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