Journal of Biology

Minireview

Mapping the protistan 'rare biosphere'

Scott C Dawson and Kari D Hagen

Addresses: Department of Microbiology, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA.

Correspondence: Scott C Dawson. Email: scdawson@ucdavis.edu

Abstract

The use of cultivation-independent approaches to map microbial diversity, including recent work published in *BMC Biology*, has now shown that protists, like bacteria/archaea, are much more diverse than had been realized. Uncovering eukaryotic diversity may now be limited not by access to samples or cost but rather by the availability of full-length reference sequence data.

See research article http://www.biomedcentral.com/1741-7007/7/72

For several decades now, microbiologists have lived with the growing realization that the majority of extant microbes are not in our culture collections. In fact, our historical reliance on cultivation to identify and quantify microbes has resulted in our missing upwards of 95% of extant bacterial and archaeal diversity. Using cultivation-independent molecular approaches to identify microbes by genetic sequence - specifically small subunit ribosomal RNA (SSU rRNA) sequences - we have begun to map the true microbial diversity of the Earth. This cultivation-independent approach to identifying diversity has recently benefited from the development of next-generation sequencing technology and a concomitant drop in sequencing costs. With respect to molecular surveys of microbial diversity in natural environmental samples, pyrosequencing approaches provide unprecedented sampling depth. Such deep sequencing has purportedly uncovered a rare and extensive biosphere of bacteria and archaea with a diversity that is perhaps several orders of magnitude greater than we had anticipated [1]. And although there may have been initial overestimations of the magnitude of the 'rare biosphere' because of the intrinsic sequence error rate pyrosequencing produces [2], many rare and novel microbes are still being discovered at taxonomic levels ranging from phyla to species.

Although we have observed protists in the wild for over three centuries and have classified them on the basis of morphology and motility, use of the modern molecular techniques that have provided insight into uncultivated bacterial/archaeal diversity has been limited in protists. Recent eukaryote-specific cultivation-independent studies to assess the extent of microbial eukaryotic diversity have identified many novel taxa at a range of taxonomic levels [3,4]. And although it may seem astounding to some that we could be unaware of phylum-level protistan taxa, the discovery of novel eukaryotic SSU rRNA genes in natural environmental samples mirrors the gaps in our understanding of bacterial and archaeal diversity. Nearly every time we have surveyed an environment using SSU rRNA cultivation-independent methods, we have found that it contains more protistan species than we know from our culture collections or sequence databases.

The extent of protistan diversity

Precisely how many protistan species have we missed? In their analysis of two marine anoxic environments using massively parallel pyrosequencing recently published in *BMC Biology*, Stoeck and colleagues [5] conclude we have indeed missed considerable protistan diversity. To determine the extent of a possible protistan 'rare biosphere', Stoeck *et al.* [5] sequenced about 250,000 eukaryotic-specific V9 variable regions of the SSU rRNA. Previous surveys of these two anoxic environments were limited to Sanger sequencing of SSU rRNA clone libraries. Novel protistan diversity was still identified, although at lower estimated levels [6]. Deep sequencing allows the extensive characterization of SSU RNA PCR amplicons, and the authors [5] thereby determined that over 90% of the SSU rRNA sequence diversity was derived from individual rare sequences, each of which was identified less than ten times. They have thus indeed revealed the existence of a protistan 'rare biosphere'. Although estimates of microbial eukaryotic diversity could be inflated because of sequencing errors [2], high copy numbers of the rRNA operon in eukaryotes [7], or simply high sequence variability or divergence in closely related organisms [7], Stoeck *et al.* [5] suggest that there are probably higher numbers of rare protistan microbes than estimated from previous molecular surveys.

Some natural environments may harbor more undiscovered protistan diversity than others, and this was a primary motivation for the analysis of anoxic environments by Stoeck *et al.* [5]. Such environments are perhaps the least studied because of the presumption that eukaryotes (including animals) require oxygen and are limited by sulfide. Yet anaerobic protists are common inhabitants of anoxic environments, deriving energy through fermentation

rather than aerobic respiration. Our anthropocentric frame of reference has probably limited our search, and thus our understanding of eukaryotic diversity and ecology, by focusing primarily on oxic environments. Importantly, the authors [5] detected eukaryotic microbes from all major protistan groups in their anoxic sediment samples, indicating that these environments harbor the same types of eukaryotic microbes as more familiar oxic environments.

Defining and quantifying eukaryotic microbial diversity using rRNA

The availability and cost-effectiveness of high-throughput sequencing - albeit of relatively short DNA fragments - forces the issue of how to define diversity. If we use only the V9 regions, definitions of eukaryotic species or operational taxonomic units (OTUs) will be based solely on a single short variable region of SSU rRNA. Traditionally, we have been confident in our assessment of novelty because we have used longer full-length rRNA sequences for phylogenetic analyses. Larger fragments (over 1,000 nucleotides) contain more phylogenetic signal and allow us to compare and classify environmental sequences relative to known protistan rRNAs in genetic databases [8]. A primary motivation for using deep sequencing to understand microbial diversity is that such strategies may obviate potential erroneous estimations of diversity resulting from the construction of clone libraries of larger fragments. Currently, we can achieve massive sequencing only of shorter DNA sequences (less than 400 nucleotides) with pyrosequencing. Thus, we are left with the tradeoff for environmental surveys of having either massive sequence numbers or longer sequence lengths but not both.

