CellPress

From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity

Melania E. Cristescu

Department of Biology, McGill University, Montreal, QC H3A 1B1, Canada

DNA-based species identification, known as barcoding, transformed the traditional approach to the study of biodiversity science. The field is transitioning from barcoding individuals to metabarcoding communities. This revolution involves new sequencing technologies, bioinformatics pipelines, computational infrastructure, and experimental designs. In this dynamic genomics landscape, metabarcoding studies remain insular and biodiversity estimates depend on the particular methods used. In this opinion article, I discuss the need for a coordinated advancement of DNA-based species identification that integrates taxonomic and barcoding information. Such an approach would facilitate access to almost 3 centuries of taxonomic knowledge and 1 decade of building repository barcodes. Conservation projects are time sensitive, research funding is becoming restricted, and informed decisions depend on our ability to embrace integrative approaches to biodiversity science.

From barcoding single individuals to metabarcoding communities

Evolutionary and ecological studies often rely on our ability to identify the species involved in the process under investigation or our capacity to provide robust biodiversity estimates [1,2]. Managing the health of global ecosystems requires detailed inventories of species and a good understanding of the patterns and trends of biodiversity [3]. For approximately 3 centuries, the acquisition of biodiversity data was based on morphological characterization of plants and animals. The idea of identifying species on the basis of molecular markers emerged soon after the advent of molecular biology. Early methods involved the use of hybridization, restriction enzyme digestion, or other molecular probes [4,5]. DNA-based species identification was introduced by Arnot *et al.* [6] and was firmly advanced and standardized by Hebert *et al.* [7]. The simple idea of using a short DNA fragment as a barcode (see Glossary) for identifying species across the Metazoa has been both strongly embraced and vigorously scrutinized over the past decade [8,9]. Nevertheless, the efforts led by Paul Hebert, and supported by the Consortium for the Barcode of Life (CBoL; http://www.barcodeoflife.org/) resulted in a global enterprise that combined molecular tools with valuable but scarce taxonomic expertise [10,11]. Today, DNA barcodes are being used commonly to identify specimens and the approach has wide applications in biodiversity conservation, environmental management, invasion biology, the study of trophic interactions, and food safety [12–14]. Despite its inherent challenges, which stem mainly from the difficult front-end curation and verification of voucher specimens [15], this approach has attracted large amounts of funding, prompted numerous taxon-specific projects, and has been used to generate over three million barcode

Glossary

DNA barcode: a small piece of the genome (marker) found in a broad range of species. The standardized barcode for most animals is a fragment of the mitochondrial *COI* gene, the standardized barcode for plants is a fragment of the plastid gene ribulose 1,5-bisphosphate carboxylase gene (*rbcL*) combined with a fragment of the maturase (*matK*) gene, whereas the barcode for fungi is the nuclear internal transcribed spacer (ITS) of the ribosomal DNA. CBOL (http://www.barcodeoflife.org/) has standardized this method of species identification, and has developed the corresponding sequence reference database for these markers [10].

DNA barcoding: the identification of species using standardized DNA fragments. The ideal DNA barcoding procedure starts with well-curated voucher specimens deposited in natural history collections and ends with a unique sequence deposited in a public reference library of species identifiers that could be used to assign unknown sequences to known species [7,43].

Metabarcoding: a rapid method of high-throughput, DNA-based identification of multiple species from a complex and possibly degraded sample of eDNA or from mass collection of specimens. The metabarcoding approach is often applied to microbial communities, but can be also applied to meiofauna or even megafauna.

Operational taxonomic unit (OTU): the taxonomic level selected to be used in a study, such as individuals or bacterial strains, populations, species, or genera [44,45].

Taxonomy: the science of discovering, describing, classifying, and naming organisms [36].

Corresponding author: Cristescu, M.E. (melania.cristescu@mcgill.ca).

Keywords: sequence biodiversity; DNA barcoding; metabarcoding; environmental DNA (eDNA); second-generation sequencing; taxonomy crisis.

