Cell PRESS

The golden age of DNA metasystematics

Mehrdad Hajibabaei

Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, Ontario N1G 2W1, Canada

The convergence of next-generation sequencing and DNA barcoding has sparked a golden age of 'DNA metasystematics', allowing researchers to understand the biodiversity of an entire ecosystem solely through DNA information, and transforming the way we view the living world around us.

As a molecular evolutionary geneticist I feel I am experiencing a golden age of DNA metasystematics, an era that started almost 10 years ago but which has multiple roots that go back for centuries. Advances in data generation and analysis – especially from DNA and genomes – have arguably transformed the biological sciences in the past 30 years, but many of these advances have been incremental and foreseeable, owing to the omnipresence of powerful computers and market-driven endeavors to develop cheaper and more advanced molecular biology procedures. However, I believe that two paradigm shifts in the past decade have led to the blossoming of a new form of molecular systematics, which brings with it a wide range of applications. Below, I describe these two advances and discuss how their convergence is impacting upon biodiversity research and changing our understanding of ecosystems.

The first paradigm shift occurred in 2003 as a result of the proposal to use information embedded in standardized DNA sequences, termed 'DNA barcodes', for species identification as a tool to circumvent limitations of traditional taxonomic methods (Box 1) [1]. Although establishing standardized DNA sequences as barcodes was a significant advancement, the true paradigm shift was the transformation of the taxonomic method of inquiry for a wide range of organisms [2]. This transformation has occurred gradually since 2003, with some groups of organisms still waiting for their designation of a 'standard' DNA barcode. Furthermore, standards may change over time, evolving to use additional DNA information to resolve difficult taxonomic puzzles. Nevertheless, taxonomic research and routine identification for nearly all taxa, from microbes to mammals, is on an irreversible course towards DNA-based solutions. The existence of a well-organized and enthusiastic movement of scientific and other user communities (e.g., governmental agencies and non-governmental organizations) that is championing the adoption and advancement of DNA barcoding [2–4] is strong evidence of this paradigm shift.

The second paradigm shift was brought on by the revolutionary advancements in DNA sequencing technologies since 2005. Traditional Sanger sequencing, invented in 1977, led to a breakthrough in genomics, providing the capacity to sequence whole genomes, including the human genome. Although technological advances increased the speed and lowered the labor-intensiveness of DNA sequencing, at its core the Sanger-based method remained unchanged until the emergence of next-generation sequencing (NGS) technologies in the mid-2000s. There is now a steady stream of advances that allow DNA to be sequenced at ever-increasing volumes and speeds while simultaneously driving the cost per sequenced base lower and lower [5,6]. Although faster and cheaper DNA sequencing is of key importance for many applications, the massively parallel capacity of NGS machines is at least equally important for environmental applications, which require the identification and tracking of multiple organisms in a habitat simultaneously.

Because methods of inquiry in systematic biology and taxonomy (and, in the broader sense, comparative biology) have traditionally focused on studying single specimens or a cloned population of the same species/strain (for microbes), the ability to study the taxonomic composition and phylogenetic relationships of communities through comparative analysis of mixtures of DNA is a major step forward. The study of pooled DNA from uncultured microorganisms began before the advent of NGS, and involved cloning target sequences (from a community) and then sequencing those clones one by one [7]. However, NGS has dramatically changed the research landscape by allowing the analysis of DNA/RNA pools from communities, thereby abrogating the need to clone and individually analyze members of those communities [8]. As a result, biodiversity-related applications such as the human microbiome analysis [9] blossomed after the introduction of NGS technologies. The NGS paradigm shift accelerated the work of many scientists who were stuck in a sample processing and identification bottleneck due to the difficulty of separating and studying individual organisms one at a time. Current biodiversity-related applications of NGS go beyond microbial groups and involve a wide range of investigations from diet analysis, to studying ancient biota, to the rapid biomonitoring of different habitats [10].

The power of these two major advances for biosystematics investigations has already been demonstrated by many exemplar applications [10]. However, the potential for even more profound socioeconomic utility is now becoming a reality as researchers link the concept of DNA barcodes to the power of NGS technologies to perform community analyses. This area of research has already been referred to by at least four terminologies: microbial ecologists refer to this concept as a branch of 'metagenomics' through

Corresponding author: Hajibabaei, M. (mhajibab@uoguelph.ca).
Keywords: metagenomics; DNA barcoding; phylogenetics.

Trends in Genetics November 2012, Vol. 28, No. 11

Box 1. Exploring biodiversity through DNA barcodes

For over 30 years, DNA variation has been used to identify species and reconstruct evolutionary histories. Many genes or non-coding segments of the genome have been used for biosystematics. For example, the ribosomal gene 16S rDNA is a widely used marker for biodiversity analysis in prokaryotes. DNA barcoding uses a small, standardized region of the genome, referred to as a DNA barcode, for broad-scale identification of species in different taxonomic groups [1,3]. The originally proposed DNA barcode, a ~650 bp-long fragment of the mitochondrial *cytochrome c oxidase 1* gene (*COI*), can distinguish more than 95% of animal species [3]. The quest to establish a standard DNA barcode for plants and fungi took several years. Although partial sequences of two genes (*rbcL* and *matK*) from the chloroplast genome have been designated as the core plant DNA barcode, in many cases additional sequences are required to distinguish between plant species [4]. The situation is similar for fungi, where use of the *internal transcribed spacer* (*ITS*), a nuclear ribosomal sequence that has been recently designated as a DNA barcode, is usually supplemented with information from additional sequences [15]. The vast diversity and long evolutionary history of protists will necessitate the employment of different genes for DNA barcoding. A single gene would not suffice for proper species identification, but ribosomal markers such as 18S rDNA could provide broad insights into protist diversity.

