only possible, approach to documenting prokaryotic diversity.

Second, there is a substantial body of literature on the ecology, physiology, and life cycles of microbial eukaryotes that is not cross-referenced to a molecular sequence. Very little of the known biodiversity of microbial eukaryotes has been characterised genetically: for example, only 7% of known species of *Gymnodinium*, an abundant, cosmopolitan, marine, and freshwater dinoflagellate genus, are represented by a sequence in GenBank [6]. The analysis of sequences without effectively cross-referencing has led to reports of novel 'phylum level' lineages (reviewed in [7]), later found to have been known for over a century, for example, [8]. Unless considerable effort is made to link traditional and new approaches to documenting biodiversity, much of this context will remain disconnected and will need to be reassembled.

Finally, enormously divergent rates of evolution have occurred among different genes and in different lineages of eukaryotes [9], driven by a variety of processes [10]. For this reason, large-scale sequencing initiatives such as the Barcode of Life have assessed multiple reference genes, sometimes in combination, for groups such as plants [11]. Because the scope of the evolutionary diversity of protists eclipses that of plants, animals, or fungi [12], it should come as no surprise that no single marker gene or percentage difference is likely to be informative to the same degree for all groups.

We propose that the solution is to greatly increase the application of high throughput sequencing to species defined by typological means. Multiple potential marker genes can be assessed for each protist group to determine their evolutionary rates. Field sampling is needed to obtain new reference material and discriminate between intra- and inter-specific variation. Novel methods will be required to obtain reference sequences from unculturable groups, for example, single cell transcriptomics coupled with microscopic identification. This will be a massive task. Currently, there is no single repository of all described protist names, while bacteria, archaea, fungi, plants, and increasingly animals, have good on-line taxonomic treatments available. Few known species are cultured, and the development of a base-line will require considerable work. Such an approach has already begun: one example is the Moore Foundation's Marine Microbial Eukaryote Transcriptome project (http://marinemicroeukaryotes.org/), which will sequence ~360 typologically defined protists. We believe that this represents a promising start to the difficult path to unlocking the real diversity and ecology of eukaryotes through the integration of transcriptomics, field studies, and typologically defined species.

References

1 Bik, H.M. *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol. Evol.* 27, 233–243

2 Corliss, J.O. (1982) Numbers of species comprising the phyletic groups assignable to the kingdom Protista. *J. Parotozool.* 29, 499

3 Rossello-Mora, R. and Amann, R. (2001) The species concept for prokaryotes. *FEMS Microbiol Rev.* 25, 39–67

4 Casabianca, S. *et al.* (2012) Population genetic structure and connectivity of the harmful dinoflagellate *Alexandrium minutum* in the Mediterranean Sea. *Proc. R. Soc. B* 279, 129–138

5 Doolittle, W.F. and Zhaxybayeva, O. (2009) On the origin of prokaryotic species. *Genome Res.* 19, 744–756

6 Thessen, A. *et al.* The taxonomic significance of species that have only been observed once: the genus *Gymnodinium* (Dinoflagellata) as an example. *PLoS ONE* (in press)

7 Moreira, D. and López-García, P. (2002) The molecular ecology of microbial eukaryotes unveils a hidden world. *Trends Microbiol.* 10, 31–38

8 Lohmann, H. (1908) Untersuchung zur Feststellung des vollständigen Gehaltes des Meeres an Plankton. *Wissenschaftliche Meeresuntersuchungen* N.F. 10, 129–370 (in German)

9 Parfrey, L.W. *et al.* (2010) Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Syst. Biol.* 59, 518–533

10 Petit, R.J. and Excoffier, L. (2009) Gene flow and species delimitation. *Trends Ecol. Evol.* 24, 386–393

11 CBOL Plant Working Group (2009) A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 12794–12797

12 Patterson, D.J. (1999) The diversity of eukaryotes. *Am. Nat.* 154, S96–S124

0169-5347/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.tree.2012.08.011 Trends in Ecology and Evolution, December 2012, Vol. 27, No. 12

Metagenomics will highlight and drive links to taxonomic data: reply to Murray

Holly M. Bik[1] and W. Kelley Thomas[2]

[1] UC Davis Genome Center, One Shields Ave, Davis, CA 95616, USA
[2] Hubbard Center for Genome Studies, University of New Hampshire, 35 Colovos Road, Durham, NH 03824, USA

Although our understanding of microbial eukaryote taxa is already being transformed by high-throughput sequencing approaches, the research community is keenly aware of the looming computational challenges and pressing need to integrate disparate, complementary, data types such as taxonomy, morphology, and functional knowledge.

Before the field can progress further, Murray [1] argues the need to first understand eukaryotic species concepts, link molecular data with the vast body of taxonomic literature, and develop sophisticated computational metrics for identifying divergent evolutionary rates. In contrast to Murray's view that historic knowledge must be prioritized for digitization before further sequence data is generated, we argue that genomic analysis of historic knowledge can

Corresponding author: Bik, H.M. (hbik@ucdavis.edu)

(and will) catch up and be computationally linked to biodiversity knowledge derived from metagenomic approaches. This drive toward integration will still benefit from conscious efforts of all scientists in the field.

