DNA barcoding, an effective tool for species identification: a review

Sandeep Antil[1] · Jeeva Susan Abraham[1] · S. Sripoorna[1] · Swati Maurya[1] · Jyoti Dagar[1] · Seema Makhija[1] · Pooja Bhagat[1] · Renu Gupta[2] · Utkarsh Sood[3] · Rup Lal[3] · Ravi Toteja[1]

Abstract

DNA barcoding is a powerful taxonomic tool to identify and discover species. DNA barcoding utilizes one or more standardized short DNA regions for taxon identification. With the emergence of new sequencing techniques, such as Next-generation sequencing (NGS), ONT MinION nanopore sequencing, and Pac Bio sequencing, DNA barcoding has become more accurate, fast, and reliable. Rapid species identification by DNA barcodes has been used in a variety of fields, including forensic science, control of the food supply chain, and disease understanding. The Consortium for Barcode of Life (CBOL) presents various working groups to identify the universal barcode gene, such as COI in metazoans; rbcL, matK, and ITS in plants; ITS in fungi; 16S rRNA gene in bacteria and archaea, and creating a reference DNA barcode library. In this article, an attempt has been made to analyze the various proposed DNA barcode for different organisms, strengths & limitations, recent advancements in DNA barcoding, and methods to speed up the DNA barcode reference library construction. This study concludes that constructing a reference library with high species coverage would be a major step toward identifying species by DNA barcodes. This can be achieved in a short period of time by using advanced sequencing and data analysis methods.

Keywords Biodiversity · DNA barcoding · Next-generation sequencing (NGS) · Oxford Nanopore Technologies (ONT)'s MinION™ · PacBio sequencing · Consortium for barcode of life (CBOL)

Introduction

The existence of life is one of the most unique aspects of Earth, and the diversity of life is the most astonishing feature of life. Biological diversity refers to the variation among living organisms from all sources comprising terrestrial, marine, and other aquatic ecosystems and ecological complexes of which they are a part; it encompasses diversity within species, between species, and within ecosystems [1]. Biodiversity plays a key role in maintaining

✉ Ravi Toteja
ravitoteja@andc.du.ac.in

Sandeep Antil
sandeep@andc.du.ac.in

Jeeva Susan Abraham
jeevasa@andc.du.ac.in

S. Sripoorna
sripoorna@andc.du.ac.in

Swati Maurya
swatimaurya@andc.du.ac.in

Jyoti Dagar
jyotidagar@andc.du.ac.in

Seema Makhija
seemamakhija@andc.du.ac.in

Pooja Bhagat
poojabhagat@andc.du.ac.in

Renu Gupta
rgupta17@maitreyi.du.ac.in

Utkarsh Sood
c_utkarsh.sood@teri.res.in

Rup Lal
rup.lal@teri.res.in

1   Acharya Narendra Dev College, University of Delhi, New Delhi, Delhi, India

2   Maitreyi College, University of Delhi, New Delhi, Delhi 110 021, India

3   The Energy and Resources Institute, IHC Complex, New Delhi 110003, India

ecological balance [2]. The diversity of species described so far is a very tiny portion of biodiversity present on the earth (approximately 9 million). This means that it is very hard to estimate the diversity of life that go extinct every day, as scientists have only described 10–15% of the total diversity of the earth [3]. A considerable portion around 86–91% (~ 7.2 million) of diversity remains undescribed due to several reasons such as scarcity of funds for taxonomy, a very less number of the trained taxonomist, and absence of accurate species identification methods. For describing the remaining diversity an accurate method for taxonomic identification, trained taxonomist, and funding are required. Traditionally, taxonomic assessment has relied on the basis of morphological character which is time-consuming, requires taxonomic specialists, and gives false identification when cryptic species and phenotypic plasticity are concerned [4, 5]. Furthermore, globally, the number of traditional taxonomists is declining [6]. Therefore, the majority of the diversity of microorganisms and invertebrates may have to be distinguished solely by DNA-based molecular techniques, without accompanying live cultures or physical specimens. These molecular techniques have several advantages over traditional approaches. Because, molecular tools are standardized tools which allow direct comparison among different users and they do not require taxonomic expertise, can be applied to environmental samples which comprise a mix of several species, like soil or a water sample and can be used in early warning allowing detection of low concentration of potential invaders, or even imprints of potential invader [7].

Over the last decade, DNA barcoding emerges as a new molecular tool for taxonomists to identify species. DNA barcoding utilizes one or more standardized short genetic markers in an organism's DNA to recognize it as belonging to a particular species, and through this strategy, DNA sample from the unidentified species is compared to identified sequences present in a DNA barcode reference library, developed by Hebert and his collaborators [8]. DNA barcoding is based on the principle of barcoding gap that refers to the difference between mean intra- and interspecific genetic distances. The wider the barcoding gap is, the more reliable species discrimination will be achieved. DNA barcoding is budget-friendly, less time-consuming, objective method, and a powerful tool for species identification when cryptic species and phenotypic plasticity is a concern or morphology keys are not available [9]. Due to the better precision and ease of DNA barcoding, this technique is gaining popularity, and it can be used to identify species in any stage of life (i.e. both adults and immature stage including eggs). DNA barcoding mostly differs from other molecular tools by use of standard markers, such as COI in metazoans; rbcL, matK, and ITS in plants; ITS in fungi; 16S rRNA gene in bacteria and archaea [7]. For DNA barcoding, the selection of the barcoding gene is crucial. A barcoding gene must satisfy three criteria (1) A distinct 'barcoding gap' between maximum intra-specific and inter-specific divergence within a group of organisms; (2) conserved flanking sites for creating universal PCR primers; (3) short sequence length to facilitate current capabilities of DNA extraction and amplification [8].

This technique involves a collection of a sample from the field, extracting DNA, selection of barcoding gene for amplification by using a universal primer (Table 1), amplified DNA molecule is sequenced by Sanger sequencing or High-throughput sequencing for assessing the diversity and analysis of obtained data by using data analysis software such as Mothur, Qiime2 etc. (Fig. 1) [10, 11].

The Consortium for the Barcode of Life (CBOL, https://www.ibol.org) is an international organisation that was founded in 2004 to facilitate the establishment and use of DNA barcodes as a global standard for biological species identification. CBOL includes various group, such as Plant working group for plants, Protist working group for eukaryotic microorganism, Fungal working group for fungi, to identify the universal barcode gene and creating a reference DNA barcode library [12]. A reference database, Barcode of life data system (BOLD, http://www.boldsystems.org), has been developed that aids in acquiring, storage, analysis, and publication of DNA barcode and allows a significant number of species to be identified [13]. The present study aims to review the various proposed/available DNA barcodes the for animals, plants, fungi, bacteria, virus, and protists (more specifically ciliates). Over the period of time, significant advances have been made in DNA barcoding. One important advancement in barcoding by mitogenomics and nuclear ribosomal RNA repeats obtained by genome skimming. The same has been discussed in the present review along with the mitogenomics approach for species identification. In the end, strengths and limitations of the technique have also been briefly described.