0169-5347/

 http://dx.doi.org/10.1016/j.tree.2014.08.001

records in the Barcode of Life Database (BOLD) [10]. More recently, the technical advancements provided by the genomic revolution have enabled more direct evaluation of biodiversity compared with screening one specimen at a time. Metabarcoding extends DNA-based species identification to communities of individuals belonging to many groups of species with distinct roles in the ecosystem [16]. This multispecies identification method uses massive parallel sequencing of bulk samples (total DNA) or potentially degraded DNA from environmental samples (eDNA) for which species identification is not practical [11,17]. This rapidly growing, high-throughput, and sensitive method is likely to generate an increase in the speed, accuracy, and resolution of species identification [12,16,18]. The significant decrease in the cost associated with sampling and sequencing bulk samples instead of individual specimens at a time has the potential to enable a global network of biodiversity surveillance and monitoring [17]. However, such a global effort would require highly standardized, international monitoring networks and integrated, multidisciplinary approaches that build on the traditional ecological and taxonomic knowledge while integrating state-of-the-art technologies enabled by the genomics revolution.

In this article, I provide perspectives on the most pressing challenges of the metabarcoding field by focusing on the problems that directly hinder our ability to extract species-level signals from a bulk sample in a reproducible, accurate, and comparable manner. Many of these challenges are well recognized, continue to receive critical attention, and stimulate new research directions. Less appreciated is the need to develop a strongly integrative research plan that would enable molecular ecologists to embrace emerging metagenomics tools, corroborate traditional approaches, and launch global biodiversity initiatives. I finish with a discussion on the major steps needed towards advancing global biodiversity monitoring programs.

A research agenda for metabarcoding

As with other rapid technological advancements, the metabarcoding approach faces challenges that can hinder our ability to produce robust, comparable biodiversity estimates (Box 1). Many of these problems stem from dependency on the intermediate PCR step, which enriches the DNA templates extracted from a bulk sample. This step generates amplification biases and contributes to errors that can influence biodiversity estimates [19]. These problems are further amplified by errors introduced by the second-generation sequencing platforms. Another set of challenges stems from the need to build appropriate bioinformatics tools [19] and infrastructure to accommodate robust algorithms and efficient pipelines for data analysis [20,21]. The sheer volume of data generated creates the need for appropriate, centralized storage. The processed data are sometimes deposited to the National Center for Biotechnology Information (NCBI) GenBank (http://www.ncbi.nlm.nih.gov/genbank/), or to the Dryad Digital Repository (http://datadryad.org). However, storing original data remains largely the responsibility of individual laboratories or genomic centers. Although the cost of sequencing continues to drop, the cost for data analyses and storage remains more or less constant [20]. Therefore, the large gap between sequencing and analytic capabilities is expected to grow.

Box 1. Essential steps in the metabarcoding approach

Sampling design
- Replicated sampling schemes that capture community diversity [19].

Experimental design
- Need for technical replicates, including independent extractions and PCR amplifications;
- Need for appropriate markers (Table 1, main text);
- Need for appropriate statistics with corrections for multiple hypothesis testing [22].

Validating pipelines for de-noising and clustering the reads into OTUs
- Using benchmarked algorithms for quality control, de-noising, chimera removal, and OTU picking;
- Using appropriate distance levels for defining species calibrated for the taxonomic groups studies, the marker that is sequenced, and the algorithm used [46,47].
- Robust method of taxonomic assignment and phylogenetic placement with confidence estimates at each taxonomic level.

Ensuring sound interpretation of data
- Validation against standard biodiversity censuses [16];
- Validation against independent markers [19].

Ensuring data transferability and comparability
- Robust OTU recognition system responsive to input from global users and enabling community validation and annotation [7,21]; this is particularly useful in 'taxonomy-free' groups, such as bacteria and fungi, as well as in other groups with difficult morphology-based taxonomy;
- Comprehensive reference DNA library based on voucher specimens that enables access to the Linnean taxonomic system [37].

Most urgent is the need to promote best practices for data analysis that can promote informed recommendations. Current metabarcoding studies provide biodiversity estimates that are highly dependent on the resolution of the marker used, the quality of the sequenced libraries, bioinformatics pipelines, and the parameters used. Moreover, the operational taxonomic units (OTUs) obtained are not easily reconcilable across sites or studies and inferences regarding species distribution are difficult to make. Estimates of biodiversity are also not directly transferable or comparable. Often, metabarcoding projects involve markers that do not overlap with the standardized barcodes used to build reference libraries derived from morphologically identified specimens. This generates a growing gap between morphological and DNA-based identification. For all these reasons, a coordinated global initiative for advancing biodiversity research is much needed. Such an initiative would improve data transferability, comparability, and interpretability and would prompt the emergence of a global biodiversity-monitoring program. Data generated by a global network of samples could help identify ecological and genomics drivers of diversification and extinction.

