words of classic books and generated model microbial communities using word usage distributions from these books. These datasets are analogous to individual populations and communities with a well-known richness, evenness, diversity, and membership. By using words and their frequency in books as a substitute for 16S rRNA gene sequences, the analysis is accessible and can be generalized to other fields. Furthermore, these large datasets of words provide surrogates for complex biological datasets that are beyond our current means to depict fully, and analysis of the surrogate datasets will generate hypotheses and direct the design of future experiments in biological systems.

THE BOOK MODEL

Within the book model, each word in a book represents a 16S rRNA gene sequence. Each distinct word that the author used represents a different OTU (perhaps species) in a sequence collection (the species richness). The frequency of each word in the book represents the frequency of OTUs found in a 16S rRNA gene sequence collection (the frequency distribution). The combined frequency and vocabulary of words used in a book therefore represents community structure and can be used to make comparisons among different books or communities.

We are not the first to propose combining literature and statistics. Efron & Thisted (15) used the Shakespearean canon to estimate the number of words in Shakespeare's vocabulary (i.e., the richness of his vocabulary). When a putative Shakespearean poem was found, they reapplied their model to determine whether the addition of the poem fit their model (37). When the poem fit their model, they concluded that the poem was an authentic Shakespearean piece on the basis of the richness of the author's vocabulary. The problem with this approach is analogous to those faced in microbial ecology: The true richness of Shakespeare's vocabulary will never be known because an exhaustive census cannot be completed.

Our approach is to view each book as a distinct community with a known richness, evenness, diversity, membership, and structure. With this approach, we can ask questions such as, "How many words must one read to estimate the overlap in word usage?" or "How many words must be read from two books to know how much of the two books' vocabularies are shared?" Because we know the identity and the number of words used in the book, it is possible to validate any decision rule we

construct to determine when to stop reading or sampling.

DESCRIBING "COMMUNITIES" OF WORDS

To demonstrate our ability to estimate the word usage richness in a book, we selected seven books: Charles Darwin's books *On the Origin of Species*, *The Voyage of the Beagle*, and *The Descent of Man*, Henry James' *The Portrait of a Lady*, Mark Twain's *The Adventures of Tom Sawyer* and *The Adventures of Huckleberry Finn*, and Margaret Wise Brown's *Goodnight Moon* (**Table 1**). These books were selected because they varied in length, richness, and literary style. Distributions such as logseries, lognormal (natural or base 10), broken stick, and uniform (25), which are commonly applied in ecology, did not describe the distribution of word usage in these books (χ^2 goodness of fit test, all $p < 0.001$), except for the word usage in *Goodnight Moon*, which fits all these distributions except the uniform distribution ($p < 0.001$). As the shapes of these distributions differ considerably, we suspect that *Goodnight Moon* is too short to differentiate among the various distributions.

Instead of using a parametric model to describe the shape of the word usage distribution for these books, we constructed empirical distributions based on the number of times the author used each word. For example, *Goodnight Moon* contains 151 words distributed among 55 different words. "Goodnight" was used 20 times, "and" 17 times, "a" 10 times, and so forth (**Figure 1*a***). The author used 28 words only once (i.e., singletons) and 20 words twice in the book (i.e., doubletons) (**Table 1**). To sample the community distribution, we randomly sampled from this distribution so that the probability of sampling the word "goodnight" was 20/151 or 0.132 and each singleton had a probability of 0.007 of being selected. We needed to sample 840 random words in order to achieve 95% confidence of observing all 51 words. Next, we tried using nonparametric richness estimators that make use of the frequency distribution and the number of observed words to estimate the word usage richness for the entire book without an exhaustive sampling of every word in the book. The simplest nonparametric richness estimator, Chao1, predicts the total richness of a community as a function of the observed richness (S_{obs}), the number of OTUs observed once (n_1), and the number of OTUs observed twice (n_2):

$$S_{Chao1} = S_{obs} + \frac{n_1 (n_1 - 1)}{2 (n_2 + 1)}.$$

Using the Chao1 (8), asymptotic coverage-based estimator (ACE) (10), and interpolated Jackknife estimators (7), we needed to sample less than 20 words in order for the 95% confidence interval to include 55; however, further sampling increased the precision of the estimate (**Figure 1*b–d***). For example, sampling 360 words resulted in a Chao1 95% confidence interval between 50 and 60.

