Review

Sequencing our way towards understanding global eukaryotic biodiversity

Holly M. Bik[1], Dorota L. Porazinska[2], Simon Creer[3], J. Gregory Caporaso[4], Rob Knight[5,6] and W. Kelley Thomas[1]

[1] Hubbard Center for Genome Studies, University of New Hampshire, 35 Colovos Rd, Durham, NH 03824, USA
[2] Fort Lauderdale Research and Education Center, University of Florida, IFAS, 3205 College Avenue, Fort Lauderdale, FL 33314, USA
[3] School of Biological Sciences, Environment Centre Wales, Deiniol Road, College of Natural Sciences, Bangor University, Gwynedd, LL57 2UW, UK
[4] Department of Computer Science, Northern Arizona University, Flagstaff, AZ 86011, USA
[5] Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309, USA
[6] Howard Hughes Medical Institute, Boulder, CO 80309, USA

Microscopic eukaryotes are abundant, diverse and fill critical ecological roles across every ecosystem on Earth, yet there is a well-recognized gap in understanding of their global biodiversity. Fundamental advances in DNA sequencing and bioinformatics now allow accurate *en masse* biodiversity assessments of microscopic eukaryotes from environmental samples. Despite a promising outlook, the field of eukaryotic marker gene surveys faces significant challenges: how to generate data that are most useful to the community, especially in the face of evolving sequencing technologies and bioinformatics pipelines, and how to incorporate an expanding number of target genes.

Microscopic eukaryotes: global dominance, scant knowledge

Microscopic eukaryotic taxa are abundant and diverse, playing a globally important role in the functioning of ecosystems [1,2] and host-associated habitats [3]. Here, we consider taxa generally represented by individuals <1 mm in size; the term 'microscopic eukaryotes' thus encompasses meiofaunal metazoans (e.g. Nematoda, Platyhelminthes, Gastrotricha and Kinorhyncha; see Glossary), microbial representatives of fungi and deep protist lineages (Alveolata, Rhizaria, Amoebozoa, algal taxa in the Chlorophyta and Rhodophyta, etc.), and eggs and juvenile stages of some larger metazoan species. These ubiquitous eukaryote groups play key roles as decomposers, predators, producers and parasites, yet little is known about their biology, ecology and diversity. Analyses of eukaryotic community structure often reveal divergent lineages [4–6] and long lists of previously undiscovered sequences [7,8]. Nematodes, for instance, account for 80–90% of all metazoans on Earth, yet <4% of the estimated >1 million species are formally known and described [9]. This discrepancy between known and estimated diversity is common for all microscopic eukaryote groups and generally stems from the difficulty of applying traditional approaches in species identification to high-throughput sequence data. Traditional approaches, although well validated, do not scale to the large numbers of sequences now being collected [6,9–12].

In many ways, the problems faced in the study of microscopic eukaryotes mirror those facing studies of archaea and bacteria. The exploration of archaeal and bacterial diversity long ago adopted a molecular taxonomy [13]; early uses of high-throughput sequencing allowed the characterization of microbial taxa in environmental samples ranging from the oceans [14,15] to our own bodies [16,17]. These approaches not only illuminate a path for the exploration of eukaryotic diversity, but also highlight the pitfalls that will need to be addressed along the way. Although advances in the study of archaeal and bacterial diversity provide valuable knowledge and infrastructure for high-throughput analyses of eukaryotes, eukaryotes also have four unique features. First, for many groups of microscopic eukaryote, there is access to biologically informative morphology and a substantial body of existing taxonomic resources (expertise, keys and specimen vouchers).

Glossary

454: common term for the Roche GS platforms that use bead emulsion methods and typically return approximately 1.2 million sequences per full plate run (reads currently averaging 350–450 bp).
Illumina: company producing the newest Hi-Seq and MiSeq platforms, which uses bridge amplification to produce 1.6 billion sequences per eight-lane Hi-Seq flow cell (current max length for paired-end reads is 300 bp).
Marker gene surveys: high-throughput environmental sequencing utilizing homologous genetic loci (e.g. 16S, 18S rRNA) amplified via conserved primer sets.
Meiofauna: a loose term to define metazoan species with a body size <1 mm, although this size fraction often varies across studies.
Metagenomics: high-throughput, random sequencing of genomic DNA from environmental isolates.
Metatranscriptomics: high-throughput sequencing of expressed gene transcripts (mRNA) from environmental isolates.
OTU (operational taxonomic unit): typically defined from high-throughput sequence data that are filtered for quality and subsequently clustered under pairwise identity cutoffs.
Pyrosequencing: general term referring to light-based high-throughput sequencing techniques (e.g. 454).

Corresponding author: Bik, H.M. (holly.bik@gmail.com).

0169-5347/$ – see front matter doi:10.1016/j.tree.2011.11.010

Box 1. Intragenomic rRNA variation in eukaryotes

Ribosomal RNA in eukaryotes is encoded by 18S, 5.8S and 28S subunit genes, organized in tandemly repeated arrays within a genome. The number of gene copies can vary dramatically across taxa, with eukaryotic species exhibiting hundreds to many thousands of ribosomal arrays [18,20]; these are sometimes found at a single locus but are also known to exist in multiple distinct loci [20]. Concerted evolution results in high levels of identity among intraspecific repeats but higher divergence across interspecific gene copies [69]. However, the number of rRNA copies can vary dramatically even within species [70], confounding the ability to correlate the number of reads generated in a marker gene survey with the number of individuals in a sample. Although the phenomenon of concerted evolution [69] predicts that new mutations are rapidly propagated across the rRNA gene copies within a species, it is clear that intragenomic ribosomal variation is extensive in some cases [71] and some of these variants might represent pseudogenes [72]. Such variation can be incorporated into the appropriate OTU by clustering approaches, although levels of rRNA diversity are significantly different across taxa [31] and significant empirical data will be required to understand the pattern and consequences of intragenomic variation across diverse eukaryotes [73].