Barcodes for identification of animals

Hebert et al. had suggested a 650 bp fragment of the mitochondrial cytochrome-c oxidase subunit 1 (COI) gene as a universal marker or 'DNA barcode' for global biological identification of animal species. COI gene is a mitochondrial gene that is highly conserved [14], codes for respiratory electron transport chain protein that reduce molecular oxygen into water, present in all aerobic organisms. Mitochondrial genes are preferred over nuclear genes because mitochondrial genes are generally haploid, lack introns, and contain limited recombination. Mitochondria reproduce by binary fission and without sexual recombination, so the mitochondrial genes are subjected less to insertions, deletions or other large-scale rearrangements that introduce

Table 1 Represents list of barcoding gene for species identification and their primer pairs

Organisms	Barcode gene	Primers (5′–3′)	References
Animals	COI	F-GGTCAACAAATCATAAAGATATTGG R-TAAACTTCAGGGTGACCAAAAAATCA	[15]
	COII	F-TCACWATACTAYTAACAGATCGCAA R-AATAGCTGTACAGTGGGT	[95]
Plant	COI	F-CAACATTT ATTTTGATTTTTTGG R-TCCAATGCACTAATCTGCCATATTA	[16]
	RbcL	F-ATGTCACCACAAACAGAAAC R-TCGCATGTACCTGCA	[19]
	MatK	F-CCTCATCTGGAAATCTTGGTT R-GCTTATAATGAGAAAGATTTCTGC	[19]
	PsbA-trnH	F-CGCGCATGGTGGATTCACAATCC R-GTTATGCATGAACGTAATGCTC	[19]
	rpoC1	F-CCSATTGTATGGGAAATACTT R-CTTACAAACTAATGGATGTAA	[96]
	rpoB	F1-AAGTGCATTGTTGGAACTGG F2-ATGCAACGTCAAGCAGTTCC R1-CCGTATGTGAAAAGAAGTATA R2-GATCCCAGCATCACAATTCC	[19]
	atpF-atpH	F-ACTCGCACACACTCCCTTTCC R-GCTTTTATGGAAGCTTTAACAAT	[19]
	psbK-psbI	F-TTAGCCTTTGTTTGGCAAG R-AGAGTTTGAGAGTAAGCAT	[19]
	ITS	F-ATGCGATACTTGGTGTGAAT R-GACGCTTCTCCAGACTACAAT	[9]
Fungi	ITS	F-TCCTCCGCTTATTGATATGC R-GGAAGTAAAAGTCGTAACAAGG	[21, 24]
	D1-D2	F-ACCCGCTGAACTTAAGC R-TCCTGAGGGAAACTTCG	[22]
	RPB1	F-GARTGYCCDGGDCAYTTYGG R-CCNGCDATNTCRTTRTCCATRTA	[24]
	RPB2	F-GAYGAYMGWGATCAYTTYGG R-CCCATWGCYTGCTTMCCCAT	
	COI	F1-TTACAAGGTGATCATCAATT F2-GTATTAAAATTTCTATCTGTAAG R1-TTTCTATCTGTAAGTAACAT R2-TTTACAAGGTGATCAA	[23]
	ACT	F-ATGTGCAAGGCCGGTTTCG R-TACGAGTCCTTCTGGCCCAT	[97]
	TEF1a	F-GCYCCYGGHCAYCGTTTYAT R-ACHGTRCCRATACCACCRATCTT	[98]

Table 1 (continued)

Organisms	Barcode gene	Primers (5′–3′)	References
Protist	ITS	F-CGTAACAAGGTTTCCGTAGG R-TCCTCCGCTTACTGATATGC	[45]
	V4	F-CCAGCASCYGCGGTAATTCC R-ACTTTCGTTCTTGATYRA	[45]
	V9	F-GTACACACCGCCCGTC R-TGATCCTTCTGCAGGTTCACCTAC	[42]
	D1-D2/D2-D3	F-AGCGGAGGAAAAGAAACTA R-ACGATCGATTTGCACGTCAG	[42]
	D3-D5	F-GACCCGTCTTGAAACACGGA R-TCGGAAGGAACCAGCTACTA	[50]
	COI	F-GWTGRGCKATGATYACACC R-ACCATRTACATATGATGWCC	[5]
	RbcL	F-TTAACCTCCATCGTGGGTAACG R-CAGGCATAGAAGCCCAATCTTG	[99]
	H4	F-GGTATTACTAAGCCCGCTATCAGAAGA R-GGTCTTTCTTCTGGCGTGTTCAGTGTA	[53]
	Spliced leader RNA gene	F-GTATAAGAGACAGNNNNNNN R-TCAGTTTCTGTA	[44]
Archaea	16S	F-GCY TAA AGS RIC CGT AGC R-TTM GGG GCA TRC IKA CCT	[26]
	Type II chaperonin (High GC primers)	F-GGC CCG AAG GGC ATG GAC AAG ATG R-GGC ATG TCG TCG ATG CCC TTC TG	[26]
	Type II chaperonin (Universal primers)	F-GGI CCI MRR GGI ITI GAY AAR ATG R-GCI AII TCR TCI ATI CCY TTY TG	[26]
Bacteria	RpoB	F-TTTCCCTACACGACGCTCTTCCGATCTGGYTWYGAAGTNCGH- GACGTDCA R-GGAGTTCAGACGTGTGCTCTTCCGATCTTGACGYTGCATGT TBGM- RCCCATMA	[30]
	16S rRNA gene	F-GAGTTTGATCCTGGCTCAG R-GTATTACCGCGGCTGCTG	[29]
	cpn60	F1-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAIIIIGCIGGI- GAYGGIACIACIAC F2-GCTTCGTGGGCTCGGAGATGTGTA TAAGAGACAGYKIYKITCIC- CRAAICC IGGIGCYTT R1-TCGTCGGCAGCGTCAGATGTG TATAAGAGACAGGAIIIIGCIGGYGAC GGYACSACSAC R2-GCTTCGTGGGCTCGGAGATGTGTA TAAGAGACAGCGRCGRTCRCC GAAGCCSGGIGCCTT	[33]
	Tuf	F-GCTCCTGAAGAAARAGAACGTGG R-ACTTGDCCTCTTTCKACTCTACCAGT	[31]
	RIF	F-TACGGCTTCGACACCTTCG R-CGGTGATCTTCTTGTTGGCG	[32]
	Gnd	F-CGCGGATCCGGWCCWWSWATWATGCCWGGWGG R-CGCGGGCCCGTATGWGCWCCAAAATAATCWCKTTGWGCTTG	[31]

The lower-case letters indicate nucleotides added to the genomic sequence to facilitate PCR; D = A/G/T, Y = C/T, R = A/G, W = A/T, M = A/C, B = C/G/T, V = A/C/G, H = A/C/T, K = G/T, N = A/G/C/T

more ambiguous variation in the sequence. The mitochondrial genome evolves at a higher rate than the nuclear genome. Therefore, mitochondrial genomic sequences are more informative in differentiating or distinguishing closely related species [15]. So far COI gene has been used as

barcoding gene for moths, butterflies, collembolans, beetles, bats, spiders, wasps, ants, fishes, Reptilia, birds, chickens, musk deer, fruit fly, and crustacean larvae [4, 16]. Some primate taxonomists recommend that ND5 (Mitochondrial gene encoding NADH:Ubiquinone Oxidoreductase Core Subunit
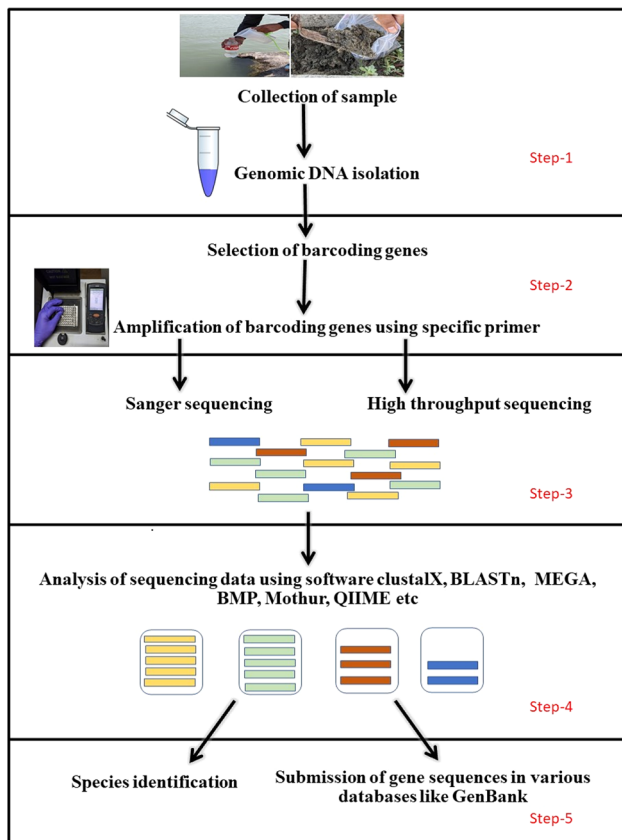
Fig. 1 Schematic representation of the methodology for DNA barcoding

5) and COII should be used as a barcode in primates species delineation and suggest that these two genes should be more appropriate markers than COI due to a more pronounced barcoding gap [17].

