Lawrence Berkeley National Laboratory

Lawrence Berkeley National Laboratory

Title

A renaissance for the pioneering 16S rRNA gene

Permalink

https://escholarship.org/uc/item/1j5679dd

Author

Tringe, Susannah

Publication Date

2009-08-17

eScholarship.org Powered by the California Digital Library
University of California

A renaissance for the pioneering 16S rRNA gene

Short title: A 16S rRNA renaissance

Susannah G. Tringe and Philip Hugenholtz

Address: DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, United States

Corresponding authors: Tringe, Susannah G (sgtringe@lbl.gov) and Hugenholtz, Philip (phugenholtz@lbl.gov)

Abstract

Culture-independent molecular surveys using the 16S rRNA gene have become a mainstay for characterizing microbial community structure over the last quarter century. More recently this approach has been overshadowed by metagenomics, which provides a global overview of a community's functional potential rather than just an inventory of its inhabitants. However, the pioneering 16S rRNA gene is making a comeback in its own right thanks to a number of methodological advancements including higher resolution (more sequences), analysis of multiple related samples (e.g. spatial and temporal series) and improved metadata and use of metadata. The standard conclusion that microbial ecosystems are remarkably complex and diverse is now being replaced by detailed insights into microbial ecology and evolution based only on this one historically important marker gene.

Introduction

16S ribosomal RNA (16S for short) holds a special place in the study of microbial evolution and ecology. By virtue of a number of uncommon properties (ubiquity, extreme sequence conservation, and a domain structure with variable evolutionary rates [1]) it spearheaded two revolutions in these fields. First, it radically changed our view of evolution from a five kingdom to three domain paradigm by providing an objective

phylogenetic framework in which to classify cellular life [1], and second, through the cloning and sequencing of 16S genes directly from the environment using conserved broad-specificity PCR primers (16S surveys), it demonstrated that microbial diversity is far more extensive than we ever imagined from culture-based studies [2].

Despite this impressive pedigree, 16S has been overshadowed in recent years by the application of high throughput shotgun sequencing to environmental DNAs (metagenomics) [3-5]. Metagenomic sequencing randomly samples all genes present in a habitat rather than just 16S, thereby providing clues to the functional capacity of a community rather than just its phylogenetic composition. "Classical" community composition profiling by 16S is now often used as a preliminary step prior to metagenomic analysis, and can be of great value in guiding decisions regarding sequencing technology to be used (454 vs. Sanger, shotgun vs. large-insert clones) and amount of sequencing necessary. However, 16S is re-emerging as a stand-alone molecular tool due to a confluence of methodological advancements.

Data generation

16S data is being generated at an unprecedented rate due to new and improved sequencing technologies that dramatically increase throughput and decrease cost. These include lower Sanger sequencing costs as well as inexpensive 454 pyrosequencing and the PhyloChip, a custom microarray for 16S surveys [6]. The flood of 16S data stemming from these advances has in most cases continued to reveal that most diversity estimates, even those based on culture-independent methodologies, fall far short of reality.

Whereas the typical 16S survey by traditional PCR clone sequencing a decade ago might have included a few dozen sequences, many today encompass thousands (e.g. [7-9]. Indeed, there has been a near-exponential increase in the size of the largest surveys (Fig. 1), though these numbers are likely underestimates as many studies only deposit unique phylotypes in the database rather than every clone sequenced. Today's 16S surveys also typically encompass multiple samples, even dozens, rather than targeting a single habitat (Fig. 1)[7,9-11]. Despite the expense of the clone-and-sequence approach, it remains the "gold standard" for identifying novel lineages as only full-length or near full-length sequences are adequate for accurate phylogenetic tree building. Such studies

continue to expand the known “tree of life” at a steady pace, and provide a valuable reference base for the high-throughput technologies discussed below.

The widespread availability of 454 pyrosequencing, a technology roughly an order of magnitude less expensive than Sanger sequencing in terms of cost per base, has changed the landscape of genomics [12]. The first commercially available pyrosequencers generated reads of just 100 bp on average, but could produce 20 Mbp of data in a single run. This first generation of pyrosequencing was termed GS20 and is already an historical footnote. To adapt pyrosequencing technology for 16S analysis, Sogin and colleagues PCR-amplified the short V6 variable region of the bacterial 16S rRNA gene from eight distinct environments using universal primers and ran them separately within a single 454 run [13]. This single run generated a total of ~118,000 sequence tags (“16S pyrotags”), more than any Sanger-based study to date. A follow-up study, also using GS20 technology, generated more than 900,000 bacterial and archaeal 16S pyrotags [14].