Therefore, researchers can (and should) collect and employ morphological metadata as a valuable component of marker gene surveys, especially when it is desirable to compare results to historical or fossil specimens from which DNA cannot be extracted. Second, the increased complexity of eukaryotic genomes is correlated with an increased number and variability of the traditional target loci for molecular taxonomy (rRNA; Box 1) [18]. Although the ribosomal locus varies in copy number (1–15) and length heterogeneity in archaea and bacteria [19], the variation can be more extensive in eukaryotes (extending to tens of thousands of copies in some taxa [18,20]) Box 2. This issue severely complicates both the clustering of sequences into operational taxonomic units (OTUs) and the use of read counts for estimating species abundances (Box 3). Third, most eukaryotes have mitochondrial genomes. In multicellular animals, the mitochondrial genome evolves rapidly (especially in the noncoding regions), offering higher resolution for detecting more recent evolutionary forces; mitochondria might thus provide a basis for large-scale analyses of gene flow. Finally, many groups of eukaryote appear to evolve with a smaller contribution of horizontal gene transfer (e.g. metazoans and fungi; [21]). Consequently, the evolutionary framework inferred from a single locus can better reflect the history of these eukaryotic genomes as a whole.

Box 2. Biases in physical and genomic sampling

A typical assessment of eukaryotic diversity derived from environmental DNA comprises field, lab and bioinformatics components (methodologically similar to archaeal and bacterial approaches; Figure 1, main text), each accompanied by specific challenges; a recent review by Creer *et al.* [9] provides a comprehensive outline of the workflow and methodological considerations involved with eukaryotic studies.

Replicated sampling schemes must effectively capture eukaryotic community diversity, given that species diversity and population densities can vary (spatially and temporally) by several orders of magnitude. Although aquatic organisms from pelagic habitats can simply be concentrated from their environment [11,28], interstitial eukaryotes are accompanied by a solid matrix (soil or sediment) that precludes direct environmental DNA extractions large enough to capture the diversity of rare taxa. Approaches for extraction have been tried and tested in unconsolidated marine [9,30] and estuarine [33] sediments, but comprehensively separating the eukaryotic specimens from terrestrial soils, including muds, clays and large amounts of organic matter and inhibitors, poses additional challenges. Biases in taxon representation should be assumed whenever such extraction protocols are adopted. A further consideration for whole-sediment extractions is the potential existence of extracellular DNA or transient fauna [74,75].

Following the separation of organisms from soil or sediment, bulk environmental DNA is extracted and specific gene targets are amplified via PCR using appropriately selected, barcoded [76,77] degenerate primers. The goal of marker gene surveys is to appraise as broad a taxonomic breadth as possible, but both DNA extraction and PCR amplification are key steps that introduce biases [11,31,74]. To minimize such biases, different DNA extraction approaches can be compared [9], the use of PCR-primer cocktails implemented [78,79] and the conservation of degenerate primer binding sites assessed using rRNA databases [45,46,51] and primer design tools.

Although both physical and genomic sampling involve many steps known to introduce biases, certain actions can be taken to reduce such discrepancies. For example, applying multiple methods to extract organisms physically and using different combinations of primer sets (and genetic loci) to minimize the potential exclusion of taxa. To date, there has been little effort towards quantifying the impact of sampling protocols in eukaryotes (although sampling bias has been exhaustively assessed in archaea and bacteria [79–82], and parasitology studies [83,84]), a robust understanding of these biases will be critical for the interpretation of community assemblages and informing practical applications for high-throughput techniques.

Outstanding questions for high-throughput marker gene studies include:

- How diverse are communities of microscopic eukaryotes?
- How geographically structured are these communities?
- Are there taxonomic or life-history biases for true cosmopolitan species?
- To what degree do environmental factors, bacteria and archaea, and eukaryotic communities interact to drive biotic assemblages?

Emerging insight from environmental data

Following earlier 16S rRNA (reference GenBank accession **X80721.1** for *Escherichia coli*) investigations of archaeal and bacterial communities [14,22], high-throughput marker gene approaches were developed for different groups of microscopic eukaryote using the 18S nuclear small subunit rRNA gene (nSSU; reference GenBank accession **X03680.1** for *Caenorhabditis elegans*), focusing on protists [11,12,23–26] and meiofauna [9,10,27]. Similar to 16S investigations, these early 18S studies uncovered concordant patterns of high eukaryotic richness and an extended rare biosphere [11,28]. Although the field is not yet mature, environmental data sets are already yielding novel molecular taxonomic insights into the magnitude and composition of the eukaryotic biosphere in a range of habitats. However, bioinformatic analyses of eukaryotic control communities indicate that additional work is needed to recover actual taxon richness [29–31].

Both 454 and Illumina sequencing data sets have suggested that the composition of marine meiofaunal [30] and protist [5,32] communities differ significantly from estimates derived from morphological taxonomy; in these sequence data sets, the unexpected prominence of turbellarian flatworms and monothalamous foraminiferans has highlighted biases stemming from sample preservation methods in traditional taxonomic approaches (see also Box 2). The isolation of deep alveolate lineages further suggests that divergent taxa identified in high-throughput datasets lack the characteristic morphological features typified by closely related clades [5]. Marker gene surveys