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

Box 3. Major challenges for eukaryotic marker gene surveys

A major challenge for marker gene surveys is the accurate identification of biological taxa across multiple samples. Although overall per-nucleotide error rates can be lower in high-throughput than in Sanger sequencing [85], larger numbers of sequences mean larger numbers of sequences containing errors. Although it is possible to predict and identify the source of errors in high-throughput data sets, such errors result from the interplay of multiple factors, such as position in sequence, presence of homopolymers and physical location of DNA on sequencing platforms [31,85,86]. Regardless of sequencing method, PCR-derived artifacts (e.g. chimeras) can artificially increase estimates of diversity [31].

A second challenge concerns the ability to quantify the absolute abundance of individuals based on sequence read counts. This relationship is complicated by the extreme variation in the number of rRNA gene copies per nucleus among species and the number of nuclei per individual. A practical solution has been to compare environmental communities only in terms of relative taxon abundance (normalizing sequence reads per OTU). Early proof-of-concept control experiments in nematodes revealed a strong consistency in rRNA patterns but highlighted the difficulty of correlating OTUs with biological species and defining absolute abundances [10,87,88]; although technical replicates agreed well with one another, the rRNA read number per individual was highly variable, even within a single specimen. OTUs corresponding to reference Sanger sequences comprise the majority of the reads from any given species, but many other variants exist, some differing from known references by >3 bp [87]. When one considers that some well-described nematode species differ at a single nucleotide in the 18S rRNA gene [87], it is clear that this intragenomic variation can confound efforts to differentiate sequencing errors and the 'rare biosphere' in environmental data sets. Although a rare biosphere [14] certainly exists, any observations of low abundance taxa need to be critically evaluated. Rare species can be important, especially for ecosystem responses, and these low-abundance taxa tend to be diverse (e.g. [89]).

thus facilitate objective comparisons of taxonomic richness at higher-level ranks, on a scale that is not possible using morphology alone. Accordingly, the cost-effective, high-throughput comparative power of marker gene surveys is ideally suited to (among other applications) biomonitoring [33–35], phylogeography [36] and exploring the relationships between biological diversity and ecosystem functioning using a 'systems ecology' approach [37].

Analyzing high-throughput data

Over the past few years, high-throughput sequencing techniques have been informed by rapid progress in sequencing technology, bioinformatics tools and analytical pipelines. Here, we present an overview of the analytical considerations for high-throughput studies. Following sample collection, extraction of environmental DNA, PCR and sequencing (Figure 1), large data sets can be processed using many existing tools (Table 1, Figure 2).

Denoising

Denoising (i.e. correcting pyrosequencing errors), using tools such as AmpliconNoise [38] or Denoiser [29], is often encouraged as a pre-processing step for 454 data sets, although it requires considerable computational resources.

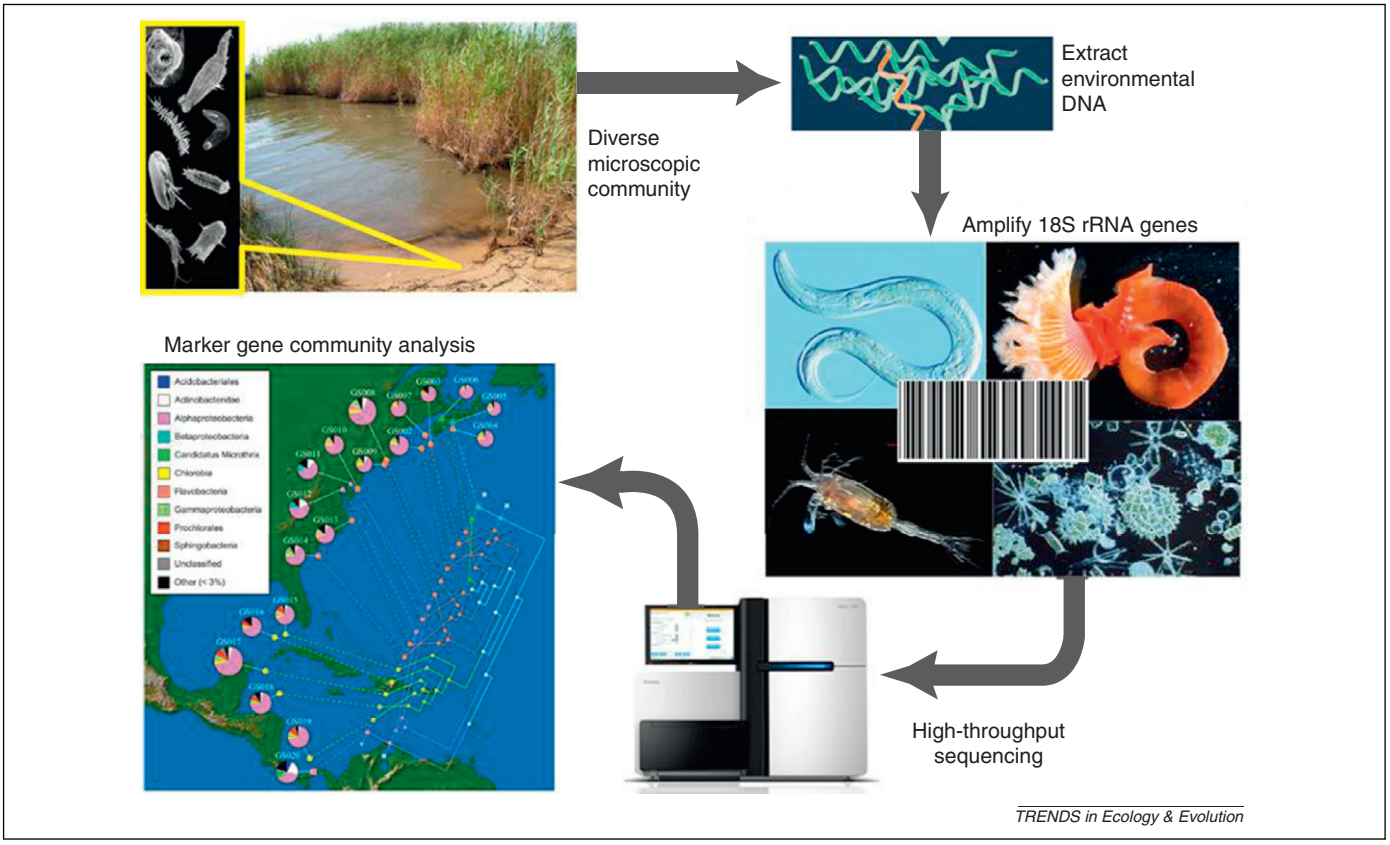

Figure 1. Typical standardized workflow (from environment to sequences) for high-throughput marker gene studies. Soils and sediments are typically frozen upon collection (–80 °C to preserve RNA) and brought back to the lab for bulk extraction of environmental DNA. Marker genes (e.g. rRNA) are amplified from genomic extracts using barcoded, conserved primer pairs. Following high-throughput sequencing (typically conducted on 454 or Illumina platforms), data sets are processed and clustered into operational taxonomic units (OTUs) under a range of pairwise identity cutoffs. OTUs are subsequently used to conduct α- and β-diversity analyses, summarize community taxonomy and interpret assemblages in a phylogenetic context. (Depiction of community analysis modified from [56].).