Commonly applied biological species concepts are an important but not always accurate indicator of species boundaries in microbial eukaryotes; for example, in nematodes convergent morphology is pervasive (DNA sequences have been shown to be a more accurate method for delimiting species [2]) and investigations of reproductive isolation are time-consuming and not always feasible (e.g., due to sampling constraints and our inability to culture most environmental species). Unexpected molecular differences can often be validated through empirical observation, providing an additional dimension that allows deep characterization of cryptic speciation even in well-known taxa [3]. Delimiting species and characterizing long-term evolutionary patterns are common goals across all forms of biodiversity research, regardless of empirical approach. Metagenomics can arguably offer a more rigorous knowledge of species boundaries and evolution, due to the depth and scale of high-throughput sequence datasets. Thus, although species delimitation in microbial eukaryotes has historically eschewed molecular approaches, the value and objectivity of DNA data must not be underestimated.

Murray cites a continued 'molecular neglect' of microbial eukaryote species, even for obviously abundant, ecologically important taxa. We agree that such taxa associated with significant ecological and physiological knowledge should be prioritized for molecular characterization, particularly in regard to functional genomics. The inaccessibility of most historical taxonomic work means that species lacking representation in public sequence databases are effectively invisible to the genomics community. Although it remains difficult to definitively separate 'rare biosphere' taxa [4] from sequencing errors and chimeras [5], high-throughput sequencing approaches can unambiguously pinpoint abundant species and track spatial/temporal changes among dominant taxa. However, just as taxonomists must strive to collect molecular data alongside morphology, genomic researchers must also emphasize integrative methods to attach informative biology to environmental sequences. Unclassified, unidentified environmental DNA sequences represent both the bane and norm of GenBank.

Finally, in the absence of a comprehensive taxonomic framework, research conducted without species names can still provide critical biological insights. Cross-sample comparisons (e.g., Principal Component Analysis) can be applied to metagenomic datasets and reveal overarching evolutionary patterns in the absence of taxonomy. For high-throughput sequence data, robust annotation of species names will require sophisticated mathematical concepts (e.g., digesting probability distributions for short reads placed over a reference phylogenetic topology [6]); using such approaches, metagenomics will foster and expand phylogenetic frameworks for biodiversity. However, these methods represent active areas of research, as it remains difficult to accurately reconcile data from sequence-based approaches with existing taxonomic frameworks. As computational and database resources evolve (and biologists make every effort to collaborate with computational researchers), we anticipate that taxonomic names will naturally become an integrated and routine component of high-throughput environmental analyses.

Although marker-based studies do not specifically inform us about function, metagenomic and metatranscriptomic approaches can be effectively used towards this goal. Transcriptomics is one way forward, but researchers must also aim to generate and link disparate data types (metagenomes, rRNA and marker gene amplicons, and metatranscriptomes), particularly for ecosystems and taxa that cannot easily be scrutinized in a laboratory setting. In the absence of typological approaches, metatranscriptomes can be used to annotate and complement environmental metagenome datasets when sequenced in parallel [7]. Furthermore, we can harness knowledge from phylogenetic and phylogenomic studies to identify and build reference datasets for conserved, phylogenetically-informative marker genes. For example, protein-coding marker genes such as those used to refine the backbone of the Eukaryote Tree of Life [8] can be identified and mined from metagenomic datasets, and phylogenetic placement algorithms subsequently applied to assign taxonomy to environmental sequences [6]. A baseline – a colossal, global pool of reference sequence data – will inevitably develop as high-throughput fields progress, with such progress accelerated by large-scale sequencing efforts emphasizing spatial/temporal sampling such as the Earth Microbiome Project (http://www.earthmicrobiome.org), NEON (http://www.neoninc.org) and TARA oceans (http://oceans.taraexpeditions.org). The clarity and complexity of this baseline will continue to evolve as resources for large-scale comparative data analysis become more readily available. High-throughput sequencing has not reduced scientific effort as much as shifted the emphasis in a typical research workflow; the onus has rapidly moved from data collection (e.g., morphological taxonomy) to data processing/analysis (e.g., bioinformatics). Although new technologies have afforded biological insights at an unprecedented scale, achieving a narrow, focused understanding of taxonomic assemblages will be critical for building accurate pictures of ecosystem function.

References

1 Murray, S. (2012) Transcriptomics and microbial eukaryotic diversity: a way forward. *Trends Ecol. Evol.* 27, 651–652

2 De Ley, P. (2000) Lost in worm space: phylogeny and morphology as road maps to nematode diversity. *Nematology* 2, 9–16

3 Derycke, S. *et al.* (2008) Disentangling taxonomy within the *Rhabditis (Pellioditis) marina* (Nematoda, Rhabditidae) species complex using molecular and morhological tools. *Zoolog. J. Linnean Soc.* 152, 1–15

4 Sogin, M.L. *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120

5 Bik, H.M. *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol. Evol.* 27, 234–244

6 Matsen, F.A. *et al.* (2010) pplacer: linear time maximum-likelihood Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinform.* 11, 538

7 Moran, M. (2009) Metatranscriptomics: eavesdropping on complex microbial communities. *Microbe* 4, 329–335

8 Parfrey, L.W. *et al.* (2010) Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Syst. Biol.* 59, 518–533

0169-5347/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.tree.2012.09.005 Trends in Ecology and Evolution, December 2012, Vol. 27, No. 12