In bacteria and archaea, full-length SSU rRNA genes with 97% or less sequence identity are generally defined as distinct OTUs [8]. In their investigations, Stoeck and colleagues [5] classified 'unique' protistan sequences using both liberal (one nucleotide difference per V9 region sequence) and conservative (five or seven nucleotide differences in the V9 region) definitions of novelty. The number of OTUs derived (several hundred to several thousand) depended on the criteria used; however, considerable diversity was detected, supporting the idea of a protistan 'rare biosphere'. For now, it seems reasonable to use the bacterial/archaeal OTU definition (97% sequence identity) for microbial eukaryotes; if only the short V9 variable region were sequenced, an amplicon with just three nucleotide differences would define an OTU.

Is the deep sequencing strategy more effective than Sanger sequencing of larger fragment clone libraries for identifying and characterizing protistan diversity? Although Stoeck and colleagues [5] reanalyzed environmental samples previously used to generate Sanger sequenced clone libraries, they did not pyrosequence larger SSU rRNA amplicons from the same libraries. A direct comparison of deep sequencing of larger full-length rRNA amplicons with the shorter V9 variable region amplicons is needed to determine the effectiveness of each sequencing strategy. In addition, the limited number of full-length sequences in our current eukaryotic SSU rRNA database (several orders of magnitude fewer than the available bacterial and archaeal rRNA sequences) complicates the taxonomic identification of the shorter V9 fragments [9].

Lastly, both methods of sequencing of rRNA genes to assess protistan diversity require PCR amplification. All PCR-based rRNA surveys rely on 'eukaryotic-specific' SSU rRNA primers that are, notably, derived from the existing (and limited) rRNA sequence data [9]. Deep sequencing strategies will uncover new sequences, but only if the regions from which the primers are designed have been sufficiently sampled and sequenced for eukaryotic diversity. Because much of known cultivated (and uncultivated) eukaryotic diversity has been determined from PCR amplification of SSU rRNA using conserved sequence regions, we still have little actual sequence data at the extreme 5' and 3' ends of SSU rRNA. We can only search for sequences similar to those that are already known. How, then, can we design PCR primers that are both specific to eukaryotes but also broad enough to identify unknown groups? Even in this study [5], in which the V9 region at the extreme 3' end of SSU rRNA was PCR amplified and sequenced, there could be more unamplified and thus hidden eukaryotic diversity. As we continue to deeply sequence more variable regions we might find that we have again underestimated eukaryotic diversity. Additional sequencing of genomes from cultivated protists or identification of eukaryotic rRNA sequences from directly sequenced metagenomic studies could provide more reference sequences.

Further challenges in assessing eukaryotic microbial diversity

With the work of Stoeck *et al.* [5], we expand the outlines of known (albeit uncultivated and uncharacterized) protistan diversity. The approach taken and adapted to protistan SSU rRNA [5] builds on previous work and harnesses the power of deep sequencing to map eukaryotic microbial diversity. Perhaps the main technological impediment to uncovering eukaryotic diversity is no longer access to samples or cost of sequence, but rather the availability of existing full-length sequence data from cultivated and uncultivated protists for use as a reference database. Shorter sequence reads enable us to identify novel sequences, but these shorter reads must either be mapped onto full-length sequences for accurate phylogenetic identification, or used as 'phylogenetic stains' in rRNA-targeted fluorescent *in situ* hybridizations to identify target organisms [10]. Despite these looming challenges, the high-throughput sequencing strategy of Stoeck *et al.* [5] confirms and expands what we surmised from previous clone-based studies that surveyed eukaryotic diversity in

anoxic environments: we have missed much of it. Application of this and other molecular strategies to assess diversity will help us to close these gaps and understand the true nature and extent of protistan diversity.

References

1. 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 USA* 2006, **103:**12115-12120.
2. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P: **Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates.** *Environ Microbiol* 2009, doi:10.1111/j.1462-2920.2009.02051.x.
3. Dawson SC, Pace NR: **Novel kingdom-level eukaryotic diversity in anoxic environments.** *Proc Natl Acad Sci USA* 2002, **99:**8324-8329.
4. Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB: **Protists are microbes too: a perspective.** *ISME J* 2009, **3:** 4-12.
5. Stoeck T, Behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora MJ, Chistoserdov A, Orsi W, Edgcomb VP: **Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities.** *BMC Biol* 2009, **7:**72.
6. Behnke A, Bunge J, Barger K, Breiner HW, Alla V, Stoeck T: **Microeukaryote community patterns along an O2/H2S gradient in a supersulfidic anoxic fjord (Framvaren, Norway).** *Appl Environ Microbiol* 2006, **72:**3626-3636.
7. Rooney AP: **Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in Apicomplexans.** *Mol Biol Evol* 2004, **21:**1704-1711.
8. Caron DA, Countway PD, Savai P, Gast RJ, Schnetzer A, Moorthi SD, Dennett MR, Moran DM, Jones AC: **Defining DNA-based operational taxonomic units for microbial-eukaryote ecology.** *Appl Environ Microbiol* 2009, **75:**5797-5808.
9. 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.
10. Amann RI, Krumholz L, Stahl DA: **Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology.** *J Bacteriol* 1990, **172:**762-770.

Published: 29 December 2009
doi:10.1186/jbiol201
© 2009 BioMed Central Ltd