Table 1. Online resources and popular software tools for eukaryotic marker gene surveys

Resource	Capabilities	Interface	Website	Refs
AmpliconNoise	Denoising	Command line executables	http://code.google.com/p/ampliconnoise/	[38]
Denoiser	Denoising	QIIME	http://www.qiime.org	[47]
QIIME	Data processing, OTU picking, taxonomy assignment and ecological analyses	Command Line and Amazon Cloud	http://www.qiime.org	[47]
OCTUPUS	Data processing, OTU picking and taxonomy assignment	Pipeline of perl scripts	http://octupus.sourceforge.net/	[30]
VAMPS	Data processing, OTU picking, taxonomy assignment and ecological analyses	Web-based tools	http://vamps.mbl.edu	NA
Galaxy	Data processing, OTU clustering and ecological analyses	Web and graphical user interface (GUI)	http://main.g2.bx.psu.edu	[55]
CANGS	Data processing, taxonomy assignment and ecological analyses	Pipeline of perl scripts	http://i122server.vu-wien.ac.at/pop/software.html	[95]
RDP	Data processing, OTU picking, taxonomy assignment and ecological analyses	Web-based tools	http://pyro.cme.msu.edu/	[45]
CLOTU	Data processing, OTU picking and taxonomy assignment	Web-based pipeline	http://www.bioportal.uio.no	[96]
Mothur	Data processing, OTU picking, taxonomy assignment, chimera checking and ecological analyses	Command line executables	http://www.mothur.org/	[42]
ESPRIT	OTU picking	Command line executables	http://www.biotech.ufl.edu/people/sun/esprit.html	[44,97]
ChimeraSlayer	Chimera checking	Command line executables	http://microbiomeutil.sourceforge.net/	[49]
Perseus	Chimera checking	Command line executables	http://code.google.com/p/ampliconnoise/	[38]
UCHIME	Chimera checking	Command line executable	http://www.drive5.com/uchime/	[50]
MEGAN4	Taxonomic assignment (BLAST-based) and exploring trees	GUI	http://ab.inf.uni-tuebingen.de/software/megan/	[98]
EPA	Tree insertion	Web and GUI	http://i12k-exelixis3.informatik.tu-muenchen.de/raxml	[40]
pplacer	Tree insertion	Command line executables	http://matsen.fhcrc.org/software.html	[39]
guppy	Edge principal coordinate analysis	Command line executables	http://matsen.fhcrc.org/software.html	[58]
TopiaryExplorer	Taxon labeling and visualization in tree topologies	GUI	http://topiaryexplorer.sourceforge.net	[67]
GenGIS	Data visualization	GUI with python console	http://kiwi.cs.dal.ca/GenGIS/Main_Page	[56]
VisTrails	Data visualization	GUI (XML-based)	http://www.vistrails.org	[57]
MG-RAST	Metagenomic analysis and data storage	Web	http://metagenomics.anl.gov	[62]
Dryad	Data repository	Web	http://datadryad.org	[99]
SILVA	Curated reference database and sequence alignment	Web	http://www.arb-silva.de	[51]
MetaBar	Metadata entry and tracking system	Graphical web tool	http://www.megx.net/metabar/	[100]

Although denoising reduces sequencing error, existing algorithms employ read abundance information [38] and, consequently, can also function to remove valuable biological signals (e.g. intragenomic rRNA variants, Box 1) and rare species from eukaryotic data sets. It seems probable that sequence 'noise' might only present concerns for specific biological questions, such as species counts, although further studies comparing the affect of denoising on a wide range of environmental data sets and applicable to different sequencing platforms are needed.

Processing raw reads

Primers, sequencing adaptors and barcode tags are typically removed from raw sequencing reads, with the relevant metadata (sample site or primer name) inserted into FASTA headers or sequence mapping files. Short and noisy reads are discarded completely, and low-quality base calls are trimmed from longer sequences (typically, the tail end of raw reads where sequencing quality deteriorates). A multitude of tools can be utilized for these basic processing steps (Table 1).

OTU picking

OTU picking tools fall into three general categories: (i) *de novo* OTU picking, where reads are clustered only against one another; (ii) closed-reference OTU picking, where reads are searched against a database and clusters are defined by the best database match for each read, with nonmatching reads discarded; and (iii) open-reference OTU picking, where clusters are similarly defined by the best database matches, but nonmatching reads are instead retained and clustered *de novo*. It is important to note that