DESCRIBING COMMUNITIES OF MICROORGANISMS

By identifying a set of shape parameters and richness that accounted for the sampling pattern observed in large collections of 16S rRNA gene sequences, we sought to engineer a word distribution that could explain the sampling distribution observed in samples from Alaskan or Minnesotan soil microbial communities (36). The overall distribution that we selected was a generic truncated lognormal distribution:

$$N_i = \frac{\frac{1}{S_i \sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln S_i - \mu}{\sigma}\right)^2\right]}{\int_0^{S_T} \frac{1}{S \sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln S - \mu}{\sigma}\right)^2\right] dS},$$

where S_i is the ith OTU and N_i is the relative abundance of individuals in that OTU. The maximum possible value of i is the total number of OTUs in the community, S_T. The shape of the distribution is affected by the values of the normal mean (μ) and standard deviation (σ).

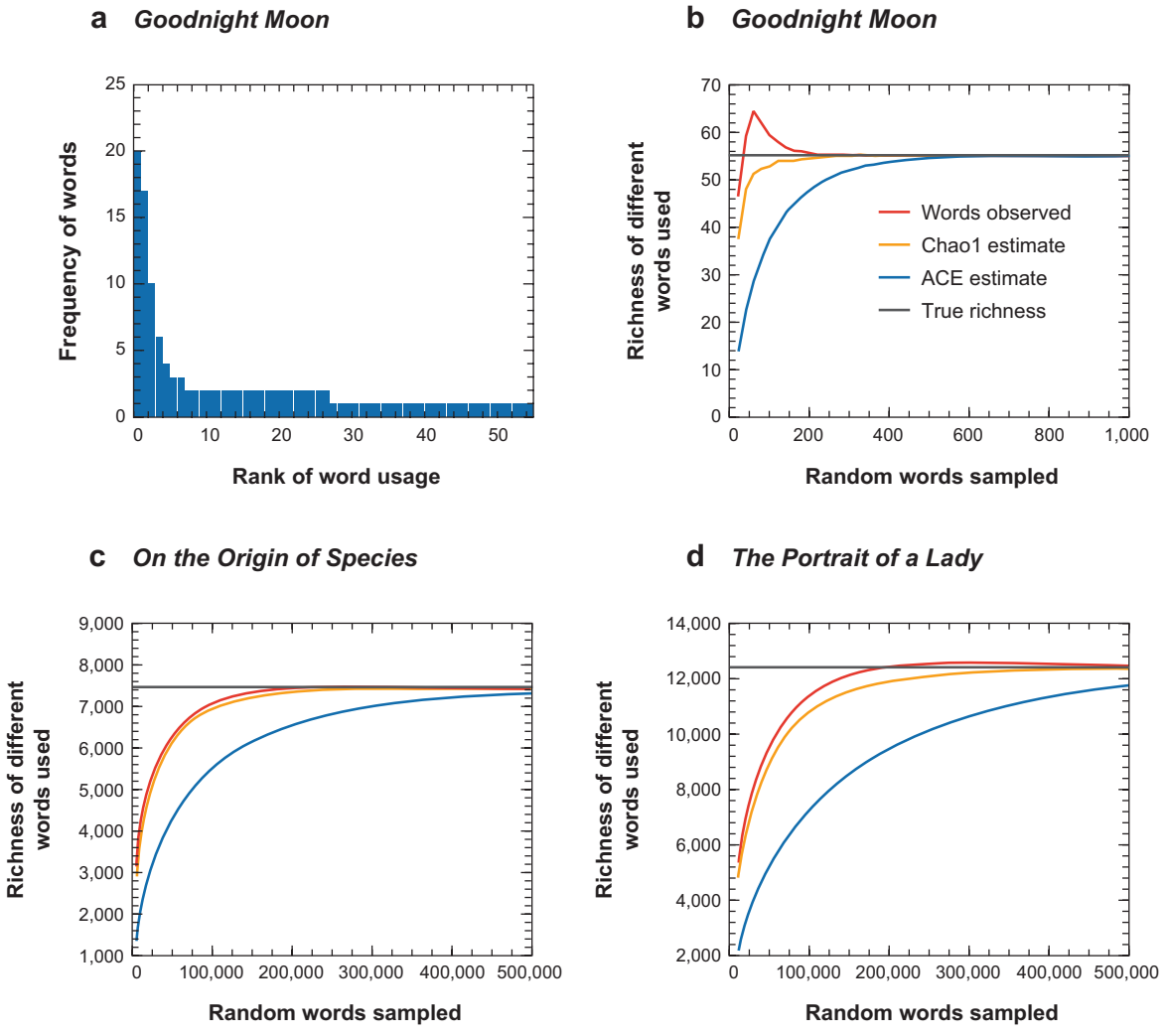

Figure 1

(*a*) Frequency of each word used and (*b*) the observed and estimated number of words used in *Goodnight Moon* as a function of sampling effort (total number of words, $N_T = 151$; richness, $S = 55$). (*c*) The observed and estimated number of words used in *On the Origin of Species* ($N_T = 150{,}951$; $S = 7426$), and (*d*) *The Portrait of a Lady* ($N_T = 230{,}485$; $S = 12{,}427$). The horizontal line indicates the true richness of words used for each book (panels *b*, *c*, and *d*). Abbreviation: ACE, asymptotic coverage-based estimator.

Because the most frequently observed species in the Alaskan collections was observed only 23 times, we were unable to obtain meaningful parameters to fit a truncated lognormal distribution to the Alaskan or Minnesotan 16S rRNA gene sequence collections using methods described elsewhere (25). Instead, we heuristically identified normal mean and standard deviation values for a lognormal distribution that would generate data for the richness estimators and the number of OTUs observed that were comparable to those observed from the Alaskan collection. Although our search was not exhaustive, we determined that a community with a truncated lognormal distribution and a richness