Barcodes for identification of plants

For plant species identification, the selection of barcoding genes remains very controversial. Plant mitochondrial genome exhibits a low rate of mutation (nucleotide substitution) that restricts COI as a universal plant barcode. Plant taxonomists have spent a large amount of time and found the Chloroplast genome as an alternative to the mitochondrial genome. In 2009, CBOL plant working group proposed seven potential barcodes such as rbcL (large subunit of ribulose 1,5 bisphosphate carboxylase), matK (maturase K), psbA-trnH (intergenic spacer region, rpoC1 (RNA polymerase C1), rpoB (RNA polymerase B), atpF-atpH (encodes for ATP synthase subunits CFO I and CFO III) and psbK-psbI (encodes for polypeptide K and L of photosystem II) [18]. Nuclear gene ITS (internal transcribed spacer) and all the chloroplast barcodes have been positively tested in plant

species [19]. Comparatively, 600–800 base-pair region of matK in association with rbcL gives the most satisfactory result and designated as core barcoding gene, while psbA-trnH work as a good marker for other plant species and identified as an important supplementary marker, but there is no single marker for identifying all the plant species [20].

Barcodes for identification of fungi

Identification of fungi through morphological methods is often difficult because they only occasionally display morphological characters suitable for identification. A molecular tool such as DNA barcoding is the best way to evaluate fungal diversity. The ITS, D1-D2 region of the large subunit of ribosomal RNA gene, RPB1 and RPB2 of the large subunit of RNA polymerase II, γ-actin (ACT), β-tubulin II (TUB2), translation elongation factor 1-α (TEF/α), DNA topoisomerase I (TOPI), phosphoglycerate kinase (PKG) are used as a barcode for identifying fungal species [20–22]. COI has a higher resolution in few groups of related species such as *Penicillium*, and *Entolomasarcopum* but in other groups it may not give satisfactory results [23]. Schoch and his group has proposed ITS as a universal barcode for the identification of fungi [24]. The length of ITS region in fungi is around 600 bp long, with two variable spacers, ITS-1 and ITS-2, interrupted by the highly conserved 5.8S rRNA gene. Another significant benefit of utilizing ITS as a barcode is that each haploid genome often contains several tandemly repeated copies of the ribosomal rRNA gene cluster (including ITS), allowing it to be amplified even from small amounts of biological materials [6]. Stielow et al. have assessed the potentiality of D1-D2 region of LSU, β-tubulin II (TUB2), γ-actin (ACT), translation elongation factor 1-α (TEF/α), the second largest subunit of RNA-polymerase II (RPB2), DNA topoisomerase I (TOPI), phosphoglycerate kinase (PKG), hypothetical protein LNS2 as an alternative DNA barcode. Among these genes TEF/α has the potential as a secondary DNA barcode due to sufficient intra- and inter-specific variation, while TOPI and PKG show high resolution for the phylum *Ascomycota*, and TOPI and LNS2 for the subphylum *Pucciniomycotina* [22].

Barcodes for identification of archaea

Archaea is a major component of microbial diversity and has a prominent place in the Tree of Life [25]. 16S rRNA gene has been widely utilised as a barcode for evaluating the diversity of archaea [25, 26]. The 16S rRNA gene is not sensitive enough to discriminate closely related microbes, particularly at the species level [27]. In a study, type 2 chaperonin or thermosome (e.g. TCP-1 ring complex/chaperonin

containing TCP-1), which are present in both archaea and eukaryotic cytoplasm, proposed as a potential complementary barcode for 16S rRNA gene to assess the archaeal diversity since it has larger barcoding gap and generate more OTUs (operational taxonomic units) than 16S rRNA gene [28].

Barcodes for identification of bacteria

16S rRNA gene is a universal marker as it is highly conserved in all the species of bacteria. The length of the 16S rRNA gene is 1600 base pairs and contains nine hypervariable regions of V1–V9. More conservative regions are valuable for identifying higher-ranking taxa, whilst more rapidly evolving ones can aid in genus or species identification. The V2–V3 region of 16S rRNA gene has higher resolution for identifying lower-ranked taxa (species and genus) [29]. The diversity of bacteria can also be accessed by using COI, rpoB, cpn60 (encodes for chaperonin protein), tuf (elongation factor), RIF (Replication initiation factor), and gnd (Gluconate-6-phosphate dehydrogenase) gene as barcode [30–33]. These genes have several benefits over frequently used 16S rRNA gene i.e. as they are frequently found in single copies in bacterial genome, and develop silent mutations owing to codon degeneracy, resulting in improved species resolution. Of these cpn60 gives better results and can be used as a possible alternative for assessing bacterial diversity [20, 26] and cpn60 is the only target that can be addressed with 'universal' PCR primers, and a curated sequence database, cpnDB, is available. For closely related species, the cpn60 gene has stronger discriminating power than the 16S rRNA gene, and the uniform size and sequence variability of the cpn60 'universal target' (UT) make sequence comparisons and other bioinformatics tasks easier [26].

Barcodes for identification of viruses

Viruses are the most abundant (approx. 10–12 times higher than the total no. of cells) life forms on earth. So far, there is no standardized barcode fragment for detection of viruses [20].

Barcodes for identification of protist

CBOL has initiated a Protist working group (ProWG) to identify barcode region across all protist lineages and setting up a reference DNA barcode library. CBOL ProWG has introduced a 2-step pipeline for protists: first, the universal pre-barcode to be used for preliminary identification; second, a group-specific barcode to be applied for species-level identification [12]. The hypervariable V4 region of 18S rRNA gene is proposed as the universal eukaryotic pre-barcode, while group-specific barcode is defined separately for each significant protistan lineage [34]. So far, ITS, COI, rbcL,18S rRNA gene, 28S rRNA gene region have been proposed as a protistan DNA barcode [35–37]. ITS, the universal barcode in fungi, also has high discriminatory power for ciliates, dinoflagellates, and oomycetes [20, 37, 38]. Mitochondrial COI, which is the universal barcode for animals and default barcode for other organisms as well, is also positively tested in protist [36]. Hypervariable regions V4 and V9 of 18S ribosomal RNA gene are promising barcodes to access the diversity and phylogenetic relationship of diatoms, dinoflagellates and ciliates [39–41]. D1-D2 and/ or D2-D3 regions at 5′ end of large subunit of rRNA gene serve as potential barcodes for many protists lineages such as diatoms, ciliates, and dinoflagellates [35, 42, 43]. Some group-specific barcodes such as rbcL and spliced leader RNA gene are also utilized in photosynthetic protists and trypanosomatids, respectively [44].

Barcodes for identification of ciliates

Large public reference libraries of DNA barcodes are being developed for animals, plants, and fungi, but no universal barcode has been accepted for ciliates species identification [37]. Various barcodes for ciliate identification are (1) mitochondrial cytochrome c oxidase subunit I gene (COI gene); (2) hypervariable regions of the small subunit (SSU) rRNA gene such as V4 and V9 region; (3) ITS region; (4) D1-D2 regions of the large subunit of rRNA gene (LSU) and (5) histone H4.