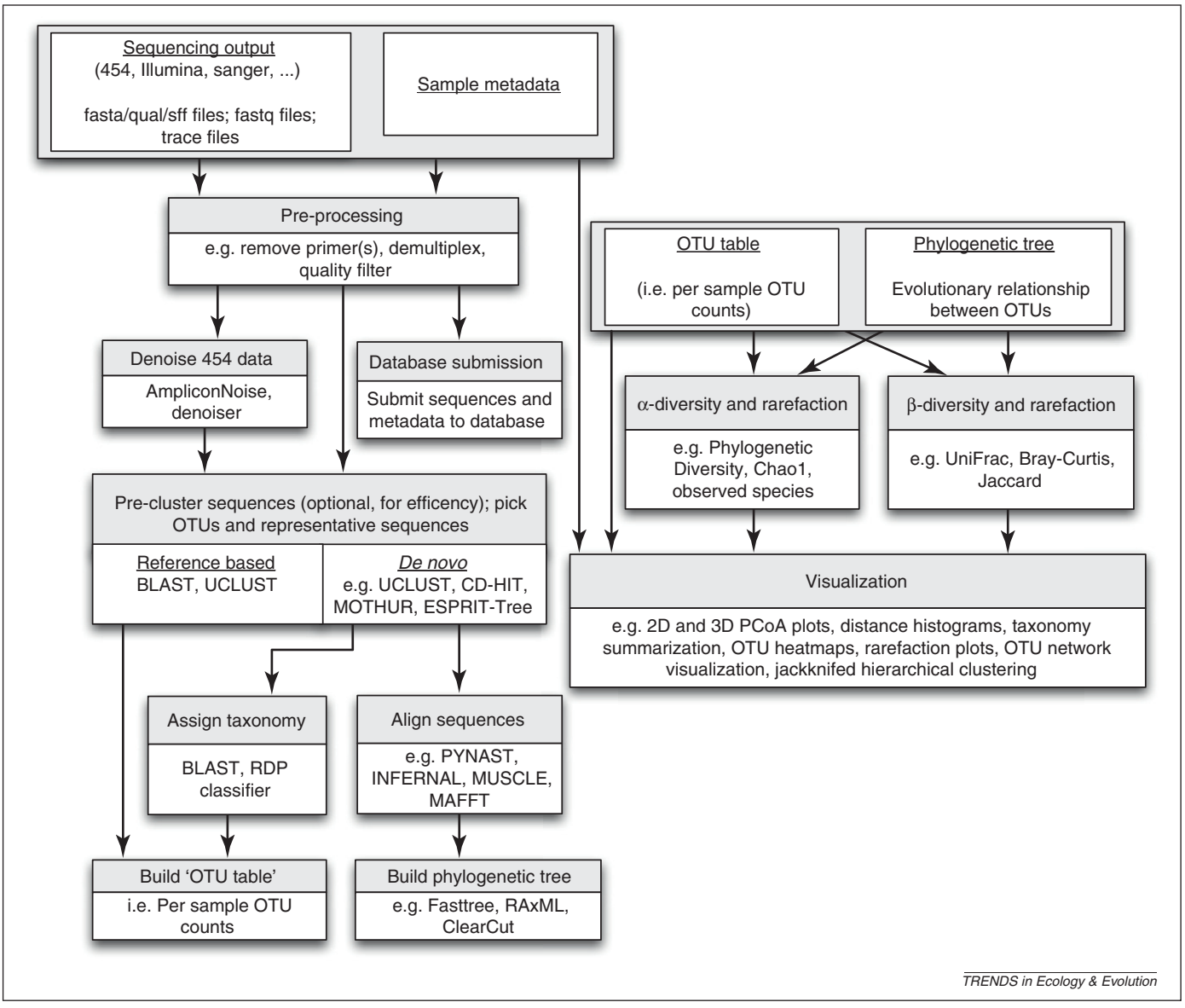

Figure 2. High-throughput studies follow a common workflow that begins with raw sequence data and sample metadata (primer barcodes and environmental data). Raw data is filtered and processed, with the option of denoising (a step currently applicable only to 454 data) before operational taxonomic units (OTUs) are picked through reference-based or *de novo* approaches. OTU picking can include pre-clustering steps such as single linkage pre-clustering (SLP, [93]), prefix–suffix filtering or collapsing of identical sequences to reduce compute time (all methods available within the QIIME pipeline [47]); the recommended and default OTU picking workflow in QIIME currently involves sorting sequences by abundance, collapsing identical reads, picking OTUs *de novo* with uclust, and subsequently inflating the 'identical reads' to recapture abundance information about the initial sequences. Taxonomy is next assigned to OTU reference sequences, followed by construction of an OTU abundance matrix and a phylogenetic tree; when working with a closed reference-based OTU picking protocol, it is not necessary to make taxonomic assignments or build a phylogenetic tree, as these can be obtained directly from the reference data set. These outputs can be subsequently utilized for ecological diversity analyses and visualization approaches.

high-throughput analyses do not necessarily depend on defining OTUs; direct phylogenetic placement methods [39,40] can handle unclustered, processed sequence reads (Figure 3). However, the sheer volume of raw sequence data means that some degree of filtering or clustering is usually necessary for minimizing downstream computational demands. In this sense, OTU picking can be viewed simply as another processing step, rather than an attempt to define biological species from sequence data *per se*.

Pre-clustering reads for computational efficiency

Pre-clustering of raw reads is commonly employed to reduce the compute time of OTU picking. Most OTU picking approaches require all-by-all comparisons (each sequence compared to every other sequence), with total computational time being proportional to the square of the number of input sequences; thus, algorithms useful for small data sets quickly become infeasible with larger data sets. The most common pre-filtering method collapses only those sequences that are 100% identical prior to OTU picking. A similar filter involves collapsing sequences that have identical prefixes or suffixes (nested sequences beginning at the 5′ or 3′ end of the gene, respectively) meeting a user-specified maximum length parameter. In all cases, abundance information (defined by unique ID tags for filtered reads) is retained for each representative sequence.

De novo OTU picking

De novo algorithms are the most ubiquitous, primarily because they do not require a specific external database