Second generation pyrosequencing technology (454-FLX) produces average read lengths of more than 200 bp and yields ~100 Mb per run, and the third generation of pyrosequencing (titanium) has recently appeared on the scene producing ~500 Mb per run and average read lengths >400 bp. These enhancements will continue to improve the throughput and resolution of 16S pyrotag investigations [15]. Barcoding, in which sequences from particular samples can be identified by unique sequences incorporated into the amplification primers, has enabled multiplexing of samples within runs and has further enhanced the usefulness of this approach [16,17].

Another major development in 16S analysis is not directly dependent on DNA sequencing but involves a high-density microarray of phylogenetically specific probes called the PhyloChip [6]. Designing such a microarray is nontrivial due to the highly conserved nature of the 16S rRNA gene; however, DeSantis *et al.* have been able to use such an array to accurately differentiate among phylotypes in diverse environmental samples, documenting not only the vast majority of taxa identified by traditional cloning and sequencing but also groups not seen in clone libraries that were subsequently confirmed by taxon-specific PCR [6]. Advantages of the PhyloChip are low cost and high speed (facilitated by dedicated software, PhyloTrac, to analyze the output) and drawbacks include only being able to identify phylotypes targeted on the chip and an inability to

determine phylotype abundance distribution in the one sample (although individual phylotypes can be tracked quantitatively across samples).

Analytical tools

With the ample data produced by these new technologies has come unprecedented statistical power in discerning similarities and differences among communities. The unidimensional diversity indices and total operational taxonomic unit (OTU) estimates commonly used in single-sample studies have given way to tools designed to directly compare the communities found in different samples. Some of these are aimed primarily at discerning overall phylogenetic similarities, while others assess the structure of the communities as well (e.g. abundance information).

Once sequences have been grouped into OTUs based on some set of similarity criteria (e.g. using DOTUR [18]), similarity indices such as Bray-Curtis can be calculated to estimate the relatedness of different communities. Regression techniques can then be applied to isolate variables that contribute significantly to community composition, as well as correlate the abundances of specific phylogenetic groups with environmental factors [19].

A recent technique more precisely tailored to 16S sequence analysis is UniFrac, a program designed to determine the fraction of unique branch lengths within a phylogenetic tree (comprising sequences from multiple samples) that is attributable to a particular sample [20]. Once this is determined, principal coordinates analysis (PCoA) can be used to identify specific environmental variables that drive differences among communities [21]. One key advantage of this approach is that it circumvents the controversial and often arbitrary process of assigning sequences to OTUs and deals entirely with tree-based metrics. Thus differences at the species or genus level receive less weight than those at the phylum level, but are still considered in the overall analysis. This method has been applied to data from a spectrum of environments and a variety of studies, often leading to new biological insights [20,21]. However, the original implementation of UniFrac takes only unique sequences into account, and thus is insensitive to changes in abundance that may be important to understanding community responses to environmental variation.

A later version of UniFrac, called weighted UniFrac, deals with this weakness by assigning weights to branches of the tree based on the abundance of specific phylotypes. Comparison of the two methods revealed that they measure very different characteristics of the communities, and thus should really be considered complementary approaches rather than different implementations of the same algorithm [22].

A recent rRNA-based study uses a new set of metrics, the phylogenetic species variability (PSV) and phylogenetic species evenness (PSE), to separate out the effects of environmental selection versus interspecies competition [23] (discussed in more detail in the next section). These metrics summarize the relatedness of species within communities such that PSV is equal to one if the members of a community are unrelated and approaches zero if all of the members are closely related. PSE incorporates abundance information in addition to prevalence, such that PSE decreases both when community members are closely related and when members are unevenly represented [23-25]. Permutation tests can then be used to indicate whether species in the communities are underdispersed, such that closely related species tend to co-occur, or overdispersed, such that closely related species tend to occur exclusively from one another. Underdispersion may indicate that environmental filtering is an important force in generating community structure, and supports the use of additional tools to correlate environmental variables with species composition.