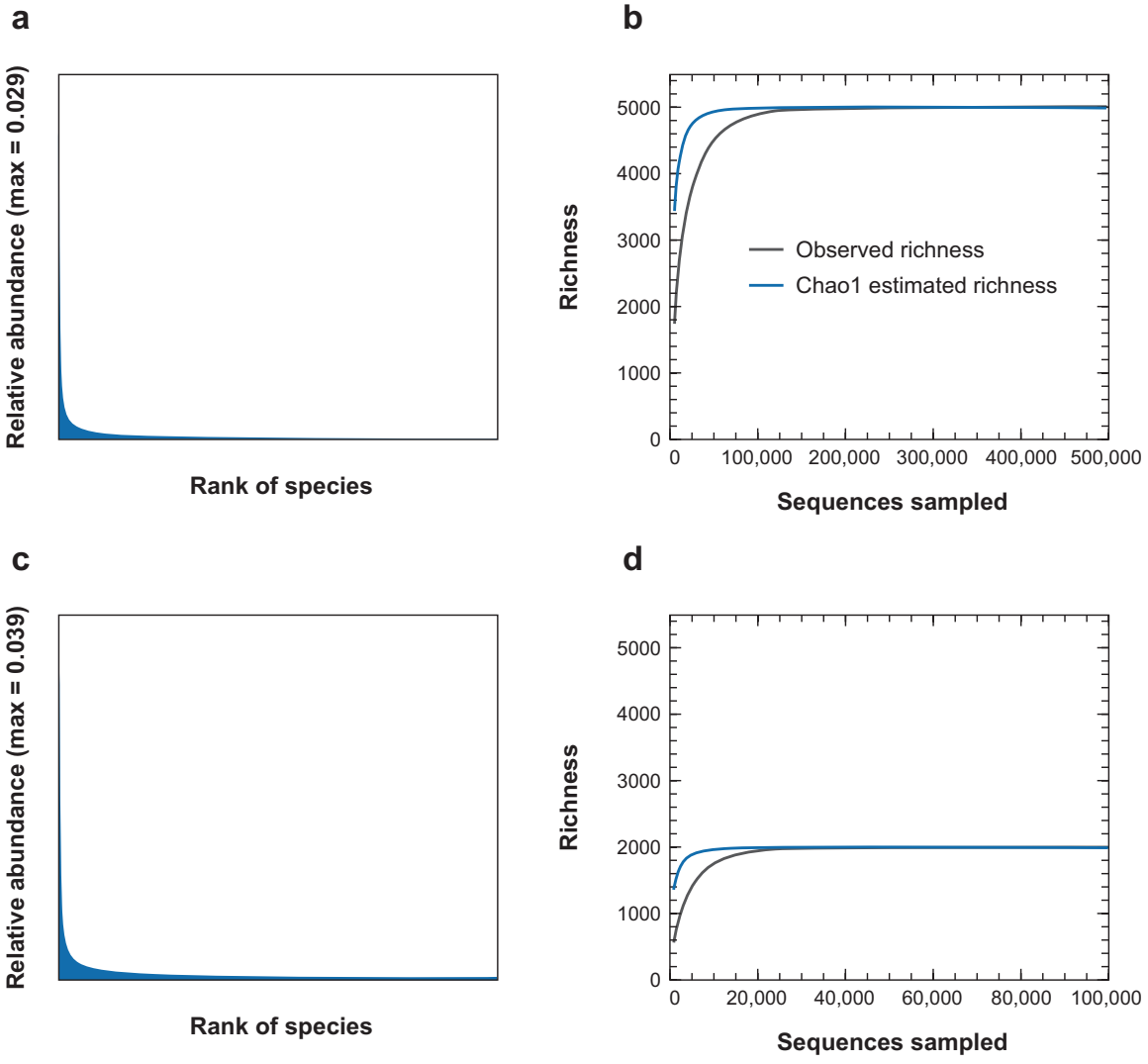

Figure 2

Description of the community structure in (*a*, *b*) Alaskan and (*c*, *d*) Minnesotan soil communities. Relative abundance of each species predicted to reside in the Alaskan and Minnesotan soil communities is given in panels *a* and *c*, respectively. The sampling effort to observe and estimate the true richness of each community is given in panels *b* and *d*, respectively.

of 5000 best explained the data observed so far from the Alaskan collection and that a community with a richness of 2000 could explain the observed data from the Minnesotan collection (**Figure 2**). To sample every member of the Alaskan community twice with 95% confidence would require 380,000 16S rRNA gene sequences and to observe 95% of the richness would require 71,000 16S rRNA genes. To obtain an estimate of the true richness using either the ACE or Chao1 estimator from the Alaskan community would require sampling 18,000 or 39,000 16S rRNA genes, respectively, which represents sampling 65 and 85% of the true estimated richness. Considering that the original sample contained

Figure 3

Unweighted pair group method with arithmetic mean-based dendrogram of the word usage dissimilarity (1-Jaccard Similarity Index) among seven books.

1033 sequences, a significant amount of effort remains. Although we have predicted the richness of these two communities, our estimates are empirical fits of data based on four parameters, and the community was assumed to follow a lognormal distribution. It remains to be seen how well these simulated books resemble microbial reality.

IS DARWIN DIFFERENT FROM TWAIN?

As with the application of the model of Shakespeare's vocabulary (15, 37), others have used statistical analysis to identify the author of 12 of the anonymously written *Federalist Papers*, which both Alexander Hamilton and James Madison claimed to have written, and other apocryphal writings (12, 16, 17, 27, 42–44). The basic premise of these analyses is that a piece of writing is a representative sample of an author's vocabulary and writing style. By comparing the vocabulary used in the apocryphal writing to an author's authenticated writings, it is possible to assign a confidence level to whether the apocryphal writing is that author's. Our strategy was to determine the word usage frequency employed in books with known authorship to assess the similarity of their vocabularies. This is analogous to determining the relatedness of two communities on the basis of the frequency of different OTUs in them and assessing the fraction of OTUs that they share.