A framework for sampling, experimental design, and data integration

Owing to the relative high cost of second-generation sequencing, early metabarcoding projects were rarely replicated, were often descriptive, and focused mainly on the

exploration of targeted taxonomic groups [22]. With a significant drop in cost, more appropriate experimental designs with technical and biological replication and more robust statistical analyses can be employed. This opportunity generates increased appreciation for the utility of metabarcoding to address fundamental questions in ecology and evolution. Given that, in many respects, metabarcoding data are similar to species–sample matrices used in ecology, many existing tools applied to identify correlations and statistically significant patterns are transferable [22]. The recovered OTUs can be also analyzed using traditional phylogenetic inferences to provide taxonomic assignment and phylogenetic placement with confidence estimates at each taxonomic level [23–25].

The need for evaluating standardized barcodes

Evaluating the performance of various DNA barcodes has been one of the main challenges of the early barcoding initiative [26]. For many targeted taxonomic groups (e.g., plants or fungi) the use of the standardized animal barcode, mitochondrial cytochrome c oxidase subunit I (*COI*), was not practical [26]. The problem of developing standardized barcodes is more severe for metabarcoding approaches that require the amplification of shorter DNA fragments, which are more appropriate for studies that target degraded eDNA, and well adapted for spanning different, often divergent, taxonomic groups [18,27]. Therefore, the metabarcoding primers have to be versatile enough to amplify equally and exhaustively different targeted groups (Table 1). Moreover, the amplified fragment has to have good taxonomic resolution and be discriminative, ideally to the species level. The barcode should have a well-understood pattern of molecular evolution to enable robust alignment and clustering. An efficient barcode must have also a comprehensive taxonomic reference database, generated with the use of the traditional barcoding approach, and based on morphologically verified and curated specimens. Given that building high-quality reference libraries is time and resource consuming, it has been suggested that metabarcodes should be designed within the standardized barcodes [11]. This would enable the validation of OTUs generated by metabarcoding pipelines against reference sequences associated with taxonomic information. The ideal scenario is to use multiple, evolutionary independent metabarcodes, one of which coincides with the standardized barcode of the particular taxonomic group. So far, a productive evaluation of the performance of various metabarcodes has been hindered by a shortage of comparable data from multiple candidate loci on a common set of samples [28].

Evaluating the pipeline of data processing

Another problem has been the lack of rigorous frameworks for data pre-processing ('de-noising'), processing raw reads (removal of primers, sequencing adaptors, barcode tags), performing analyses (clustering, BLASTing), and interpreting against standard biodiversity censuses [19,27,29]. A large number of dedicated software packages are available to assist with data processing and analyses [30], prompting the need for comprehensive, comparative performance tests against each other and against standard biodiversity data sets [31]. This array of continually emerging packages can intimidate not only ecologists entering the genomics field, but also bioinformaticians tasked to identify which method best fits a particular data set and which provides the most accurate estimation of biodiversity [19,20]. The outcome (number of OTUs and their taxonomic and ecological kinship) that various packages of filtering, clustering, or BLASTing algorithms generate can be divergent, making comparisons and broad conclusions difficult to achieve. Therefore, comprehensive comparative studies of OTU generation methods tested on

Table 1. Criteria for barcode evaluation

Criteria for evaluation	Barcoding	Metabarcoding
Size	Sizes usually longer than 500 bp	Sizes <400 bp are appropriate for degraded DNA
Specificity	At the taxon level	Specific across a divergent group of targeted taxa, but not beyond
		Broad application of single primer pair beyond targeted groups compromises depth of coverage
		Multiple primer pairs can be employed when amplification bias across divergent taxonomic groups is severe. Each targeted group can be amplified by a specific primer-pair.
Versatility	Extensive versatility beyond the taxon of interest is not essential, but can enhance projects charged with comprehensive coverage of large taxonomic groups	High versatility to amplify equally and exhaustively all target groups
Taxonomic resolution	Taxonomic resolution at the species level is desirable	Taxonomic resolution, ideally to the species level, is required; requires validation based on mock communities or similar methods
Well-understood mode of evolution	A distinct break between the intra- and interspecific levels of genetic divergence is required	Desirable for enabling good global alignments that allow valid recovery of OTUs
		Knowledge on the intra- and interspecific levels of genetic divergence across the targeted groups is required
Comprehensive taxonomic database	Building a comprehensive database is a major goal of the barcoding approach	Comprehensive taxonomic database based on verified and curated specimens is desirable; many metabarcodes used currently do not have an associated taxonomic database