Each of these approaches has its strengths and weaknesses, and no one tool can address each individual situation [26]. But the currently available tool kit of experimental and analytical approaches allows a wide variety of experimental hypotheses to be tested.

Case studies

In combination, these experimental and analytical developments are converting 16S surveys from “fishing expeditions” to hypothesis-driven studies. The ability to take multiple samples over time, space or other metrics and deeply interrogate each has enabled an entirely new class of studies, in both the environmental and medical arenas, in which 16S presence and abundance are correlated with specific factors.

A recent novel application of 16S rRNA sequencing to medical diagnosis and treatment investigated the microbial diversity of the lung in intubated patients and the

effects of antibiotic therapy [27]. It found that while the lung remained sterile during brief intubation, patients inevitably became colonized during long-term intubation. Intriguingly, though, patients who were culture-positive for *Pseudomonas aeruginosa*, a primary agent in ventilator associated pneumonia (VAP), were often colonized with a spectrum of other pathogenic and non-pathogenic bacterial species. Paradoxically, when patients were treated with antibiotics targeted to the *Pseudomonas* strains found by culture, the diversity of the flanking populations decreased and *Pseudomonas* became more dominant, potentially as a result of biofilm formation [28]. This finding held true whether the diversity was examined via cloning and sequencing or by PhyloChip, and the decreased diversity correlated with poorer patient outcome in patients with active infections [27]. This study provided an excellent example of the usefulness of PhyloChip analysis in communities dominated by a single member, as far greater diversity was revealed by microarray than could feasibly have been sampled by traditional PCR clone library. The availability of multiple samples from individual patients, taken over the course of therapy, greatly increased the ability to correlate community composition with therapeutic intervention and patient outcome.

Ribosomal RNA sequencing has also been used to study spatial variability in similar environments, both for complete microbial communities and for specific components of those communities. In one study of freshwater lakes, Newton *et al.* [23] investigated the prevalence and abundance of members of the acI lineage of Actinobacteria in 18 different locations. Using the PSV and PSE metrics (described above) they concluded that the 11 distinct acI lineages observed were significantly underdispersed. However, the pattern of dispersion displayed little dependence on distance and rather correlated with environmental variables such as pH, indicating the strong effects of environmental filtering on this set of populations [23].

One field in which the relative roles of evolutionary history and environmental selection have been difficult to sort out is the study of mammalian gut microbiota. A number of studies have revealed strong similarities among the gut communities of diverse mammals, but it has been unclear whether this was the result of similarities among the habitats and the nature of the host-microbe symbiosis, or simply the legacy of descent from a common ancestor whose gut community was already established. Ley *et al.*

recently tackled this question in a study systematically characterizing the fecal communities of 59 distinct mammalian species, from diverse phylogenetic lineages and with widely varying lifestyles, as well as numerous humans [29]. In total, the study encompassed more than 20,000 sequences from 106 samples, including some previously published data. Using UniFrac and PCoA, they found that host phylogeny had a dominant effect on community composition, while diet had a strong secondary role.

Some remaining challenges

The increased statistical power that comes with more data as well as the many tools available to correlate environmental variables with 16S and other molecular data has highlighted the need for accurate, standardized and accessible metadata (i.e. non-sequence data associated with the samples being analyzed such as biogeochemical data). Coordinated efforts are now underway to address this need, such as the Genomics Standards Consortium [30-32].

Almost all of the 16S sequence data in the public repositories to date are the products of PCR amplification of the 16S rRNA gene using 15-25 nucleotide primers broadly targeting bacteria or archaea. However, such primers are known to miss some organisms due to target mismatches (e.g. [33,34]) and the recent application of short (10 nucleotide) "miniprimers" suggests that a considerable amount of diversity may be overlooked in environmental samples using standard 16S primers [35]. Pyrosequencing of cDNAs prepared from environmental RNAs may be the way of the future. This approach not only bypasses any potential primer bias, but simultaneously provides a community profile of all three domains of life and functional information in the form of expressed messenger RNAs [36].