Mitochondrial cytochrome-c oxidase subunit 1 gene (COI)

Within ciliates, taxonomic and molecular phylogenetic studies using COI gene have been used in *Paramecium*, *Tetrahymena*, *Carchesium*, *Miamiensis*, *Sterkiella* and *Pseudokeronopsis* [5, 45]. All the above studies prove that the highly variable COI gene of ciliates can identify closely related species and cryptic species since it has a distinct barcode gap between maximum intraspecific and minimum inter-specific genetic divergence (Table 1). Within ciliates, the COI gene have been successfully sequenced from *Tetrahymena* and *Paramecium*[4, 45]. The COI gene (average 2000–2200 nucleotides long) have been found to be widely dissimilar from other eukaryotes as it includes > 300 nucleotides long insert region which has exceptional variation in a genetic distance value and intraspecific genetic divergence [46]. This insert region is used as a barcode to discriminate

closely related species based on genetic divergence [8]. Earlier studies have shown that the COI gene of ciliates has high intraspecific genetic divergence than nuclear gene [5, 46]. Park et al. (2019) have reported a 478 bp long COI sequence of 69 population of spirotricheans ciliates, which has maximal intraspecific genetic divergence ranging from 0 to 14.8% and minimal interspecific genetic variation, i.e.,13.6–47.3%. They identified three putative cryptic species, *Caudiholostichaylvatica, Diophrys scutum*, and *Euplotes vannus* [5]. COI nucleotide tree has a higher resolution to discriminate closely related and sibling species at and below the species level. Recently, Zhang et al. [36] studied the phylogenetic relationship of subclass scuticociliates with the usage of nuclear SSU-rRNA gene, mitochondrial SSU-rRNA gene and COI gene as a molecular marker and showed that sequence divergence of COI (average 24%) is more significant than mtdSSU-rRNA gene (average 21%) and nSSU-rRNA gene (average 11.5%). They proved that COI is a better choice as a molecular marker to examine phylogenetic relationships than mtdSSU-rRNA gene and nSSU-rRNA gene[36]. However, consortium for the barcode of life does not consider COI as an appropriate barcode for uncovering ciliates species because of issues like the absence of functional mitochondria in some ciliates from the anoxic environment e.g., ciliates belonging to *Metopus*and *Trimyema* genus and presence of heteroplasmy [4, 5].

Small subunit (SSU) rRNA gene

SSU rRNA gene was the first and widely used molecular marker in genealogy and systematics study of ciliates because it can be sequenced accurately, universally, availability of diverse and large database from NCBI, and includes both conserved and variable nucleotide sequences allowing combined phylogenetic reconstruction and biota recognition at various taxonomic levels. Within ciliates, the average size of 18S rRNA gene is ~1771 bp long except in litostomatea which has 1635-1641 bp [45], but this entire region of 18S rRNA gene is not used for species identification. Only the hypervariable regions (V1–V5 and V7–V9) of 18S rRNA gene are used for species identification. Among them, V4 and V9 hypervariable regions are considered the famous barcoding gene. The hypervariable region V9 is immensely used as a genetic marker for evaluating eukaryotic diversity and also a prime candidate for assessing protist lineage richness, while the V4 region of SSU rRNA gene is the primary candidate for studying the phylogenetic relationship of eukaryotes. V4 region is more extensive, more variable, and show better resolution to explore the evolutionary relationship of eukaryotes than the V9 region [39]. The secondary structure of hypervariable region V9, V7, V4, V2 of 18SrRNA gene in urostylids shows a high degree of variability and provides further evidence that the V4 region is

the most effective for revealing interspecific relationship. On the other hand, the V9 region seems appropriate at the family level or higher [47]. It is recommended to use V4 and V9 together to assess the diversity and phylogenetic relationship of eukaryotic microbes [39].

Internal transcribed spacer (ITS) region

Internal transcribed spacer (ITS) and the external transcribed region (ETR) are the flanking regions of the SSU, and the 5.8S rRNA is a non-coding part of LSU rRNA. ITS1 is present between SSU rRNA and 5.8S rRNA, and ITS2 is present between 5.8S rRNA and LSU rRNA [45]. Various studies suggest that ITS region has the potential of promising barcode for ciliate identification and investigation of intraspecific genetic diversity at species and population levels since they shows much higher rate of evolutionary changes (> 100 times) than the coding regions of the ribosomal subunit [34, 48, 49]. Usually, phylogenetic trees of ITS1-5.8S-ITS2 region usually do not differ significantly from those inferred from the 18S rRNA gene, implying that the ITS region is a viable proxy for genealogical studies. Although both the ITS1 and ITS2 have sufficient conserved and variable region, but ITS2 seems to have more information and may be more valuable for comparisons at the family, order, and even higher level. Moreover, the secondary structure of the ITS2 molecule has been employed to improve the quality of species-level phylogenetic reconstructions. Apart from phylogenetic reconstructions, the compensatory base changes (CBCs) in the ITS2 region correlate with sexual incompatibility and so can be used for species discrimination [48]. More and more studies suggest that using both primary sequence and secondary structure of ITS2 produce higher phylogenetic resolution [34, 49]. Zhan et al. used ITS1-5.8S-ITS2 and the ITS2 as a barcode to delimitates *Pseudokeronopsis* species and found that both the ITS1-5.8S-ITS2 and the ITS2 regions shows similar levels of genetic variation and substantial gaps between intraspecific and interspecific distance (0.52–3.72% for ITS2; 0.42–3.84% for ITS-5.8S-ITS2). Additionally, they also proposed a genetic divergence of 1.5% as an ideal threshold of ITS1-5.8S-ITS2 and ITS2 to distinguish *Pseudokeronopsis* species and also suggested the ITS1-5.8S-ITS2 can be used as an ideal SGS (Second generation sequencing) metabarcode for assessing ciliates environmental diversity [34].

Large subunit (LSU) of rRNA gene

LSU rRNA gene is a good barcoding gene for discriminating closely related taxa because it has a higher evolutionary rate than SSU. Similar to SSU, LSU rRNA gene has variable region such as D1-D12, of which D1-D3 region show much

higher variation than other variables such as D4, D5, D7, D8, D12 [50]. Over the last decade, the D1-D2 region of LSU rRNA gene emerged as a promising barcode marker for species identification up to species level. Santoferrara et al. has proposed D1–D2 region of LSU rRNA gene with a 1% threshold value (for tintinnid) as a barcoding marker for ciliate species identification and potentiality of this marker further assessed by Stoeck et al., Zhao et al., Forster et al. [37, 42, 51, 52]. D1–D2 region of LSU rRNA gene has several advantages over other frequently used markers such as showing a clear barcoding gap, rapid evolutionary rate enough to provide higher diversity resolution than SSU and higher universality and constant threshold value than COI [51]. LSU has less intra-clonal and intraindividual variability [45]. One study suggested that the D2 region is a suitable marker for discriminating all *Frontonia* morphospecies since it shows a clear barcoding gap with a threshold of 4.5%, while the D1 region alone is not ideal for determining because it shows the overlap between intraspecific and interspecific genetic divergence [37]. So far, D1–D2 region of LSU together have been used as a marker for diatoms, dinoflagellates, tintinnid ciliates, *Paramecium* and *Frontonia* species [37, 42, 51, 52]. All the above-discussed features such as higher universality, conserved primers for its amplification in ciliates and constant threshold value as well as the presence of high quality manually curated databases (i.e., SILVA), makes hypervariable D1–D2 region of LSU rRNA gene promising DNA barcodes for ciliates species delineation [37].

Histone H4

The histone H4 is known to be a highly conserved protein among all eukaryotes with the exception of the high degree of variation observed in the ciliate species [45]. The histone protein is responsible for the organization of eukaryotic chromatin. The ciliate histone H4 encoded by the macronuclear gene. Due to considerable difference within ciliates, histone H4 is considered an excellent molecular marker to study phylogenetic relationships and can be used as DNA Barcoding [53].