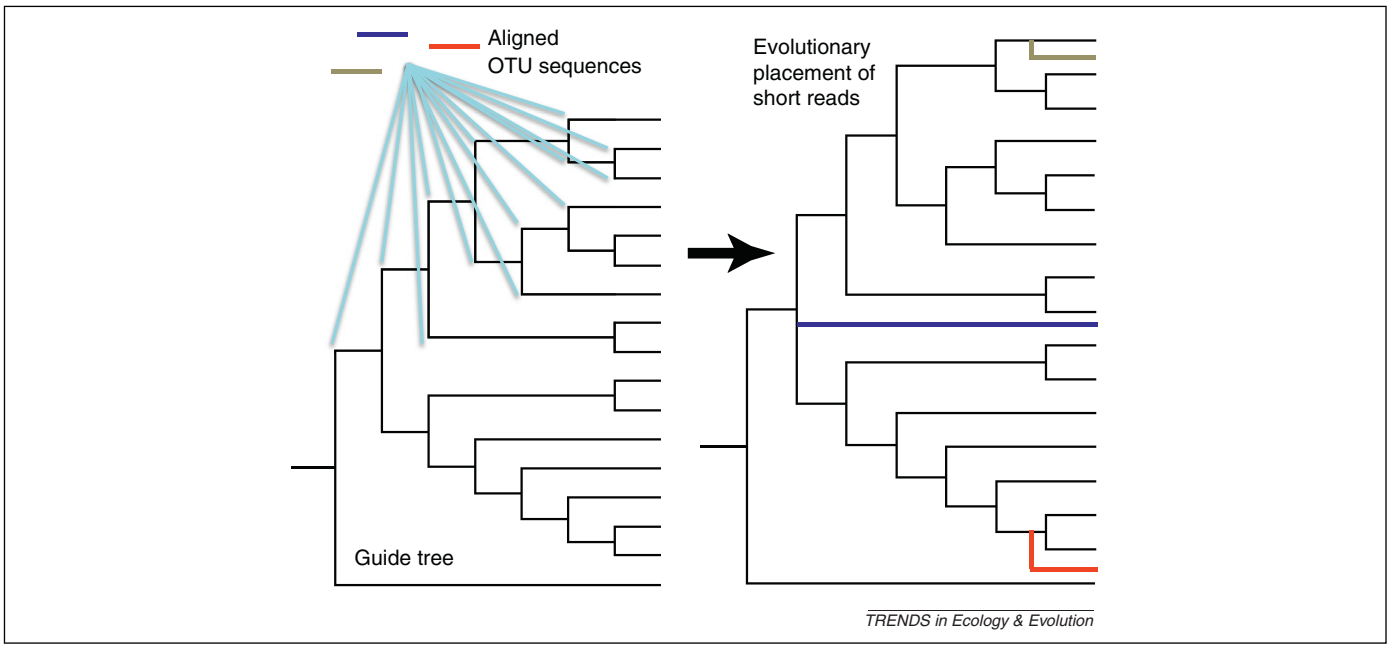

Figure 3. Operational taxonomic unit (OTU) reference sequences can be placed into an evolutionary context using tools such as pplacer [39] or the evolutionary placement algorithm (EPA [40]), which place short reads into a guide tree framework constructed from full-length reference sequences. For each pre-aligned OTU or sequencing read, likelihood scores are calculated for all possible positions in the tree, and the sequence is subsequently inserted at the node exhibiting the best score.

and inherently retain all input sequence reads; therefore, they cannot introduce biases stemming from database incompleteness. There are many existing algorithms for *de novo* OTU picking, including uclust [41], OCTUPUS [30], mothur [42], cd-hit [43] and ESPRIT-Tree [44] (Table 1). A useful strategy in *de novo* OTU picking (and the *de novo* step of open-reference OTU picking) is to pre-sort sequences by their abundance. For algorithms that sort through the sequences in order (e.g. uclust), the most abundant sequences (less likely to represent errors) are the reads initially used to form OTU clusters.

Closed-reference OTU picking

When using closed-reference OTU picking, a minimal threshold is usually applied for accepting database hits as matches, and sequences that do not match this threshold are discarded as failures. Although closed-reference OTU picking can exclude many good reads, it can also be considered a strict quality filter (and screen for contaminants) because all observed sequences must be similar to previously known sequences. Closed-reference OTU picking has primarily been applied to bacterial and archaeal studies that can take advantage of large data repositories and multiple curated databases [e.g. the Ribsomal Database Project (RDP) [45], Greengenes [46]]. This method is more difficult for eukaryotic studies, given the limited number of eukaryotic database resources (although the SILVA database [47] can be harnessed) and the patchy coverage across diverse taxa in terms of published rRNA sequences.

BLAST and uclust are both commonly used for closed-reference OTU picking, with assignments differing in the use of local (BLAST) versus global (uclust) alignments. For this reason, uclust is more conservative: for a sequence to match at 97% identity, the uclust match must span the full length of the read, whereas a BLAST-assigned sequence can exhibit 97% identity, but only cover, for example, only 50% of the total read length. Minimum aligned percentages can be applied to circumvent this issue with BLAST, although uclust is considerably faster for OTU picking. However, global alignments can be problematic when the sequence termini are especially noisy, as is the case with Sanger sequencing (where both ends tend to be lower quality) and high-throughput platforms (where quality drops off, sometimes dramatically, towards the 3′ end of the sequence).

Despite discarding valid reads that are not represented in the reference database, closed-reference OTU picking exhibits several benefits over *de novo* clustering. First, reads from multiple hypervariable gene regions (or genomic loci [48]) can be compared in a single analysis if clustered against full-length reference sequences. In *de novo* clustering, independent, non-overlapping amplicons would instead cluster independently of each other. Closed-reference approaches also facilitate direct comparisons and meta-analysis across studies utilizing the same collection of reference sequences, as new unique data sets are incrementally added into a centralized database resource. Closed-reference OTU picking additionally results in stable OTU identifiers, so the 'same' OTU can be pinpointed in different studies. Finally, this method is easily parallelized and, therefore, more amenable to increasingly large sequence outputs, such as those generated on Illumina platforms. Because no new clusters are created in the process, it is possible to split an input collection of reads into tens or hundreds of smaller input collections, process all of the sub-inputs in parallel and subsequently collate the results.

Open-reference OTU picking

Open-reference OTU picking is a combination of *de novo* and closed-reference OTU picking strategies, and any pair

of methods could theoretically be combined for tailored workflows. However, users should be wary of the differences in how OTUs are defined if combining strategies such as BLAST and *de novo* uclust for OTU picking. The open-reference approach begins the same way as closed-reference OTU picking, but sequences that fail to hit a reference sequence are retained through *de novo* clustering at the end of the process. Although the initial reference-based OTU picking can be run in parallel (as in closed-reference OTU picking), the *de novo* process must consider all remaining sequences simultaneously. For eukaryotic studies (and in general), open-reference approaches are preferable over closed-reference OTU picking, because environmental data sets typically contain a significant proportion of taxa with no close relatives in public sequence repositories. *De novo* clustering thus retains unknown diversity and supplies high-throughput database resources with new divergent lineages.

Identifying chimeras

New algorithms are emerging for the identification of chimeras (Box 4) in high-throughput data sets, including ChimeraSlayer [49], Perseus [38] and UCHIME [50]. Both Perseus and UCHIME take advantage of the intuitive assumption that chimeras (i.e. hybrid, daughter sequences) should be less frequent than parental sequences that have undergone at least one more round of PCR amplification. Control experiments show that Perseus and UCHIME (i.e. tools used without a reference database) both outperform ChimeraSlayer (which compares taxonomic lineage information from reference databases), at least in 16S data sets [38,50]. Detecting hybrid sequences between closely related taxa currently represents one of the most significant challenges. In practice, no method is 100% effective for preventing or identifying chimeras: chimeras can theoretically be generated at any position along an amplicon, and detection ability will correlate with the length of the chimeric fragment.

Box 4. The perils of chimeras

Chimeras are *in vitro* DNA artefacts derived from the mixture of two or more parent molecules in a PCR reaction of a homologous gene region [9,90], typically formed when incomplete extension occurs in a preliminary round of PCR and the resulting sequence fragment acts as a primer for a different sequence in subsequent thermal cycles. The result is an artificial recombinant molecule with discrete break points (and specific hotspots along an amplicon; [91]), corresponding to the transition between the different parent molecules [38]. Chimeras substantially overinflate richness estimates and suggest the presence of spurious taxa and OTUs, potentially across multiple samples [38]. They are additionally likely to skew the interpretation of the extent of the 'rare biosphere' [14], because chimeras usually appear as divergent, low-frequency sequences [29,92].