We compared the vocabulary used in seven books by estimating the fraction of the vocabulary used in one book and shared in each of the other books. To simplify the discussion of our results, we then calculated the Jaccard Similarity Index, which measures the similarity of vocabularies used in two books. A dendrogram, which graphically presents clusters of books that had the most similar vocabularies, shows the three books written by Charles Darwin clustered together and the two books by Mark Twain clustered together (**Figure 3**). On the basis of this analysis we can propose interesting experiments: If we found a piece of fiction from Charles Darwin, would the vocabulary be more similar to fiction by Twain or nonfiction by Darwin? What about the reverse? Does the vocabulary used by British authors differ significantly from that used by American authors? Do modern evolutionary biologists use the same vocabulary as the field's pioneer? Is there a minimum vocabulary that every book must have to represent a logical "story"?

This final question is most interesting to us. We observed that between 64 and 80% of the 55 different words used to write the book with the lowest richness, *Goodnight Moon*, were found in the other six books. Comparing the other six books to each other, we observed that between 18 and 67% of the words were used in any pair of books. Here we can begin to formulate additional hypotheses: Perhaps there are words, such as "a," "the," "and," "to," or "of," that are essential to any piece of writing. Perhaps, also, there are words that perform a functional role in a story and, although they may not be essential to forming a coherent sentence, are essential to talking about evolution (e.g., "selection"), a boy's adventure (e.g., "Huck"), or putting a child to bed (e.g., "goodnight").

ARE THIN MICE DIFFERENT FROM FAT MICE?

By analogy, we could ask, when we compare microbial communities, Are there core types

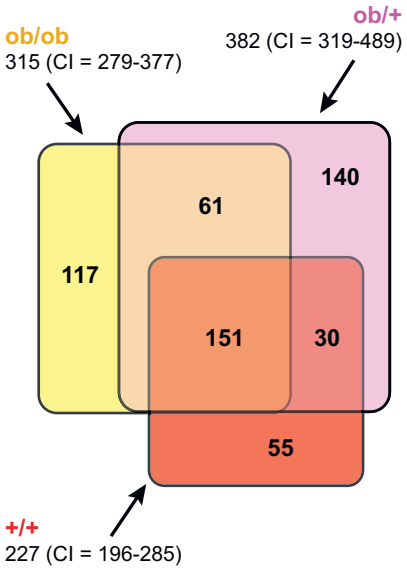

Figure 4

Venn diagram showing the overlap between microbial communities in mice with a +/+ (lean), +/ob (lean), and ob/ob (obese) genotypes. The numbers in the shaded regions indicate the richness of species that are either endemic or shared between genotypes. The overall richness and 95% confidence interval (CI) for each genotype as estimated by the Chao1 estimator are given outside of each region.

of bacteria found in similar environments and are there accessory types of bacteria that perform specialized roles only in certain environments? We reanalyzed a dataset of 16S rRNA genes sequenced from the gut microbial community of mice that varied in their genotype at a locus that affects body type. The mice were either phenotypically lean and homozygous (+/+), phenotypically lean and heterozygous (ob/+), or obese and homozygous (ob/ob) (23). Because it is not feasible to sequence every member of these communities, the dataset is necessarily an incomplete census. We employed another set of nonparametric estimators, which estimate the fraction of a community's membership that is shared with another community when there are unobserved populations. These estimators are analogous to the Chao1 richness estimator for a single community. We found that approximately 151 OTUs, or 37% of all of the OTUs in the three communities, were shared among mice of all three genotypes. Interestingly, no OTUs were predicted to be shared between the ob/ob and +/+ communities. This led to the observation that the communities in the ob/ob and ob/+ mice are more similar to each other than either is to the community in the +/+ mice, although the phenotype of the ob/+ mouse is the same as that of the +/+ mouse (**Figure 4**).

CONCLUSIONS

All models and analogies have both utility and limitations. Using words as a substitute for DNA sequences made the analysis accessible and generalizable to fields outside of microbial ecology. As we have shown, these models help to develop hypotheses that may at first seem opaque when applied to microbial communities but are simple to understand with books. Other conceptual models have been presented to describe complex issues and to assist in the interpretation of data including the use of chain letters to model genome evolution (2), versions of *The Canterbury Tales* to model phylogenetic methods (1), necktie knots to model the outcomes of a random walk (18), origami to model self-organizing systems (26), and a currency-tracking website to model intracontinental human transportation (4). Also, taking tools used in the humanities to compare books and languages and applying them to microbiology and applying the tools of microbiology to the data of the humanities generate a considerable amount of synergy. Finally, these large artificial datasets represent a surrogate for biological datasets that are beyond our current means and their analysis will direct the design of future experiments.