Advancement in DNA barcoding

By using DNA metabarcoding and microarray, it is very feasible to develop a powerful taxonomic identification tool. The development of metabarcoding was compelled by the growth of next-generation sequencing technologies capable of producing millions of sequences at a comparably low price. The Metabarcoding approach uses the same general principle as the traditional DNA barcoding, but this approach focuses on assessing the community's whole diversity instead of identifying individual taxa [54]. This advancement has overcome the limitation of traditional DNA barcoding, such as extensive sampling efforts. Metabarcoding relies on the shorter DNA fragments instead of whole 658 bp fragment (standard barcodes) used in classical DNA barcoding. Metabarcoding approaches on environmental and faecal samples have revealed population structure in a variety of species [55]. The main problem associated with standard barcodes is length, i.e., longer than 500 bp used in the traditional approach for achieving high discriminatory power at the species level. Unfortunately, metabarcoding assess the diversity up to family, order, or higher taxonomic level from environmental DNA sample [56, 57]. One of the most challenging aspects of metabarcoding on which their accuracy depends is to find new and acceptable primer pair and their corresponding markers. An ideal metabarcoding marker should have a short length (e.g., 100 bp) for easy sequencing, good conserved flanking primer binding sites to minimise taxonomic bias during PCR amplification, and a sufficiently variable intervening sequence for species identification [58]. V4 region is the primary choice metabarcode for assessing the richness and phylogenetic relationship of eukaryotic microorganisms, while COI is widely used for animals [34]. Primers with fewer template–primer mismatches are better for quantitative DNA metabarcoding, especially for species of higher relative abundance in a sample. Barcode of life DATABase (BOLD) system has a primer database (http://boldsystem.org/index.php/Public_Primer_PrimerSearch) that store all the published primers. Researchers can either determine the primers of their interest by searching in primer database or design their primer by using software like Primer3, QPRIMER, UniPrime, Primaclade, Amplicon program, Primer Hunter, Greene SCPrimer andecoPrimer etc.

Next-generation sequencing (NGS) is a cost and time-saving high throughput platform and generate millions of reads in a single run for only one environmental sample. Braukmann et al. compare the performance of three Next-generation platforms, namely Illumina MiSeq, Ion TorrentS5, and Ion Torrent PGM, and showed that they perform equally well for species recovery, although MiSeq is often recommended because of its low error rate and well-established bioinformatics methods [59]. Illumina NovaSeq is the recent advancement in sequencing technology with the same sequencing depth as MiSeq but assesses more metazoan diversity. One of the known limitations of NGS for metabarcoding is the generation of short read length, i.e., 400 bp [7]. The development of Illumina MiSeq overcome this limitation of short read length by generating longer sequence reads (600–800 bp) that provide better taxonomic resolution and phylogenetic inference [7, 54]. Metabarcoding data has significantly improved the estimates of microbial communities and offered precise information about the

structure and spatiotemporal turnover of microbial populations, particularly in the ocean. According to some estimates, there are 50,000 to 100,000 protist OTUs (operational taxonomic unit) in the world's oceans, which is five to ten times the number of bacteria and archaea combined. These OTUs have different distribution patterns, with varied ocean regions have various ecosystems in terms of taxonomic composition and relative abundances. The metabarcoding data also used to relate microbial community distribution patterns with assembly mechanisms [54].

Microarray or biochip, or gene chip are other high-throughput platforms for identifying species. The ability to identify thousands of targets in a single hybridization experiment makes microarray one of the most potent molecular tools [60]. A microarray made up of a DNA barcode that may be used to design probe sequences in microarray analysis. A DNA microarray containing a species-specific oligonucleotide probe is a viable alternative to the traditional Sanger sequencing for identifying species in food sample. Several commercial DNA chips are available to identify animal species in food samples (e.g. CarnoCheck DNA-Chip, Greiner Bio-One, Austria; LCD Array Kit MEAT 5.0, Chipron, Germany) [61]. Fish species are identified in both culinary and forensic samples using 16 S rRNA gene, Cytochrome b, and COI derived probes [61–63]. Shortly, the microarray-based identification approach will play a more prominent role in molecular species identification [56].

Third generation sequencing such as Oxford Nanopore Technologies (ONT)'s MinION™ and PacBio sequencing is an another sequencing advancement that makes DNA barcoding more feasible [64]. MinION nanopore sequencing overcome the limitation associated with the Sanger sequencing and NGS. Sanger sequencing is costly and requires well equipped molecular laboratory and ABI sequencer. On the other hand, next generation sequencing is cost-effective only when large numbers of specimens are barcoded simultaneously, generate sequence reads with high accuracy, also requires expensive equipment in laboratory and has long sequencing run time [65]. ONT MinION™ nanopore sequencing, introduced in 2014, is authentic, quick, third generation sequencing, cost-effective, generate long reads, enables real time analysis and do not require well-equipped molecular laboratory [64]. Various studies proposed that complete genome sequence of microbes can be obtained by using multiplexed reads from a single MinION™ run in combination with matched Illumina short reads such as *Staphylococcus aureus, Klebsiella pneumoniae*, and multidrug resistance encoding plasmid [64, 66]. With the introduction of MinION nanopore sequencing several full plasmid sequences can now be obtained in a single MinION run using a quick barcoding methodology. MinION™ has also been successfully used in bacterial and plant identification, microbiome characterisation, and DNA fingerprinting [67,

68]. Nanopore sequencing has also proven to be a very versatile technology, e.g., allowing for whole genome sequencing and assembly of fungal and human genomes, as well as sequencing full-length RNA transcripts using both direct RNA and cDNA sequencing [69].

PacBio sequencing, which is a single molecule real time sequencing, is an alternative DNA barcoding approach for large sample sizes: its workflow simplifies and reduces post-sequencing manipulation, generating longer read length and faster running time that provide better taxonomic resolution [70]. Due to longer reads of PacBio sequencing, one can sequence through longer repetitive sequences and detect mutations, many of which are linked to disease. Furthermore, because of its potential to sequence full-length transcripts, it is beneficial for identifying gene isoforms and allows reliable discoveries of novel genes and novel isoforms of annotated genes. Furthermore, PacBio's sequencing technique can be used to detect base modification such as methylation [71]. PacBio sequencing also has some drawbacks including costly, high error rate, and low throughput [71, 72]. The High sequencing error rate can be reduced by re-sequencing of circular molecules several times. So far PacBio sequencing has been used successfully in meta-barcoding analysis of arthropods and fungi [72]. Several researchers suggested that to use PacBio sequencing along with SGS since both of them are highly complementary in term of their advantage [70, 71].

MALDI–TOF MS (Matrix-assisted laser desorption/ionization time of flight mass spectrometry) is being more commonly employed as a novel tool for barcoding, however this method should be based on accurate species identification both morphologically and genetically. This approach is extensively used to identify arthropods [73]. Other than arthropods, MALDI TOF MS has been successfully used in identification of bacteria and archaea [74].

DNA barcoding in combination with nanotechnology is another novel approach that has been shown to be highly sensitive, allowing for rapid uniplex and multiplex detection of pathogens in food, blood, and other samples [75]. Nano-based detection methods increase the sensitivity level up to ten times as compare to PCR and other detection methods such as radio-immunoassay, microarrays, enzyme-linked immunosorbent assay (ELISA) etc. Gold nanoparticles and magnetic nanoparticles based "fluorescent bio-barcode DNA assay" has been used to probe the *Salmonella enteritidis* genes [76]. Another bacterial gene Exotoxin A has been detected by using magnetic and gold nanoparticles-based fluorescence bio-barcode DNA assay [77]. Recently, Ding et al. (2021) identified the DNA marker in liquors, condiments and milk by using gold nanoparticles [78]. Valentini et al. (2017) introduced a new approach, NanoTracer that streamlines all the analytical steps involved with traditional DNA barcoding and enabling it sequencing-free and

accessible outside the specialized laboratories. NanoTracer enables quick naked eye molecular validation of any food with simple and inexpensive processing and limited instrumentation [79]. Species-specific lateral flow dipstick (LFD) assays developed by Taboada et al. (2017) for identifying Atlantic cod, Pacific cod, Alaska pollock and ling in food products, using gold nanoparticles to enable visual identification with high sensitivity even for processed samples [80].

Alternatives to DNA barcoding

Dip-stick approach is a recent innovation in which lateral flow assay combined with species specific primer to detect wide variety of species from environmental samples [55].