Sample and taxonomic diversity, the composition and structure of the locus employed, and DNA polymerases and thermal cycling conditions will all interact to create different levels of chimeric amplicons [38,49]. Identifying optimal 'low chimera' PCR protocols for marker loci will require a series of carefully controlled experiments, but general rules include keeping the number of PCR thermal cycles to a minimum and increasing extension times [28]. The latter allows DNA polymerases to reach the 3′ terminus of the target molecule and minimizes chimera formation in longer PCR reactions.

Assigning taxonomy

Although interesting biological patterns can be investigated in the absence of species names, taxonomic frameworks are useful for relating sequences to what is known about the organisms. Several different approaches can be used for assigning taxonomy (alone or in conjunction), including BLAST, Global Alignments for Sequence Taxonomy (GAST), probabilistic classifiers, or tree-based assignments. BLAST assignments utilize pairwise alignment scores and have the advantage of being the easiest to use: a local database and single step can identify close relatives from millions of published sequences. BLAST methods can assign taxonomy from public sequence repositories (e.g. GenBank or EMBL) or use a smaller database of reference sequences with trusted taxonomy (e.g. curated from SILVA [51]). GAST approaches utilize BLAST searches against reference data sets for hypervariable gene regions, in conjunction with global alignments and RDP classifier scores; this method is presently implemented in the VAMPS pipeline (http://vamps.mbl.edu). Probabilistic approaches, such as the RDP classifier, instead apply a naïve Bayesian classifier to match eight-base sequence 'words' to a reference training set (the RDP [45] or another user-defined database when implemented in QIIME [47]), also returning confidence scores for each taxonomic assignment. Tree-based assignment (Figure 3) provides an alternate (and, in theory, more robust) approach: a predefined guide tree containing full-length reference sequences is used to place unknown OTU sequences within a known phylogeny. Using likelihood scores, aligned short reads are tested across all nodes in the reference topology and subsequently inserted as a branch at the highest scoring position. Current tools include the RAxML-based Evolutionary Placement Algorithm [40], accessible through a web interface (http://i12k-exelixis3.informatik.tu-muenchen.de/raxml), and the open-source command-line tool pplacer [39] that can be run on local machines.

Ecological analyses

Vague or absent taxonomy does not preclude ecological analyses or the testing of ecological hypotheses (e.g. phylogeographic patterns or community assemblages); the production of OTU tables and taxonomy tables are thus intermediate steps towards these end goals. A suite of new tools is on the horizon (Figure 4), and many between samples useful ecological analyses are currently incorporated within the QIIME pipeline [47], including community summaries (i.e. pie, bar or area charts detailing taxonomic proportions), heatmaps displaying OTU abundance across sample sites, α-diversity (including rarefaction based on OTU counts, Chao1 estimation, or phylogenetic diversity), phylogenetically informed β-diversity and ordination (Principal Coordinate Analysis and jackknifed UPGMA analysis based on UniFrac distances [52] between samples), and OTU network analysis (cytoscape [53]). High-throughput data can also be imported into other ecological workbenches (e.g. Primer-E [54] or Galaxy [55]), which provide some additional multivariate statistics. Other methods include visualization workbenches (Figure 4b), such as GenGIS [56] and VisTrails [57], which allow graphical explorations of high-throughput sequence

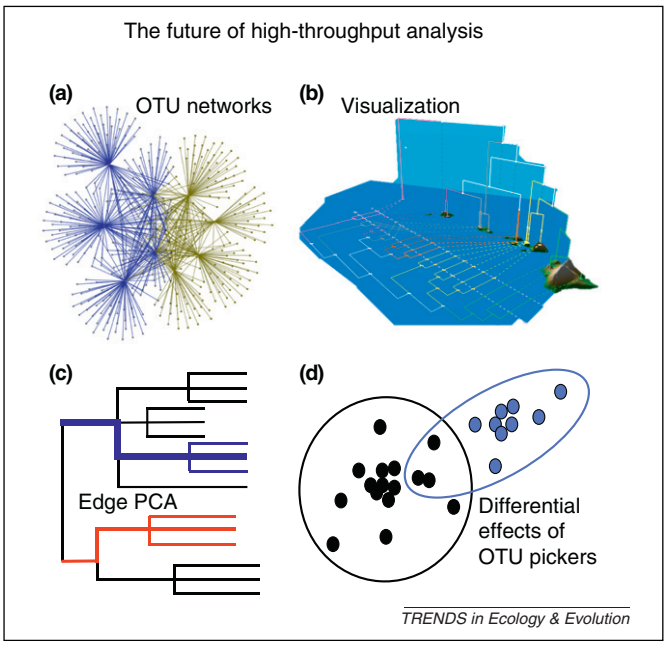

Figure 4. High-throughput biodiversity research is an active and rapidly evolving field. Future analytical tools will expand towards several exciting, emerging research areas, including: **(a)** operational taxonomic unit (OTU) network analysis; **(b)** visualization as an exploratory tool; **(c)** edge principal component analysis (PCA) to identify biological lineages that define community assemblages; and **(d)** quantifying the impact of different OTU picking strategies on cluster formation. Modified from [94] **(b)** and [58] **(c)**.

data, and Edge Principal Component Analysis (Figure 4c) for visually identifying key community lineages in tree topologies [58].

Comparative meta-analyses

As publicly available high-throughput data accumulate, there will soon be available an unprecedented view of microscopic eukaryote species on a truly global scale. Unfortunately, the ability to carry out such comparative eukaryotic meta-analyses is presently hindered by methodological heterogeneity, data accessibility and underdeveloped computational infrastructure. The differential use of primer sets or gene regions (varying according to taxon and investigator preference) presents a substantial long-term challenge. Conserved primers targeting informative 18S regions [9] will provide a broad taxonomic view of eukaryote communities (although with consistent amplification biases) and provide useful data for biodiversity research, whereas other primers [e.g. internal transcribed spacers (ITS) for fungi [59]] can deliver species-level data and address certain specific biological questions.

Future meta-analyses will ideally harness a robust subset of biologically meaningful OTUs, potentially made accessible through reference-based OTU picking. This evokes an urgent question: how should the output of different OTU pickers be compared and the deposition of 'real' OTUs (e.g. species analogues) in reference databases be ensured? The goal of clustering is not to maximize or minimize the final OTU count, yet current approaches do not allow researchers to visualize and quantify how OTU picking affects the biological interpretation of sequence data [60]. New tools (Figure 4d) or tree-based algorithms (employing phylogenetic species concepts) could be applied to dynamically define species 'clouds' from high-throughput rRNA sequence reads.