A good quality reference taxonomy based on phylogenetic inference of full-length 16S sequences is required to classify pyrotag and PhyloChip data. Unfortunately, such a reference tree remains somewhat elusive due to the rate of data accumulation (Fig. 1) and difficulties associated with producing and managing trees with hundreds of thousands of taxa. The problem is particularly acute for environmental sequences that are mostly unclassified in the public databases. The issue is currently being addressed through dedicated 16S databases (e.g. [37-39]) and tools developed to handle large sequence

datasets (e.g. [40,41] http://www.microbesonline.org/fasttree/). It is important to note, however, that a number of the new analytical tools can provide biological insights through correlative analyses without the need to classify the underlying 16S data [20,24].

Conclusion

A continuing central role for 16S rRNA in microbial ecology and evolution looks certain thanks to methodological advances in the field.

Acknowledgements

We thank Norman Pace for feedback on the manuscript. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

References and recommended reading

1. Woese CR: **Bacterial evolution**. *Microbiol Rev* 1987, **51**:221-271.
2. Pace NR: **A molecular view of microbial diversity and the biosphere**. *Science* 1997, **276**:734-740.
3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al.: **Environmental genome shotgun sequencing of the Sargasso Sea**. *Science* 2004, **304**:66-74.
4. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment**. *Nature* 2004, **428**:37-43.
5. Tringe SG, Rubin EM: **Metagenomics: DNA sequencing of environmental samples**. *Nat Rev Genet* 2005, **6**:805-814.
6. DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, Andersen GL: **High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment**. *Microb Ecol* 2007, **53**:371-383.
7. Ley RE, Turnbaugh PJ, Klein S, Gordon JI: **Microbial ecology: human gut microbes associated with obesity**. *Nature* 2006, **444**:1022-1023.

8. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA: **Diversity of the Human Intestinal Microbial Flora**. *Science* 2005.

9. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI: **Obesity alters gut microbial ecology**. *Proc Natl Acad Sci U S A* 2005, **102**:11070-11075.

10. Dunn AK, Stabb EV: **Culture-independent characterization of the microbiota of the ant lion Myrmeleon mobilis (Neuroptera: Myrmeleontidae)**. *Appl Environ Microbiol* 2005, **71**:8784-8794.

11. Schauer M, Hahn MW: **Diversity and phylogenetic affiliations of morphologically conspicuous large filamentous bacteria occurring in the pelagic zones of a broad spectrum of freshwater habitats**. *Appl Environ Microbiol* 2005, **71**:1931-1940.

12. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al.: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005.

••13. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ: **Microbial diversity in the deep sea and the underexplored "rare biosphere"**. *Proc Natl Acad Sci U S A* 2006, **103**:12115-12120. This study introduced 16S pyrosequencing, an approach that will likely supercede Sanger-based 16S sequencing in the near term and have long ranging effects on our understanding of microbial diversity.

14. Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML: **Microbial population structures in the deep marine biosphere**. *Science* 2007, **318**:97-100.

15. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R: **Short pyrosequencing reads suffice for accurate microbial community analysis**. *Nucleic Acids Res* 2007, **35**:e120.

•16. Parameswaran P, Jalili R, Tao L, Shokralla S, Gharizadeh B, Ronaghi M, Fire AZ: **A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing**. *Nucleic Acids Res* 2007, **35**:e130. This study demonstrated that it is possible to multiplex hundreds of samples in a single pyrosequencing run.

17. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R: **Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex**. *Nat Methods* 2008, **5**:235-237.

18. Schloss PD, Handelsman J: **Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness**. *Appl Environ Microbiol* 2005, **71**:1501-1506.

19. Brodie EL, DeSantis TZ, Parker JP, Zubietta IX, Piceno YM, Andersen GL: **Urban aerosols harbor diverse and dynamic bacterial populations**. *Proc Natl Acad Sci U S A* 2007, **104**:299-304.

20. Lozupone C, Knight R: **UniFrac: a new phylogenetic method for comparing microbial communities**. *Appl Environ Microbiol* 2005, **71**:8228-8235.

21. Lozupone CA, Knight R: **Global patterns in bacterial diversity**. *Proc Natl Acad Sci U S A* 2007, **104**:11436-11440.