standard set of data sets are needed, and best practices in quality filtering, clustering, and defining OTUs need to be established. Several metabarcoding studies conducted on microbial communities have reported much higher diversity of rare species than expected based on classical morphology [32]. Further verifications have suggested that these high estimates are due to a combination of sequencing errors and inadequate data processing [33–35]. Collectively these studies promote rigorous analysis of metabarcoding data sets, including data de-noising and sequence clustering with parameters adjusted to the particular dataset. One of the main goals of metabarcoding projects is to cluster sequences into OTUs that correspond to ecological species, unique in their particular niche, or biological species as unique reproductive pools. The reluctance to assign species-level designation based on DNA sequences is in part a consequence of the limited amount of reference sequences with taxonomic validity presently available for many groups of organisms. The parallel advancement of barcoding and metabarcoding approaches facilitates species-level assignment and allows us to investigate how genotypic and phenotypic diversity relate.

Advancing data transferability and interpretability

There is an increased need to attach meaning to the growing amount of sequence information entering databases and to link OTUs with ecology or particular physiological traits [21,24]. Assigning OTUs to the species level enables us to monitor changes in community composition at a fine taxonomic scale and identify the early stages of a disturbance. It also enables us to extend metabarcoding as a tool for early detection of invasive species. The ability to discriminate OTUs at the species level is essential for such applications. Most importantly, a robust taxonomic inference helps us calibrate biodiversity estimates and prevent erroneous interpretations. The consequence of overestimating the genetic diversity of challenged environments can have detrimental managerial consequences. By contrast, underestimation resulting from pooling divergent sequences derived from related species into one OTU can hinder our ability to discover closely related species, low abundance species with key roles in the ecosystem, or early invasions. Therefore, cross-validation using multiple markers and mock communities with well-understood species composition and relevant complexity can help mitigate these problems.

Unfortunately, taxonomic information based on morphology is only sporadically available for many groups of organisms that are minute (meio- and micro-fauna and flora) and difficult to inspect visually [24]. However, the absence of taxonomy can still enable researchers to conduct basic ecological analyses and address relevant ecological questions [19]. For 'taxonomy-free' groups, such as bacteria and fungi, as well as for other groups with scarce taxonomic information, researchers need a robust way of surveying the taxonomic level, phylogenetic position, and ecological role of a particular OTU. Is a particular OTU frequently encountered across samples or studies? Is its geographic distribution widespread or limited? Does it represent a native or invasive species? Identifying a particular OTU across studies can be facilitated by implementing a robust OTU recognition system [21]. A global identification system would enable the scientific community to close the growing gap between taxonomy and ecology [36] by allowing researchers to capitalize on almost 3 centuries of taxonomic research and one decade of building repository barcodes. Currently, most metabarcoding projects use markers that are not standardized and are not associated with a reference DNA library. Such projects do not allow easy association with the Linnean taxonomic system. Molecular markers often reveal taxa that have not been described or require further taxonomic evaluation. An efficient recognition system should be open to annotation based on morphological, physiological, or behavioral traits and validation based on other molecular markers.

DNA barcoding and metabarcoding as highly complementary approaches

Metabarcoding has been described as a promising technique that is rendering the more traditional DNA barcoding irrelevant [12]. In this section, I argue that barcoding and metabarcoding are complementary techniques and that the future of biodiversity studies would benefit from a harmonizing approach to biodiversity research. The two techniques are similar in that they both use DNA-based identification of species, but they have divergent assets that are determined by their distinct sequencing technologies and specific goals. Whereas DNA barcoding involves sequencing one well-curated individual at a time, metabarcoding entails massive parallel sequencing of complex bulk samples for which morphological identification and curation is not practical. The metabarcoding technique has the potential to capitalize on the enormous advantage offered by second-generation sequencing technology, capable of generating millions of sequences in one run. However, our ability to interpret the results depends not only on further bioinformatical refinements, but also on the availability of well-populated databases that contain reference sequences of the taxa of interest [16,37]. The DNA barcoding approach has been designed to build essential bridges between molecular ecologists and morphological taxonomists by generating reference databases based on verified and curated specimens [38,39]. This basic alliance enables molecular ecologists to access the Linnean taxonomic system. Fortunately, this association can be extended to the metabarcoding approach by designing metabarcodes within the standardized barcodes. The taxonomic information associated with the standard barcodes would allow researchers to place the OTUs within meaningful ecological, physiological, and evolutionary contexts. With the continued generation of barcoded and curated specimens of museum collections, we will be able to link reference sequences to the enormous functional biological knowledge [16,37].