Non-targeted NGS is an alternative to DNA barcoding for species identification, phylogenetics, and phylogeography. Non-targeted NGS methods, such as whole genome sequencing, metagenomics and mitogenomics, do not rely on amplification. Therefore, problems like primer biases and non-standard amplification have no effect on these methods [81].

Mitogenomics is a variant of metagenomics, shotgun sequencing approach that uses mitochondrial genomes as references rather than nuclear genomes. Mitogenomes are easily amenable to genome skimming, in which a high copy region of the genome is assembled into longer contigs from low coverage shotgun sequencing of a specimen mixture [82]. This method is desirable because of its advantages. Firstly, a mitogenome and its genes are commonly used molecular markers. Secondly, the mitogenomes structure are conserved, whereas sequences can be extremely diverse. Thirdly, mitogenomes are small and easy to obtain and can be reconstructed directly using bioinformatics methods. Fourthly, large numbers of mitogenomes are available in public databases [83]. Furthermore, this approach is not affected by problems like Primer biases and non-specific amplification. Several studies have shown that mitogenomics outperforms metabarcoding in terms of discriminatory power [83, 84]. However, the utility of mitogenomic is limited as it is quite expensive because each sample requires an individually prepared library, samples must be sequenced more deeply than for metabarcoding, and assembling a mitogenome reference database incurs additional costs for specimen acquisition, sequencing, and assembly [84]. It has been found that phylogentics constructed on the basis of mitogenomics or nuclear ribosomal RNA repeats are well resolved and with this, one can distinguish between closely related species.

Bayesian inference under the multispecies coalescent model is also an alternative to DNA barcoding. This method can discriminate species with high power when multi-locus data are used, even if the species is represented by a single specimen [85].

All of the advancements discussed above, particularly HTS sequencing, whole genome sequencing, and metagenomics, have been viewed as a threat to DNA barcoding. HTS sequencing, whole genome sequencing, and metagenomics produce massive amounts of genomic data. The genomic data analysis takes more time, requires more bioinformatic expertise compared to standardized DNA barcodes, requires more energy for data computation and storage, and is difficult to control quality when shared [86]. Therefore, DNA barcoding remains the preferred method for species identification and biomonitoring, while genomics is useful for understanding genome complexity, diversity, and function. Rather than being a threat, barcoding and genomics have clear mutual benefits, with DNA barcoding establishing a platform for well-identified samples in genome sequencing projects and genomic studies contributing insights that may identify new barcode regions in groups where the standard regions are suboptimal [55].

Reference library construction

Currently, DNA Barcoding (Metabarcoding) is the most effective approach for identifying species, and its accuracy is relied on the resolution of DNA barcodes and the reference library. BOLD is the largest reference library or database and its growth has been exponential over the last decades. The International Barcode of Life Consortium (iBOL) launches several projects to expand the DNA barcode reference library or database, including 500K (completed in 2015), BIOSCAN (launched in 2019), and the Earth Biogenome Project [55]. Despite this, very few such libraries have been developed.

Constructing a reference library with extensive species coverage presents several challenges. The first challenge is the high expense of collecting raw data, which can be accomplished through DNA sequencing. Conventional sanger sequencing is expensive and of low efficiency [11]. This obstacle must be overcome by acquiring NGS and third generation sequencing platforms such as PacBio and Nanopore. Another challenge is selecting a critical sequencing platform for obtaining high quality results at a low cost, which can be accomplished by taking into consideration base quality, data sizes, sequencing depth, and cost efficiency [87]. There are several NGS platforms but the most appropriate choice for DNA barcoding is Roche-454 [88], which is no longer available. In terms of high base quality and low cost, the Illumina system and Ion Torrent S5 platform are currently the most suitable NGS platform for conventional DNA barcoding than third generation sequencing platforms [87]. Several studies have compared the performance of the

Illumina and Ion torrent platforms, but researchers are still unsure which one is better suitable for DNA barcoding. Both the Illumina system and the Ion Torrent S5 generate massive amounts of data, posing new challenges for data analysis [89, 90]. Several software packages have been developed, including Vsearch, Usearch, Mothur, Zotu, DADA2, and others. However, these current softwares is not perfect for creating DNA barcodes, and it was not designed for conventional DNA barcode data analysis. A new data analysis method called Cotu has been developed for conventional DNA barcodes, and its performance outperforms other commonly used methods like Zotu and DADA2 [87]. However, more research is needed to confirm and adopt Cotu for data analysis. Using an appropriate NGS platform and advanced data analysis methods, a regional or even global DNA barcoding reference library with high species coverage is likely to be developed within a few years.

The majority of current work on DNA barcoding has been done in Europe and North America, which could be another reason for the limited reference library/database. Financial assistance is also required for the creation of a high-quality reference library. Funding for DNA barcoding research should encourage the creation and curation of a reference library. A large number of national and global collaborations will aid in financial support as well as to combine local knowledge on species identification with sequencing capacity [91]. There are several curated natural history museums around the world that house a large number of vouchered specimens. Obtaining DNA barcoding data from these vouchered specimens should significantly improve the quality of the reference database [55]. Another possible step would be to incorporate reference barcodes on a regular basis. To improve the reference barcode library, make it mandatory to submit the reference barcode when describing a new species.

Strength and limitation of DNA barcoding

Apart from taxonomists, the DNA barcoding technique can benefit scientists from other fields such as biotechnology, food industries, forensic science, and animal diet [57]. Taxonomist uses a *sensu-stricto* (refers to the identification of species level using a single standardized DNA fragment) approach of DNA barcoding, while other scientists use a *sensu-lato* (refers to the identification of any taxonomic group using any DNA fragment) approach. The main application of DNA barcode in taxonomy to accelerate the species identification and revealing cryptic species. DNA barcode data can provide a comprehensive foundation for organizing and identifying species-rich groups in the tree of life, serving as a good starting point for taxonomy, biodiversity assessments, and biomonitoring [55].

This technique can also help to settle enduring nomenclatural debates, leading to the taxonomic revision of inadequately defined morphospecies. DNA barcoding approach is also widely used by ecologist due to several reasons. First, the diversity of ecologically essential life forms such as ciliates and nematodes are mostly unknown and the DNA barcoding approach is a better way to assess the biodiversity of such life forms [34, 92]. Second, DNA barcode can also detect endangered species from hair and faeces sample left behind by animals [57]. Third, illegal trade in animal by-product can be monitored with the help of DNA barcoding technique [93]. Fourth, DNA barcoding can be advantageous in the field of biosecurity. This is one of the available technique to identify invasive species at a very early stage of their life cycle, such as an egg or larval stage [7]. Fifth, the past environment can be reconstructed by using this technique. Finally, by using the DNA barcoding approach diet of animals can be analysed from faeces or stomach content [57]. Within the food industry, DNA barcoding reveals mislabelling of processed food that may lead to health hazards. Recently COI gene is used as a DNA barcode to reveal mislabelling of seafood in the European market [94]. DNA barcoding can be highly useful in forensic science [20, 57]. Some species of plants are poisonous in nature, such as *Datura* sp., *Brugmansia* sp., and *Cannabis sativa*, which cause serious health problems to humans and animals when ingested. Rapid identification of the poisonous plant is required for appropriate treatment, and identification from vomited or excreted samples by visual observation is not feasible because most of the plant part can be degraded. So, DNA barcoding will be useful for identification from these degraded samples. Recently rbcL and ITS2 genes are used as a barcoding marker for identifying poisonous plant species [20].