22. Lozupone CA, Hamady M, Kelley ST, Knight R: **Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities**. *Appl Environ Microbiol* 2007, **73**:1576-1585.
23. Newton RJ, Jones SE, Helmus MR, McMahon KD: **Phylogenetic ecology of the freshwater Actinobacteria acI lineage**. *Appl Environ Microbiol* 2007, **73**:7169-7176.
•24. Helmus MR, Savage K, Diebel MW, Maxted JT, Ives AR: **Separating the determinants of phylogenetic community structure**. *Ecol Lett* 2007, **10**:917-925. A paper describing useful metrics for assessing the relative influences of environmental selection and interspecies competition.
25. Helmus MR, Bland TJ, Williams CK, Ives AR: **Phylogenetic Measures of Biodiversity**. *Am Nat* 2007, **169**. A paper describing useful metrics for assessing the relative influences of environmental selection and interspecies competition.
26. Schloss PD: **Evaluating different approaches that test whether microbial communities have the same structure**. *Isme J* 2008, **2**:265-275.
27. Flanagan JL, Brodie EL, Weng L, Lynch SV, Garcia O, Brown R, Hugenholtz P, DeSantis TZ, Andersen GL, Wiener-Kronish JP, et al.: **Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with Pseudomonas aeruginosa**. *J Clin Microbiol* 2007, **45**:1954-1962.
28. Tart AH, Wozniak DJ: **Shifting paradigms in Pseudomonas aeruginosa biofilm research**. *Curr Top Microbiol Immunol* 2008, **322**:193-206.
•29. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, et al.: **Evolution of Mammals and Their Gut Microbes**. *Science* 2008. One of the largest 16S studies to date, this manuscript addresses the relative contributions of multiple factors to gut microbial community membership and structure.
30. Garrity GM, Field D, Kyrpides N, Hirschman L, Sansone SA, Angiuoli S, Cole JR, Glockner FO, Kolker E, Kowalchuk G, et al.: **Toward a standards-compliant genomic and metagenomic publication record**. *Omics* 2008, **12**:157-160.
31. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, et al.: **The minimum information about a genome sequence (MIGS) specification**. *Nat Biotechnol* 2008, **26**:541-547.
•32. Hirschman L, Clark C, Cohen KB, Mardis S, Luciano J, Kottmann R, Cole J, Markowitz V, Kyrpides N, Field D: **Habitat-Lite: A GSC Case Study Based on Free Text Terms for Environmental Metadata**. *Omics* 2008. This paper discusses the important issue of properly collecting and indexing environmental metadata in 16S and metagenomic studies.
33. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO: **A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont**. *Nature* 2002, **417**:63-67.
34. Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE, Banfield JF: **Lineages of acidophilic archaea revealed by community genomic analysis**. *Science* 2006, **314**:1933-1935.
35. Isenbarger TA, Finney M, Rios-Velazquez C, Handelsman J, Ruvkun G: **Miniprimer PCR, a new lens for viewing the microbial world**. *Appl Environ Microbiol* 2008, **74**:840-849.

••36. Urich T, Lanzen A, Qi J, Huson DH, Schleper C, Schuster SC: **Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome**. *PLoS ONE* 2008, **3**:e2527. This study shows the potential for RNA pyrosequencing to circumvent PCR for unbiased community profiling.

37. Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM: **The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data**. *Nucleic Acids Res* 2007, **35**:D169-172.