The roadmap for the future of sequencing-based biodiversity analysis is hard to predict but few directions transpire. Single-marker sequencing will be likely replaced by multi-marker approaches. It is easy to imagine that eventually, specific markers for taxon identification will be substituted by whole genome information. Since genomes will be fragmented in pools of scattered genomes, bioinformatic tools will have to be refined to allow for comparing

homologous genome fragments across divergent taxa. Single-molecule sequencing technologies will likely eliminate the need of PCR amplification step and will eradicate an entire suite of PCR-induced errors. Reference libraries will probably not be constructed on a genome assembly, but rather a series of sequences derived from representative genomes. Although the future holds considerable technological advancements, the present demands responsible decisions that maximize our chance of sustaining future endeavors.

Concluding remarks: The time-sensitive nature of the biodiversity decline requires coordinated effort

Human ingenuity and fascination with technological advancements have prompted progress and have radically transformed cultural, economic, and physical landscapes. We became global environmental drivers [40], a role that comes with great responsibility and a need for foresight. Our future as a species depends on our ability to find the fine balance between what we want and what nature can provide. A direct way to evaluate our environmental impact is by documenting the biological richness of our planet and monitoring biodiversity trends at a global level [41]. Metabarcoding is emerging as a promising way of scrutinizing the health of our ecosystems and advancing biodiversity research [18]. However, generating reliable, verifiable, and easily interpretable biodiversity estimates requires robust, well-probed methods and generally accepted standards. It also requires a global, integrative, and interdisciplinary program for biodiversity research (Box 2, Figure 1). Such a program would enable us to capitalize on the efforts of the International Barcode of Life (IBOL; http://ibol.org/) to build reference libraries, and would facilitate access to taxonomic knowledge acquired over the past 3 centuries. Much can be learned from the early phase of the IBOL program about ways of promoting standardization, reproducibility, and reusability in barcoding approaches. The potential consequences of erroneous biodiversity estimates and poor taxonomy are commonly underestimated, but might have serious ecological and economic implications [16,36,42]. Developing sustainable action plans requires a solid understanding of the patterns, trends, origins, and functions of biodiversity, as well as the underlying drivers [3]. This can only be achieved through global, coordinated efforts that integrate traditional

Box 2. Steps towards enabling a global biodiversity program

Major steps

- Establish museums of genomes and metagenomes (biodiversity biorepositories) equipped with a portal for sample collections around the world and developed standards for DNA sharing [48];
- Establish global sampling networks for terrestrial and aquatic life [49];
- Promote targeted sampling and preservation of samples from representative regions and sites;
- Adopt sets of markers, design sets of primers suitable for metabarcoding approaches, and link to reference libraries derived from morphologically identified specimens;
- Continue to develop barcode reference libraries for standardized barcodes;
- Develop comprehensive genomics analyses and promote best practices for data analysis that can promote informed recommendations;
- Develop an infrastructure for metabarcoding data deposition and OTU recognition open to annotation and cross-validation;
- Build partnership with experts from different geographical regions and with various taxonomic foci;
- Inform negotiations leading to international agreements.

Potential long-term outcomes

- Provide robust catalogs of animal and plant life and solid local or regional biodiversity estimates;
- Enable fundamental ecological and evolutionary research by providing essential information on community assembly, trophic linkage, energy flow, distribution, and origin of diversity;
- Couple measures of diversity with spatial, ecological, and climatic information;
- Quantify status and trends in global biodiversity;
- Identify drivers of biodiversity change;
- Direct conservation efforts effectively, and inform remediation and intervention actions.
- Facilitate a broad range of applications in invasion biology, food safety and human health.