Data storage and access presents another significant barrier towards the effective use of published sequences. The fluctuating status and accessibility issues of the NCBI Short Read Archive (SRA) makes this resource less than suitable for meeting long-term community needs. Repositories such as Dryad (http://datadryad.org) and CAMERA [61] exist as alternatives, but are limited by file size (10 GB limit for Dryad) or are targeted towards a different purpose (e.g. microbial metagenomics for CAMERA). At present, MG-RAST [62] accepts amplicon datasets in addition to its primary focus on metagenomes, and can be used as a resource for sequence deposition. Increasing data volumes also present an increasing need for data standards: for both sequences and metadata, standard workflows and documentation can provide a solid foundation to drive innovation and promote information discovery. The Genomic Standards Consortium [63] (http://gensc.org) is working to coordinate efforts such as MIMARKS standards (minimum information about a marker gene sequence [64]), promoting the collection of rich, contextual metadata to ensure the long-term utility of published data sets and encourage comparative studies.

The need for robust guide trees and reference databases

Limited eukaryote reference databases and inconsistent taxonomic levels currently hinder the development of robust computational pipelines for marker gene data (e.g. reference-based OTU picking and confident taxonomy assignments [60]), and limit the use of tree-based methods and deeper sequencing technologies with shorter sequence reads (such as those derived from the Illumina platforms). Microscopic eukaryotic taxa have been historically underrepresented in public repositories, with some divergent phyla represented by a single published rRNA sequence (e.g. the phylum Loricifera contains one reference sequence in SILVA release 106 [51]). The accuracy of BLAST-derived taxonomy depends on the database coverage for a given taxonomic group: for well-sampled groups (e.g. Arthropoda or Annelida), it is possible to obtain genus-level accuracy, whereas only phylum-level accuracy (at best) might be possible for neglected 'minor' phyla (e.g. Loricifera, Gnathostomulida or deep protist lineages). Divergent lineages, common in poorly characterized environmental samples, often return few or no 'good' matches (>90%) from public repositories. Underpopulated databases also prevent the use of tools requiring taxonomic lineage information (e.g. RDP classifier [65]) and decrease the accuracy of probabilistic tools [66]. Databases for microscopic eukaryotes covering alternate loci are smaller (28S rRNA [51]) or almost nonexistent (mtDNA), deterring the use of additional genes in high-throughput studies. Ultimately, a much larger collection of full-length eukaryotic reference sequences (or whole genomes) will be necessary for identifying erroneous reads, and providing a strong link between sequence data and morphology.

In theory, tree-insertion methods [e.g. pplacer [39] or the evolutionary placement algorithm (EPA) [40]; Figure 3] will circumvent many of the issues that confound BLAST assignments, and these tools have recently become more

broadly accessible with their incorporation into the QIIME pipeline [45]. The availability of software such as TopiaryExplorer [67] will greatly aid the visualization and interpretation of OTUs within phylogenetic trees. No high-quality, densely sampled guide tree exists for the eukaryotic domain, although recent efforts have substantially improved the backbone of deep splits across major taxa [68] (a eukaryotic guide tree currently exists for the SILVA reference database within the ARB software suite, but this tree also suffers from the above-mentioned taxon sampling issues, and represents only an approximated phylogenetic topology). In the future, tree-based tools will be a critical component of high-throughput analyses, providing an additional line of evidence to supplement BLAST hits (particularly where reference sequences are unclassified or misnamed), helping to identify divergent lineages (long branch taxa with no close reference sequences) and aiding the development of phylogenetic species concepts to delineate OTUs as putative species.

Future outlook and challenges

Although substantial progress is being made with high-throughput eukaryotic studies, many challenges lie ahead. A strong emphasis on morphological and environmental data collection, guide trees and reference sequence databases, and open-access repositories for high-throughput data sets is urgently needed. Large-scale sequencing methods offer substantial promise for basic and applied biodiversity research, yet the wider adoption of these approaches will probably hinge on the ease-of-use and accuracy of analytical tools and pipelines. Traditional ecologists and biologists typically have limited computational backgrounds, yet computer scientists rarely design cutting-edge tools with this fact in mind. Similarly, computational pipelines are not always subjected to rigorous benchmarking, unit testing, or proof-of-concept validation in relation to real-world biology. Encouraging an ongoing dialog between computer scientists (who want to develop clever algorithms) and biologists (who want knowledge of ecology and taxonomy) will enable interdisciplinary collaborations that promote the development of easy-to-use, reliable, and well-documented software.

Concluding remarks

The promise and accessibility of high-throughput sequencing is now poised to attract increasing numbers of non-computationally trained researchers. With ongoing declines in the price of sequencing, deep sequencing will inevitably represent the most cost-effective approach for elucidating ecological and functional roles of complex communities. However, exploiting the data will require the continued refinement of bioinformatics pipelines and database resources, which will in turn require an ongoing and reciprocal collaboration between computational and biological scientists (Figure 5). A key challenge for high-throughput studies is to move beyond descriptions of differences between ecosystems, towards capturing whole-ecosystem function. For any given sample, marker gene surveys will define community assemblages, metagenomics will reveal the genomic potential of a community, and metatranscriptomics, metaproteomics and metabolomics will provide a snapshot in time of community function. Combined with environmental information, these complementary approaches will enable the driving forces to be defined that govern species distributions and drive ecological assemblages on scales ranging from microscopic to global.

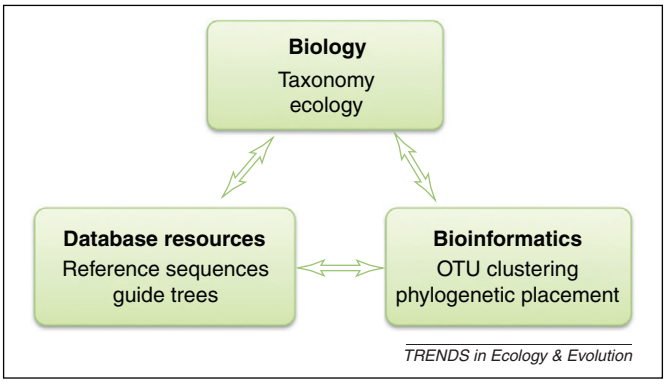

Figure 5. For future success, biodiversity research must adhere to a trifecta of biology, bioinformatics and database resources; none of these foci can exist in isolation and each must serve to inform the others. Biological questions drive high-throughput studies, and so computational pipelines and cyberinfrastructure must function to provide knowledge of ecosystem processes. Likewise, computational resources must be complementary, whereby bioinformatic outputs are effectively data-based, and evolving database resources produce continuing refinements in analytical pipelines. Seamless integration between these sectors will be crucial for enabling comparative metadata analyses and untangling complex ecological patterns; for example, mining published data sets for co-occurring species, or linking specific operational taxonomic units (OTUs) with environmental parameters (pH, salinity, temperature, etc.).