DNA barcoding tool overcomes the limitation of the classical identification method, but this approach itself has certain restrictions. One of the most significant drawbacks of the DNA barcoding method is that there is no universal primer or universal gene found in all forms of life and has enough sequence divergence to allow for species differentiation [56]. Very less number of reference DNA barcode library, and Loss of quantitative information due to primer and polymerase biases [84]. DNA barcoding distinguishes species based on intraspecific and interspecific genetic variation, although the ranges of such variation are unclear and may differ between taxa [31]. The existence of pseudogenes and heteroplasmy reduces the accuracy of DNA barcoding and increases the complexity of database. A pseudogene can result in the erroneous division of single species into several species. Pseudogenes can produce heteroplasmy, which causes more than one kind of mtDNA to coexist in the same individual and limiting species identification by DNA barcoding [56].

Conclusion

Through the rapid development in the last 2 decades, DNA barcoding has emerged as a highly effective molecular tool for taxonomic classification. It relies on barcoding gap within a short and standardized region of the genome for assessing species diversity and phylogenetic relationship. The DNA barcoding allows more accurate and cost-effective biodiversity characterization and its use in accelerating species discovery is becoming increasingly important, given the current threats to biodiversity and elevated rates of extinction. Several DNA barcodes have been extensively used for biological species identification, including the mitochondrial COI gene, rbcL, matK, trnH-psbA, 16S rRNA, V4, D1–D2 region, and ITS (nuclear internal transcribed spacer). But there is no single barcode for all the species and it is very hard to find because of differences in evolutionary rates. Over the years, the DNA Barcoding approach has become more accurate, sensitive and faster due to several advancement such as next generation sequencing, third generation sequencing, and Nanotracer. A large-scale DNA barcoding research using an appropriate NGS platform and advanced data analysis methods will surely help to create a reference DNA barcode library of all organisms in order to avoid misidentification and definitely simplify the interpretation of sequencing results. Also, barcoding by mitogenome and snRNAs can upgrade current barcoding strategies.

Data availability Not applicable.

Declarations

Conflict of interest The authors declare that there are no conflicts of interest.

Ethical approval This article does not contain any studies conducted on humans and or vertebrate animals.

Consent to participate Not applicable.

Consent to publish Not applicable.

References

1. Feio MJ, Filipe AF, Garcia-Raventós A et al (2020) Advances in the use of molecular tools in ecological and biodiversity assessment of aquatic ecosystems. Avanços no uso de ferramentas moleculares na avaliação ecológica e biodiversidade dos ecossistemas aquáticos. Limnetica. https://doi.org/10.23818/limn.39.27

2. Schweiger AK, Cavender-Bares J, Townsend PA et al (2018) Plant spectral diversity integrates functional and phylogenetic components of biodiversity and predicts ecosystem function. Nat Ecol Evol 2:976–982. https://doi.org/10.1038/s41559-018-0551-1

3. Rico-Sánchez AE, Sundermann A, López-López E et al (2020) Biological diversity in protected areas: not yet known but already threatened. Glob Ecol Conserv 22:e01006. https://doi.org/10.1016/j.gecco.2020.e01006

4. Chantangsi C, Lynn DH, Brandl MT et al (2007) Barcoding ciliates: a comprehensive study of 75 isolates of the genus Tetrahymena. Int J Syst Evol Microbiol 57:2412–2423. https://doi.org/10.1099/ijs.0.64865-0

5. Park M-H, Jung J-H, Jo E et al (2019) Utility of mitochondrial CO1 sequences for species discrimination of *Spirotrichea ciliates* (Protozoa, Ciliophora). Mitochondrial DNA Part A 30:148–155. https://doi.org/10.1080/24701394.2018.1464563

6. Xu J (2017) Fungal DNA barcoding. In: The 6th international barcode of life conference 01:913–932. https://doi.org/10.1139/gen-2016-0046@gen-iblf.issue01

7. Comtet T, Sandionigi A, Viard F, Casiraghi M (2015) DNA (meta) barcoding of biological invasions: a powerful tool to elucidate invasion processes and help managing aliens. Biol Invasions 17:905–922. https://doi.org/10.1007/s10530-015-0854-y

8. Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc Lond B 270:313–321. https://doi.org/10.1098/rspb.2002.2218

9. Yao H, Song J, Liu C et al (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS ONE 5:e13102. https://doi.org/10.1371/journal.pone.0013102

10. Stoeck T, Behnke A, Christen R et al (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. BMC Biol 7:72. https://doi.org/10.1186/1741-7007-7-72

11. Sire L, Gey D, Debruyne R et al (2019) The challenge of DNA barcoding saproxylic beetles in natural history collections—exploring the potential of parallel multiplex sequencing with illumina MiSeq. Front Ecol Evol. https://doi.org/10.3389/fevo.2019.00495

12. Pawlowski J, Audic S, Adl S et al (2012) CBOL Protist Working Group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. PLoS Biol 10:e1001419. https://doi.org/10.1371/journal.pbio.1001419

13. Ratnasingham S, Hebert PDN (2007) bold: the barcode of life data system. Mol Ecol Notes 7:355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

14. Mueller RL (2006) Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. Syst Biol 55:289–300. https://doi.org/10.1080/10635150500541672

15. Hebert PDN, Ratnasingham S, de Waard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc Lond B 270:S96–S99. https://doi.org/10.1098/rsbl.2003.0025

16. Hussain K, Rashid K, Hafeez F et al (2020) Molecular identification of sugarcane black bug (Cavelarius excavates) from Pakistan using cytochrome C oxidase I (COI) gene as DNA barcode. Int J Trop Insect Sci 40:1119–1124. https://doi.org/10.1007/s42690-020-00144-5

17. Jackson A, Nijman V (2020) DNA barcoding of primates and the selection of molecular markers using African Great Apes as a model. J Anthropol Sci. https://doi.org/10.4436/JASS.98017

18. Group CPW, Hollingsworth PM, Forrest LL et al (2009) A DNA barcode for land plants. PNAS 106:12794–12797. https://doi.org/10.1073/pnas.0905845106

19. Soledispa P, Santos-Ordóñez E, Miranda M et al (2021) Molecular barcode and morphological analysis of Smilax purhampuy Ruiz. Ecuador PeerJ 9:e11028. https://doi.org/10.7717/peerj.11028

20. Ahmed S, Ibrahim M, Nantasenamat C et al (2022) Pragmatic applications and universality of DNA barcoding for substantial organisms at species level: a review to explore a way forward. Biomed Res Int 2022:e1846485. https://doi.org/10.1155/2022/1846485

21. Bradshaw M, Grewe F, Thomas A et al (2020) Characterizing the ribosomal tandem repeat and its utility as a DNA barcode in lichen-forming fungi. BMC Evol Biol 20:2. https://doi.org/10.1186/s12862-019-1571-4

22. Stielow JB, Lévesque CA, Seifert KA et al (2015) One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. Persoonia 35:242–263. https://doi.org/10.3767/003158515X689135

23. Aoki W, Watanabe M, Watanabe M et al (2020) Discrimination between edible and poisonous mushrooms among Japanese *Entoloma sarcopum* and related species based on phylogenetic analysis and insertion/deletion patterns of nucleotide sequences of the cytochrome oxidase 1 gene. Genes Genet Syst. https://doi.org/10.1266/ggs.19-00032

24. Schoch CL, Seifert KA, Huhndorf S et al (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. PNAS 109:6241–6246. https://doi.org/10.1073/pnas.1117018109

25. Adam PS, Borrel G, Brochier-Armanet C, Gribaldo S (2017) The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J 11:2407–2425. https://doi.org/10.1038/ismej.2017.122

26. Chaban B, Hill JE (2012) A 'universal' type II chaperonin PCR detection system for the investigation of Archaea in complex microbial communities. ISME J 6:430–439. https://doi.org/10.1038/ismej.2011.96

27. Zeigler DRY (2003) Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int J Syst Evol Microbiol 53:1893–1900. https://doi.org/10.1099/ijs.0.02713-0

28. Liu Y-F, Mbadinga SM, Gu J-D, Mu B-Z (2017) Type II chaperonin gene as a complementary barcode for 16S rRNA gene in study of Archaea diversity of petroleum reservoirs. Int Biodeterior Biodegrad 123:113–120. https://doi.org/10.1016/j.ibiod.2017.04.015