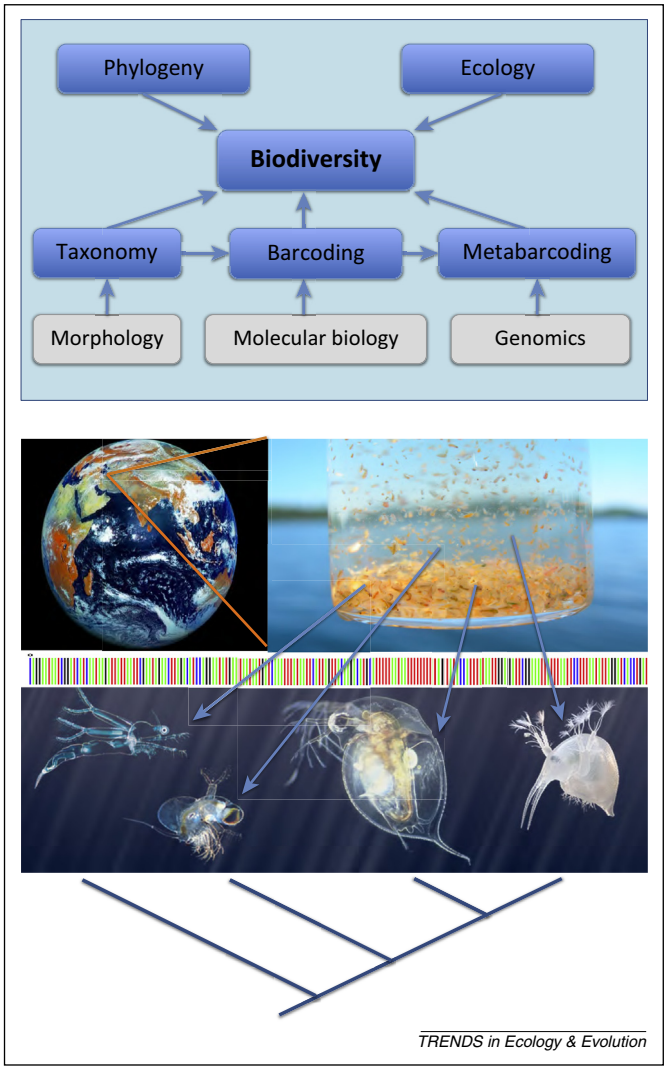

Figure 1. An integrative approach to advance biodiversity research in the genomic era. Closing the gap between traditional approaches derived from ecology, taxonomy, and phylogeny and fast-evolving molecular techniques would enable researchers to capitalize on almost 3 centuries of taxonomic research and one decade of building repository barcodes.

approaches and effectively implement emerging technologies.

Acknowledgments

I would like to thank P. Hebert and all participants of the Planetary Biodiversity Project for inspiring discussions and the Canadian Institute for Advanced Research (CIFAR) for supporting the meeting. The work benefited also from insightful comments provided by three anonymous reviewers. I would also like to thank E.L. Clare, T. Crease, D. DeMill, E. Domaizon, S. Ratnasingham, and J. Witt for helpful comments; C. Abbott, E. Brown, F. Chain, J Flynn, H. MacIsaac, and G. Rolshausen provided stimulating discussions. This work was supported by a Natural Sciences and Engineering Research Council of Canada Discovery Grant and the Canada Research Chair program.