Once a reference DNA barcode library from well-characterized species is assembled, it is possible to identify newly obtained or preserved (e.g., museum) specimens by comparing their DNA barcodes to those in the reference library [3]. Global or specialized libraries of DNA barcodes (and their metadata) are essentially DNA-based maps of biodiversity. DNA barcode reference libraries for various applications (e.g., pests, pathogens, and disease vectors, bioindicator species) are now being established, and DNA barcoding can routinely be used to aid taxonomic identifications. Leveraging the massively parallel next-generation sequencing (NGS) technology, DNA barcodes can now be generated from bulk environmental samples such as air, water, soil, sediments, and gut contents [10]. DNA barcode information gained through NGS allows biodiversity exploration on a much larger spatiotemporal scale and is fundamental to a DNA metasystematics framework for community analysis.

marker gene(s) analysis [8]; researchers who study marine meiofaunal and eukaryotic microbial communities have termed their approach 'metagenetics' [11]; a spin-off of the DNA barcoding initiative has termed their approach 'environmental barcoding' [12]; and finally, a group of molecular ecologists have termed their approach 'DNA metabarcoding' [13]. These approaches share two common elements: (i) DNA sequences are the main source of taxonomic (or functional diversity) information; and (ii) DNA sequences are gathered from communities through bulk/environmental samples containing mixtures of DNA (or environmental DNA).

Although similar, there are key differences among the above-mentioned approaches. For example, environmental barcoding focuses on obtaining standard DNA barcode sequence information to gain species-level resolution through comparisons to a reference DNA barcode library. This is somewhat similar to marker gene metagenomics analyses of bacterial 16S rDNA, which is linked to a vast reference library. By contrast, the DNA metabarcoding approach employs short markers that may not provide species-level resolution or be linked to a standardized sequence library. Finally, the metagenetics approach has focused on a single marker, 18S rDNA, because of its relative ease of amplification using available primers. This marker can be utilized to gain biodiversity insights from a large number of eukaryotic taxa, but it suffers from sequence heterogeneity within individuals' genomes and is not considered a good species-level DNA barcode [14].

Despite such differences, these four techniques can be united under a common framework: 'DNA metasystematics'. Within this framework, genetic information is obtained directly from environmental samples to understand biodiversity at different levels of organization and taxonomic groups. Metasystematics studies can provide species-level biodiversity information through the use of standardized DNA barcode markers that are linked to reference sequence libraries, voucher specimens and, ultimately, all available taxonomic knowledge gained through past investigations. Different DNA markers can provide deeper phylogenetic diversity or functional diversity measures (i.e., genes involved in various pathways). This common framework will foster collaboration and cross-disciplinary studies and the development of new applications.

Historically, golden ages reflected periods in which important tasks have been accomplished in a specific endeavor. I have outlined how the convergence of DNA-based identifications with NGS technologies has led to a new suite of methods – that I group under the umbrella term 'DNA metasystematics' – that will revolutionize our ability to measure, study, and understand complex communities of organisms. The impact of these advancements will be felt across many disciplines, from human health to environmental science, and will transform the way we view the living world around us.

Acknowledgments

I thank Donal Hickey (Concordia University), Dan Janzen (University of Pennsylvania), Ian King (University of Guelph), and Greg Singer (Ryerson University) for valuable comments on an earlier version of this manuscript. My research is funded by the Government of Canada through Genome Canada, the Ontario Genomics Institute, and the Natural Sciences and Engineering Research Council.

References

1 Hebert, P.D.N. *et al.* (2003) Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. Ser. B: Biol. Sci.* 270, 313–321
2 Marshall, E. (2005) Taxonomy. Will DNA bar codes breathe life into classification? *Science* 307, 1037
3 Hajibabaei, M. *et al.* (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet.* 23, 167–172
4 Hollingsworth, P.M. *et al.* (2009) A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U. S. A.* 106, 12794–12797
5 Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141
6 Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380
7 Handelsman, J. *et al.* (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, R245–R249
8 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
9 Ley, R.E. *et al.* (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022–1023
10 Shokralla, S. *et al.* (2012) Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* 21, 1794–1805
11 Creer, S. *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Mol. Ecol.* 19 (Suppl. 1), 4–20

12 Hajibabaei, M. *et al.* (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE* 6, e17497
13 Taberlet, P. *et al.* (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* 21, 2045–2050
14 Creer, S. and Sinniger, F. (2012) Cosmopolitanism of microbial eukaryotes in the global deep seas. *Mol. Ecol.* 21, 1033–1035
15 Schoch, C.L. *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U. S. A.* 109, 6241–6246

0168-9525/$ – see front matter
http://dx.doi.org/10.1016/j.tig.2012.08.001 Trends in Genetics,
November 2012, Vol. 28, No. 11