It is important to note that rare species may not be detected until the sequencing effort is complete. For example, although we estimated that 37% of the bacterial species estimated to be in the three mouse communities were shared, it is impossible to identify

those species unless we sequence every one. In our analysis of the simple book, *Goodnight Moon*, whose richness resembles that of the simple microbial community in the gypsy moth midgut (5), the entire book must be read to encounter a word that starts with an "e"—the last sentence of the book is "Goodnight noises *everywhere*." In microbial ecology, is the "last word," or one more bacterial taxon, important? In the Alaskan sequence collection, two sequences belonging to the Sediment-1 candidate phylum (3) were found only after sampling 832 sequences. We suspect that members of many poorly sampled candidate phyla are rare in microbial communities (34) but may play a significant functional role in the microbial community. Sequencing 20,000 or 200,000 16S rRNA genes will probably lead to the discovery of many more new phyla and species.

There are numerous other ways that we could use the book model to describe microbial communities. For instance, our analysis has focused on lexical data (20), such as the raw number of times different words were used. We could also consider content data, which assigns individual words a function, context, and tone (27). Instead of measuring the context of words, we might be interested in understanding the organization of a community at the gene, operon, genome, and metagenome levels. A common analogy for the human genome is a collection of 23 volumes that tell the story of each of us (33). Considering that the number of bacteria that live within and on us exceeds our own human cells by a factor of 10 to 100, perhaps it is time to start thinking about the ways in which the other books on the bookshelf of life affect that 23-volume work.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Murray Clayton for reviewing an earlier version of the manuscript. This work was supported by the USDA (2003-35107-13856), the National Science Foundation, the Howard Hughes Medical Institute, and the University of Wisconsin-Madison College of Agricultural and Life Sciences.

LITERATURE CITED

1. Barbrook AC, Howe CJ, Blake N, Robinson P. 1998. The phylogeny of *The Canterbury Tales*. *Nature* 394:839
2. Bennett CH, Li M, Ma B. 2003. Chain letters and evolutionary histories. *Sci. Am.* 288:76–81
3. Bowman JP, Rea SM, McCammon SA, McMeekin TA. 2000. Diversity and community structure within anoxic sediment from marine salinity meromictic lakes and a coastal meromictic marine basin, Vestfold Hills, Eastern Antarctica. *Environ. Microbiol.* 2:227–37
4. Brockmann D, Hufnagel L, Geisel T. 2006. The scaling laws of human travel. *Nature* 439:462–65
5. Broderick NA, Raffa KF, Goodman RM, Handelsman J. 2004. Census of the bacterial community of the gypsy moth larval midgut by using culturing and culture-independent methods. *Appl. Environ. Microbiol.* 70:293–300