References

1 Brook, B.W. *et al.* (2008) Synergies among extinction drivers under global change. *Trends Ecol. Evol.* 23, 453–460
2 Barnosky, A.D. *et al.* (2011) Has the Earth's sixth mass extinction already arrived? *Nature* 471, 51–57
3 Larigauderie, A. *et al.* (2012) Biodiversity and ecosystem services science for a sustainable planet: the DIVERSITAS vision for 2012–20. *Curr. Opin. Environ. Sust.* 4, 101–105
4 Nanney, D.L. (1982) Genes and phenes in *Tetrahymena*. *BioScience* 32, 783–788
5 Gibson, W.C. *et al.* (1988) Species-specific DNA probes for the identification of African trypanosomes in tsetse flies. *Parasitology* 97, 63–73
6 Arnot, D.E. *et al.* (1993) Digital codes from hypervariable tandemly repeated DNA sequences in the *Plasmodium falciparum* circumsporozoite gene can genetically barcode isolates. *Mol. Biochem. Parasitol.* 61, 15–24
7 Hebert, P.D.N. *et al.* (2003) Biological identification through DNA barcodes. *Proc. Biol. Sci.* 270, 313–321
8 Blaxter, M. (2003) Molecular systematics – counting angels with DNA. *Nature* 421, 122–124
9 DeSalle, R. (2007) Phenetic and DNA taxonomy; a comment on Waugh. *Bioessays* 29, 1289–1290
10 Ratnasingham, S. and Hebert, P.D.N. (2007) BOLD: the barcode of life data system. *Mol. Ecol. Notes* 7, 355–364
11 Taberlet, P. *et al.* (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* 21, 2045–2050
12 Taylor, H.R. and Harris, W.E. (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Mol. Ecol. Res.* 12, 377–388
13 Valentini, A. *et al.* (2009) DNA barcoding for ecologists. *Trends Ecol. Evol.* 24, 110–117
14 Clare, E.L. *et al.* (2014) Trophic niche flexibility in *Glossophaga soricina*: how a nectar seeker sneaks an insect snack. *Funct. Ecol.* 28, 632–641
15 Borisenko, A.V. *et al.* (2009) The front-end logistics of DNA barcoding: challenges and prospects. *Mol. Ecol. Res.* 9, 27–34
16 Ji, Y. *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.* 16, 1245–1257
17 Bohmann, K. *et al.* (2014) Environmental DNA for wildlife biology and biodiversity. *Trends Ecol. Evol.* 29, 358–367
18 Gibson, J. *et al.* (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8007–8012
19 Bik, H.M. *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends Ecol. Evol.* 27, 233–243
20 McPherson, J.D. (2009) Next-generation gap. *Nat. Methods* 6, S2–S5
21 Ratnasingham, S. and Hebert, P.D.N. (2013) A DNA-based registry for all animal species: the Barcode Index Number (BIN) System. *PLoS ONE* 8, e66213
22 Thomas, T. *et al.* (2012) Metagenomics – a guide from sampling to data analysis. *Microb. Inform. Exp.* 2, 3
23 Matsen, F.A. *et al.* (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11, 538
24 Blaxter, M. *et al.* (2005) Defining operational taxonomic units using DNA barcode data. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 360, 1935–1943
25 Taib, N. *et al.* (2013) Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: new insights into the freshwater protist diversity. *PLoS ONE* 8, e58950
26 Hollingsworth, M.L. *et al.* (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Mol. Ecol. Res.* 9, 439–457
27 Coissac, E. *et al.* (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Mol. Ecol.* 21, 1834–1847
28 Tang, C.Q. *et al.* (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16208–16212
29 Bonder, M.J. *et al.* (2012) Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* 28, 2891–2897
30 Caporaso, J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336
31 Edgar, R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998
32 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
33 Kunin, V. *et al.* (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* 12, 118–123
34 Bachy, C. *et al.* (2013) Accuracy of protist diversity assessments: morphology compared with cloning and direct pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid ciliates as a case study. *ISME J.* 7, 244–255
35 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
36 Wägele, H. *et al.* (2011) The taxonomist – an endangered race. A practical proposal for its survival. *Front. Zool.* 8, 25
37 Janzen, D.H. *et al.* (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. *Proc. R. Soc. Lond. B* 360, 1835–1845
38 Hebert, P.D.N. and Barrett, R.D.H. (2005) Reply to the comment by L. Prendini on 'Identifying spiders through DNA barcodes'. *Can. J. Zool.* 83, 505–506
39 Hebert, P.D.N. and Gregory, R.T. (2005) The promise of DNA barcoding for taxonomy. *Syst. Biol.* 54, 852–859
40 Barnosky, A.D. *et al.* (2012) Approaching a state in Earth's biosphere. *Nature* 486, 52–58
41 Bottrill, M.C. *et al.* (2008) Is conservation triage just smart decision making? *Trends Ecol. Evol.* 23, 649–654
42 Goldstein, P.Z. and DeSalle, R. (2010) Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. *Bioessays* 33, 135–147
43 Floyd, R. *et al.* (2002) Molecular barcodes for soil nematode identification. *Mol. Ecol.* 11, 839–850
44 Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy. The Principles And Practice Of Numerical Classification*, W.H. Freeman
45 Wooley, J.C. *et al.* (2010) A primer on metagenomics. *PLoS Comput. Biol.* 6, e1000667
46 Sun, Y. *et al.* (2012) A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief. Bioinform.* 13, 107–121
47 Hebert, P.D.N. *et al.* (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. B: Biol. Sci.* 270, 96–99
48 Droege, G. *et al.* (2013) The global genome biodiversity network (GGBN) data portal. *Nucleic Acids Res.* 42, D607–D612
49 Henry, P–Y. *et al.* (2008) Integrating ongoing biodiversity monitoring: potential benefits and methods. *Biodivers. Conserv.* 17, 3357–3382