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments that significantly helped to improve an earlier version of the manuscript. Development of this manuscript was made possible by a Catalysis Meeting award (HB and WKT) from the National Evolutionary Synthesis Center. HB and WKT supported through NSF (DEB-1058458 and NIH (NIH-1P20RR030360-01)). SC supported by a Natural Environment Research Council (NERC) New Investigator Grant (NE/E001505/1), a Post Genomic and Proteomics Grant (NE/F001266/1) and a Molecular Genetics Facility Grant (MGF-167). DP acknowledges funding from USDA/CSREES – TSTAR (grants 2006-04347 and 2008-34135-19505), NSF (DEB-0450537) and the CR-USA Foundation. RK and JGC supported in part by the National Institutes of Health, Bill and Melinda Gates Foundation, the Crohns and Colitis Foundation of America, the Sloan Indoor Environment program and the Howard Hughes Medical Institute.

References

1 Danovaro, R. *et al.* (2008) Exponential decline of deep-sea ecosystem functioning linked to benthic biodiversity loss. *Curr. Biol.* 18, 1–18
2 Wardle, D.A. (2006) The influence of biotic interactions on soil biodiversity. *Ecol. Lett.* 9, 870–886
3 Wegner Parfrey, L. *et al.* (2011) Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Front. Microbiol.* 2, 1–16
4 Behnke, A. *et al.* (2006) Microeukaryote community patterns along an O_2/H_2S Gradient in a supersulfidic Anoxic Fjord (Framvaren, Norway). *Appl. Environ. Microbiol.* 72, 3626–3636
5 Groisillier, A. *et al.* (2006) Genetic diversity and habitats of two enigmatic marine alveolate lineages. *Aquat. Microb. Ecol.* 42, 277–291
6 Richards, T. and Bass, D. (2005) Molecular screening of free-living microbial eukaryotes: diversity and distribution using a meta-analysis. *Curr. Opin. Microbiol.* 8, 240–252
7 Rosling, A. *et al.* (2011) Archaeorhizomycetes: unearthing an ancient class of ubiqutous soil fungi. *Science* 333, 876–879
8 Jones, M.D.M. *et al.* (2011) Discovery of novel intermediate forms redefines the fungal tree of life. *Nature* 474, 200–203

9 Creer, S. *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls, and promises. *Mol. Ecol.* 19, 4–20
10 Porazinska, D.L. *et al.* (2009) Evaulating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. *Mol. Ecol. Resour.* 9, 1439–1450
11 Stoeck, T. *et al.* (2010) Multiple marker parallel tag enviornmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* 19, 21–31
12 Stoeck, T. *et al.* (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol.* 7, 72
13 Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science* 276, 734–740
14 Sogin, M.L. *et al.* (2006) Microbial diversity in the deep sea and the unexplored 'rare biosphere'. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120
15 Amaral-Zettler, L. *et al.* (2010) A global census of marine microbes. In *Life in the World's Oceans: Diversity, Distribution and Abundance* (McIntyre, A.D., ed.), pp. 223–245, Blackwell Publishing
16 Fierer, N. *et al.* (2008) The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 105, 17994–17999
17 Turnbaugh, P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature* 457, 480–484
18 Prokopowich, C.D. *et al.* (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome* 46, 48–50
19 Pei, A.Y. *et al.* (2010) Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl. Environ. Microbiol.* 76, 3886–3897
20 Zhu, F. *et al.* (2005) Mapping of picoeukaryotes in marine ecosystems with a quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.* 52, 79–92
21 Andersson, J.O. (2005) Lateral gene transfer in eukaryotes. *Cell. Mol. Life Sci.* 62, 1182–1197
22 Huber, J.A. *et al.* (2007) Microbial population structures in the deep marine biosphere. *Science* 318, 97–100
23 Amaral-Zettler, L. *et al.* (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* 4, e6372
24 Medinger, R. *et al.* (2010) Diversity in a hidden world: potential an limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Mol. Ecol.* 19, 32–40
25 Orsi, W. *et al.* (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. II. Habitat specialization. *ISME J.* 5, 1357–1373
26 Edgcomb, V. *et al.* (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J.* 5, 1344–1356
27 Bik, H.M. *et al.* (2012) Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Mol. Ecol.* 21, 1048–1059
28 Nolte, V. *et al.* (2010) Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Mol. Ecol.* 19, 2908–2915
29 Reeder, J. and Knight, R. (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat. Methods* 7, 668–669
30 Fonseca, V.G. *et al.* (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nat. Commun.* 1, 98
31 Behnke, A. *et al.* (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environ. Microbiol.* 13, 340–349
32 Lecroq, B. *et al.* (2011) Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13177–13182
33 Chariton, A. *et al.* (2010) Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Front. Ecol. Environ.* 8, 233–238
34 Hajibabaei, M. *et al.* (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE* 6, e17497
35 Pfrender, M.E. *et al.* (2010) Assessing macroinvertebrate biodiversity in freshwater ecosystems: advances and challenges in DNA-based approaches. *Q. Rev. Biol.* 85, 319–340
36 Emerson, B. *et al.* (2011) Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. *Philos. Trans. R. Soc. B: Biol. Sci.* 366, 2391–2402
37 Purdy, K.J. *et al.* (2010) Systems biology for ecology: from molecules to ecosystems. *Adv. Ecol. Res.* 43, 87–149
38 Quince, C. *et al.* (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinform.* 12, 38
39 Matsen, F.A. *et al.* (2010) pplacer: linear time maximum-likelihood Bayesian phyogenetic placement of sequences onto a fixed reference tree. *BMC Bioinform.* 11, 538
40 Berger, S.A. and Stamatakis, A. (2011) Aligning short reads to reference alignments and trees. *Bioinformatics* 27, 2068–2075
41 Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461
42 Schloss, P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541
43 Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659
44 Cai, Y. and Sun, Y. (2011) ESPRIT-Tree: hierarchial clustring analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res.* 39, e95
45 Cole, J.R. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37, D141–D145
46 DeSantis, T. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072
47 Caporaso, J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336
48 Caporaso, J.G. *et al.* (2011) Host-associated and free-living phage communities differ profoundly in phylogenetic composition. *PLoS ONE* 6, e16900
49 Haas, B.J. *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504
50 Edgar, R.C. *et al.* (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200
51 Pruesse, E. *et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188–7196
52 Lozupone, C. *et al.* (2011) Unifrac: an effective distance metric for microbial community comparison. *ISME J.* 5, 169–172
53 Cline, M.S. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* 2, 2366–2383
54 Clarke, K. and Gorley, R. (2006) PRIMER v6: User Manual/Tutorial, PRIMER-E.
55 Blankenberg, D. *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* 19.10.1-21
56 Parks, D.H. *et al.* (2009) GenGIS: a geospatial information system for genomic data. *Genome Res.* 19, 1896–1904
57 Callahan, S.P. *et al.* (2006) VisTrails: visualization meets data management, In *Proceedings of ACM SIGMOD, June 27–29, 2006*, ACM Press. pp. 745–747
58 Matsen, F.A. and Evans, S. (2011) Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. ar Xiv:1107.5095v1101.
59 Bellemain, E. *et al.* (2010) ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. *BMC Microbiol.* 10, 189
60 Christen, R. (2008) Global sequencing: a review of current molecular data and new methods available to assess microbial diversity. *Microbes Environ.* 23, 253–268
61 Seshadri, R. *et al.* (2007) CAMERA: a community resource for metagenomics. *PLoS Biol.* 5, e75
62 Meyer, F. *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* 9, 386