6. Bunge J, Epstein SS, Peterson DG. 2006. Comment on "Computational improvements reveal great bacterial diversity and high metal toxicity in soil". *Science* 313:918
7. Burnham KP, Overton WS. 1979. Robust estimation of population size when capture probabilities vary among animals. *Ecology* 60:927–36
8. Chao A. 1984. Non-parametric estimation of the number of classes in a population. *Scand. J. Stat.* 11:265–70
9. Chao A, Hwang WH, Chen YC, Kuo CY. 2000. Estimating the number of shared species in two communities. *Stat. Sin.* 10:227–46
10. Chao A, Lee SM. 1992. Estimating the number of classes via sample coverage. *J. Am. Stat. Assoc.* 87:210–17
11. Chao A, Ma MC, Yang MCK. 1993. Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika* 80:193–201
12. Collins J, Kaufer D, Vlachos P, Butler B, Ishizaki S. 2004. Detecting collaborations in text: comparing the authors' rhetorical language choices in *The Federalist Papers*. *Comput. Hum.* 38:15–36
13. Curtis TP, Sloan WT, Scannell JW. 2002. Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. USA* 99:10494–99
14. Dykhuizen DE. 1998. Santa Rosalia revisited: Why are there so many species of bacteria? *Antonie Van Leeuwenhoek* 73:25–33
15. Efron B, Thisted R. 1976. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* 63:435–47
16. Elliott WEY, Valenza RJ. 1991. A touchstone for the bard. *Comput. Hum.* 25:199–209
17. Elliott WEY, Valenza RJ. 1996. And then there were none: winnowing the Shakespeare claimants. *Comput. Hum.* 30:191–245
18. Fink TM, Mao Y. 1999. Designing tie knots by random walks. *Nature* 398:31–32
19. Head IM, Saunders JR, Pickup RW. 1998. Microbial evolution, diversity, and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms. *Microb. Ecol.* 35:1–21
20. Holmes DI. 1994. Authorship attribution. *Comput. Hum.* 28:87–106
21. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJM. 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* 67:4399–406
22. Kent AD, Triplett EW. 2002. Microbial communities and their interactions in soil and rhizosphere ecosystems. *Annu. Rev. Microbiol.* 56:211–36
23. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. 2005. Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. USA* 102:11070–75
24. Lunn M, Sloan WT, Curtis TP. 2004. Estimating bacterial diversity from clone libraries with flat rank abundance distributions. *Environ. Microbiol.* 6:1081–85
25. Magurran AE. 2004. *Measuring Biological Diversity*. Malden, MA: Blackwell Pub. 256 pp.
26. Mahadevan L, Rica S. 2005. Self-organized origami. *Science* 307:1740
27. Martindale C, McKenzie D. 1995. On the utility of content analysis in author attribution: *The Federalist*. *Comput. Hum.* 29:259–70
28. McCaig AE, Glover LA, Prosser JI. 1999. Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. *Appl. Environ. Microbiol.* 65:1721–30
29. Ovreas L, Jensen S, Daae FL, Torsvik V. 1998. Microbial community changes in a perturbed agricultural soil investigated by molecular and physiological approaches. *Appl. Environ. Microbiol.* 64:2739–42
30. Ovreas L, Torsvik V. 1998. Microbial diversity and community structure in two different agricultural soil communities. *Microb. Ecol.* 36:303–15

31. Pace NR, Stahl DA, Lane DJ, Olsen GJ. 1985. Analyzing natural microbial populations by rRNA sequences. *ASM News* 51:4–12
32. Rappé MS, Giovannoni SJ. 2003. The uncultured microbial majority. *Annu. Rev. Microbiol.* 57:369–94
33. Ridley M. 1999. *Genome: The Autobiography of a Species in 23 Chapters*. New York: HarperCollins. 344 pp.
34. Schloss PD, Handelsman J. 2004. Status of the microbial census. *Microbiol. Mol. Biol. Rev.* 68:686–91
35. Schloss PD, Handelsman J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* 71:1501–6
36. Schloss PD, Handelsman J. 2006. Toward a census of bacteria in soil. *PLoS Comput. Biol.* 2:e92
37. Thisted R, Efron B. 1987. Did Shakespeare write a newly-discovered poem? *Biometrika* 74:445–55
38. Torsvik V, Goksoyr J, Daae FL. 1990. High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* 56:782–87
39. Torsvik V, Sorheim R, Goksoyr J. 1996. Total bacterial diversity in soil and sediment communities: a review. *J. Indust. Microbiol.* 17:170–78
40. Volkov I, Banavar JR, Maritan A. 2006. Comment on "Computational improvements reveal great bacterial diversity and high metal toxicity in soil". *Science* 313:918
41. Ward DM. 1998. A natural species concept for prokaryotes. *Curr. Opin. Microbiol.* 1:271–77
42. Wilhelm A, Sander M. 1998. Interactive statistical analysis of dialect features. *J. R. Stat. Soc. D* 47:445–55
43. Yang ACC, Hseu SS, Yien HW, Goldberger AL, Peng CK. 2003. Linguistic analysis of the human heartbeat using frequency and rank order statistics. *Phys. Rev. Lett.* 90:108103
44. Yang ACC, Peng CK, Yien HW, Goldberger AL. 2003. Information categorization approach to literary authorship disputes. *Physica A* 329:473–83