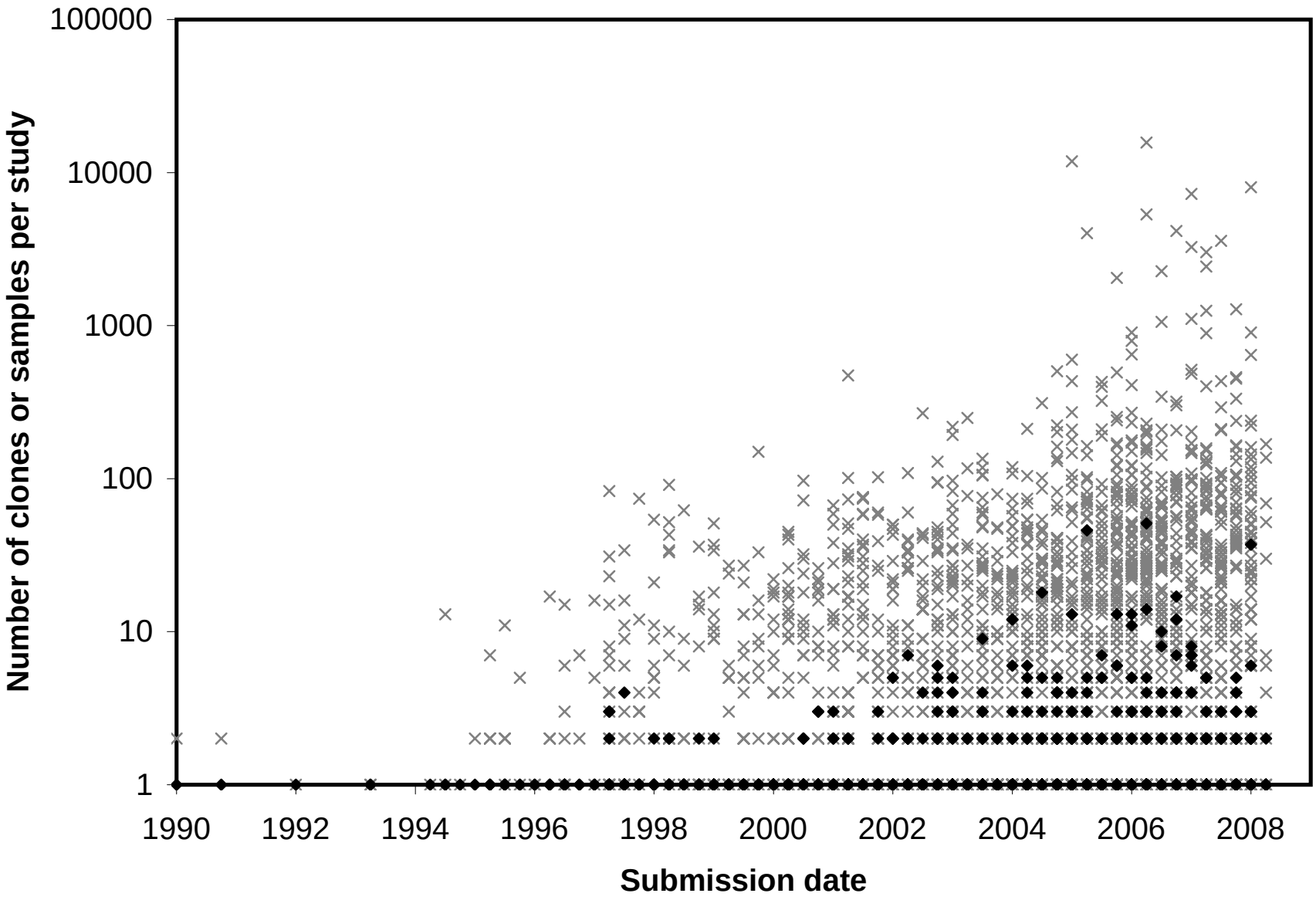
38. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL: **Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB**. *Appl Environ Microbiol* 2006, **72**:5069-5072.

39. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO: **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB**. *Nucleic Acids Res* 2007, **35**:7188-7196.

40. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006, **22**:2688-2690.

41. Dalevi D, Desantis TZ, Fredslund J, Andersen GL, Markowitz VM, Hugenholtz P: **Automated group assignment in large phylogenetic trees using GRUNT: GRouping, Ungrouping, Naming Tool**. *BMC Bioinformatics* 2007, **8**:402.

Figure 1: Increases in the number of full-length (>1200 nt) Sanger-sequenced 16S clones (grey Xs) and samples (black diamonds) per study since 1990. Note logarithmic scale on Y-axis. Data were obtained from the greengenes database [38] and parsed such that each entry for a given study was assumed to be an independent clone, and each unique entry in the "isolation source" field was assumed to be an independent sample. No attempt was made to further manually curate the clone or sample counts based on information in publications or manual examination of clone names or other information, beyond spot-checking the accuracy of the parsing scripts. Submission dates were converted to numeric format by taking the submission year and adding zero for submissions in January through March, 0.25 for April-June, 0.5 for July-September and 0.75 for October-December.


100000
10000
1000
100
10
1
Number of clones or samples per study
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
Submission date