63 Field, D. *et al.* (2011) The Genomic Standards Consortium. *PLoS Biol.* 9, e1001088
64 Yilmaz, P. *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* 29, 415–420
65 Wang, Q. *et al.* (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267
66 Werner, J.J. *et al.* (2012) Impact of training sets on classification of high-throughput bacerial 16S rRNA gene surveys. *ISME J.* 6, 94–103
67 Pirrung, M. *et al.* (2011) TopiaryExplorer: visualizing large phylogenetic trees with environmental metadata. *Bioinformatics* DOI: 10.1093/bioinformatics/btr517
68 Wegener Parfrey, L. *et al.* (2010) Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Syst. Biol.* 59, 518–533
69 Dover, G. (1982) Molecular drive: a cohesive mode of species evolution. *Nature* 299, 111–117
70 Averbeck, K.T. and Eickbush, T.H. (2005) Monitoring the mode and tempo of concerted evolution in *Drosophila melanogaster* rDNA locus. *Genetics* 171, 1837–1846
71 James, S.A. *et al.* (2009) Repetitive sequence variation and dynamics in the ribosomal DNA array of *Saccharomyces cerevisiae* as revealed by whole-genome resequencing. *Genome Res.* 19, 626–635
72 Santos, S.R. *et al.* (2003) Molecular characterization of nuclear small subunit (18S)-rDNA pseudogenes in a symbiotic dinoflagellate (*Symbiodinium*, Dinophyta). *J. Eukaryot. Microbiol.* 50, 417–421
73 Schlöetterer, C. and Tautz, D. (1994) Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. *Curr. Biol.* 4, 777–783
74 Pawlowski, J. *et al.* (2011) Eukaryotic richness in the abyss: insights from Pyrotag sequencing. *PLoS ONE* 6, e18169
75 Dell'Anno, A. and Danovaro, R. (2005) Extracellular DNA plays a key role in deep-sea ecosystem functioning. *Science* 309, 2179
76 Binladen, J. *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2, e197
77 Hamady, M. *et al.* (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* 5, 235–237
78 Ivanova, N.V. *et al.* (2007) Universal primer cocktails for fish DNA barcoding. *Mol. Ecol.* 7, 544–548
79 Matzen da Silva, J. *et al.* (2011) Systematic and evolutionary insights derived from mtDNA COI barcode diversity in the Decapoda (Crustacea: Malacostraca). *PLoS ONE* 6, e19449
80 Feinstein, L.M. *et al.* (2009) Assessment of bias associated with incomplete extraction of microbial DNA from soil. *Appl. Environ. Microbiol.* 75, 5428–5433
81 Engelbrektson, A. *et al.* (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J.* 4, 642–647
82 Hong, S. *et al.* (2009) Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* 3, 1365–1373
83 Valkiunas, G. *et al.* (2008) A comparative analysis of microscopy and PCR-based detection methods for blood parasites. *J. Parasitol.* 94, 1395–1401
84 Desquesnes, M. and Dávila, A.M.R. (2002) Applications of PCR-based tools for detection and identification of animal trypanosomes: a review and perspectives. *Vet. Parasitol.* 109, 213–231
85 Huse, S.M. *et al.* (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8, R143
86 Gilles, A. *et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 12, 245
87 Porazinska, D.L. *et al.* (2010) Linking operational clustered taxonomic units (OCTUs) from parallel ultra sequencing (PUS) to nematode species. *Zootaxa* 2427, 55–63
88 Porazinska, D.L. *et al.* (2009) Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. *Mol. Ecol. Resour.* 10, 666–676
89 Pester, M. *et al.* (2010) A 'rare biosphere' microorganism contributes to sulfate reduction in a peatland. *ISME J.* 4, 1591–1602
90 Creer, S. (2010) Second-generation seequencing derived insights into the temporal biodiversity dynamics of freshwater protists. *Mol. Ecol.* 19, 2829–2831
91 Smyth, R.P. *et al.* (2010) Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene* 469, 45–51
92 Quince, C. *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Methods* 6, 639–641
93 Huse, S.M. *et al.* (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* 12, 1889–1898
94 Shapiro, L.H. *et al.* (2006) Molecular phylogeny of *Banza* (Orthoptera: Tettigoniidae), the endemic katydids of the Hawaiian Archipelago. *Mol. Phylogenet. Evol.* 41, 53–63
95 Pandey, R. *et al.* (2010) CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies. *BMC Res. Notes* 12, 182
96 Kumar, S. *et al.* (2011) CLOTU: an online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. *BMC Bioinform.* 12, 182
97 Sun, Y. *et al.* (2009) ESPRIT: estimating species richness uing large collections of 16S rRNA pyrosequences. *Nucleic Acids Res.* 37, e76
98 Huson, D.H. *et al.* (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560
99 Greenberg, J. (2009) Theoretical considerations of lifecycle modeling: an analysis of the Dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging Classification Q.* 47, 380–402
100 Hankeln, W. *et al.* (2010) MetaBar – a tool for consistent contextual data acquisition and standards compliant submission. *BMC Bioinform.* 11, 358