29. Bukin YS, Galachyants YP, Morozov IV et al (2019) The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci Data 6:190007. https://doi.org/10.1038/sdata.2019.7

30. Ogier J-C, Pagès S, Galan M et al (2019) rpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. BMC Microbiol 19:171. https://doi.org/10.1186/s12866-019-1546-z

31. Purty RS, Chatterjee S (2016) DNA barcoding: an effective technique in molecular taxonomy. Austin J Biotechnol Bioeng 3:1059

32. Schneider KL, Marrero G, Alvarez AM, Presting GG (2011) Classification of plant associated bacteria using RIF, a computationally derived DNA marker. PLoS ONE 6:e18496. https://doi.org/10.1371/journal.pone.0018496

33. Vancuren SJ, Santos SJD, Hill JE, Team the MMLP (2020) Evaluation of variant calling for cpn60 barcode sequence-based microbiome profiling. PLoS ONE 15:e0235682. https://doi.org/10.1371/journal.pone.0235682

34. Zhan Z, Li J, Xu K (2019) Ciliate Environmental diversity can be underestimated by the V4 Region of SSU rDNA: insights from species delimitation and multilocus phylogeny of pseudokeronopsis (Protist, Ciliophora). Microorganisms 7:493. https://doi.org/10.3390/microorganisms7110493

35. Lin S, Hu Z, Deng Y et al (2020) An assessment on the intrapopulational and intraindividual genetic diversity in LSU rDNA in the harmful algal blooms-forming dinoflagellate Margalefidinium (= Cochlodinium) fulvescens based on clonal cultures and bloom samples from Jiaozhou Bay. China Harmful Algae 96:101821. https://doi.org/10.1016/j.hal.2020.101821

36. Zhang T, Fan X, Gao F et al (2019) Further analyses on the phylogeny of the subclass Scuticociliatia (Protozoa, Ciliophora) based on both nuclear and mitochondrial data. Mol Phylogenet Evol 139:106565. https://doi.org/10.1016/j.ympev.2019.106565

37. Zhao Y, Yi Z, Gentekaki E et al (2016) Utility of combining morphological characters, nuclear and mitochondrial genes: An attempt to resolve the conflicts of species identification for ciliated protists. Mol Phylogenet Evol 94:718–729. https://doi.org/10.1016/j.ympev.2015.10.017

38. Stern RF, Andersen RA, Jameson I et al (2012) Evaluating the ribosomal internal transcribed spacer (ITS) as a candidate dinoflagellate barcode marker. PLoS ONE 7:e42780. https://doi.org/10.1371/journal.pone.0042780

39. Choi J, Park JS (2020) Comparative analyses of the V4 and V9 regions of 18S rDNA for the extant eukaryotic community using the Illumina platform. Sci Rep 10:6519. https://doi.org/10.1038/s41598-020-63561-z

40. Mordret S, Piredda R, Vaulot D et al (2018) dinoref: a curated dinoflagellate (Dinophyceae) reference database for the 18S rRNA gene. Mol Ecol Resour 18:974–987. https://doi.org/10.1111/1755-0998.12781

41. Zimmermann J, Jahn R, Gemeinholzer B (2011) Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. Org Divers Evol 11:173. https://doi.org/10.1007/s13127-011-0050-6

42. Forster D, Filker S, Kochems R et al (2019) A comparison of different ciliate metabarcode genes as bioindicators for environmental impact assessments of salmon aquaculture. J Eukaryot Microbiol 66:294–308. https://doi.org/10.1111/jeu.12670

43. Hamsher SE, Evans KM, Mann DG et al (2011) Barcoding diatoms: exploring alternatives to COI-5P. Protist 162:405–422. https://doi.org/10.1016/j.protis.2010.09.005

44. Cuypers B, Domagalska MA, Meysman P et al (2017) Multiplexed spliced-leader sequencing: a high-throughput, selective method for RNA-seq in trypanosomatids. Sci Rep 7:3725. https://doi.org/10.1038/s41598-017-03987-0

45. Abraham JS, Sripoorna S, Maurya S et al (2019) Techniques and tools for species identification in ciliates: a review. Int J Syst Evol Microbiol 69:877–894. https://doi.org/10.1099/ijsem.0.003176

46. Strüder-Kypke MC, Lynn DH (2010) Comparative analysis of the mitochondrial cytochrome c oxidase subunit I (COI) gene in ciliates (Alveolata, Ciliophora) and evaluation of its suitability as a biodiversity marker. Syst Biodivers 8:131–148. https://doi.org/10.1080/14772000903507744

47. Wang P, Gao F, Huang J et al (2015) A case study to estimate the applicability of secondary structures of SSU-rRNA gene in taxonomy and phylogenetic analyses of ciliates. Zool Scr 44:574–585. https://doi.org/10.1111/zsc.12122

48. Shazib SUA, Vďačný P, Kim JH et al (2016) Molecular phylogeny and species delimitation within the ciliate genus Spirostomum (Ciliophora, Postciliodesmatophora, Heterotrichea),

using the internal transcribed spacer region. Mol Phylogenet Evol 102:128–144. https://doi.org/10.1016/j.ympev.2016.05.041

49. Sun P, Clamp JC, Xu D et al (2013) An ITS-based phylogenetic framework for the genus Vorticella: finding the molecular and morphological gaps in a taxonomically difficult group. Proc R Soc B 280:20131177. https://doi.org/10.1098/rspb.2013.1177

50. Wylezich C, Nies G, Mylnikov AP et al (2010) An evaluation of the use of the LSU rRNA D1–D5 domain for DNA-based taxonomy of eukaryotic protists. Protist 161:342–352. https://doi.org/10.1016/j.protis.2010.01.003

51. Santoferrara LF, McManus GB, Alder VA (2013) Utility of genetic markers and morphology for species discrimination within the order Tintinnida (Ciliophora, Spirotrichea). Protist 164:24–36. https://doi.org/10.1016/j.protis.2011.12.002

52. Stoeck T, Przybos E, Dunthorn M (2014) The D1–D2 region of the large subunit ribosomal DNA as barcode for ciliates. Mol Ecol Resour 14:458–468. https://doi.org/10.1111/1755-0998.12195

53. Tasneem F, Shakoori FR (2017) Phylogenetic relationship of locally isolated paramecium species inferred from histone H4 genes. Pak J Zool. https://doi.org/10.17582/journal.pjz/2017.49.5.1767.1774

54. Santoferrara L, Burki F, Filker S et al (2020) Perspectives from ten years of protist studies by high-throughput metabarcoding. J Eukaryot Microbiol 67:612–622. https://doi.org/10.1111/jeu.12813

55. Grant DM, Brodnicke OB, Evankow AM et al (2021) The future of DNA barcoding: reflections from early career researchers. Diversity 13:313. https://doi.org/10.3390/d13070313

56. Gong S, Ding Y, Wang Y et al (2018) Advances in DNA barcoding of toxic marine organisms. Int J Mol Sci 19:2931. https://doi.org/10.3390/ijms19102931

57. Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. Trends Ecol Evol 24:110–117. https://doi.org/10.1016/j.tree.2008.09.011

58. Liu M, Clarke LJ, Baker SC et al (2020) A practical guide to DNA metabarcoding for entomological ecologists. Ecol Entomol 45:373–385. https://doi.org/10.1111/een.12831

59. Braukmann TWA, Ivanova NV, Prosser SWJ et al (2019) Metabarcoding a diverse arthropod mock community. Mol Ecol Resour 19:711–727. https://doi.org/10.1111/1755-0998.13008

60. Medlin LK, Orozco J (2017) Molecular techniques for the detection of organisms in aquatic environments, with emphasis on harmful algal bloom species. Sensors 17:1184. https://doi.org/10.3390/s17051184

61. Kappel K, Eschbach E, Fischer M, Fritsche J (2020) Design of a user-friendly and rapid DNA microarray assay for the authentication of ten important food fish species. Food Chem 311:125884. https://doi.org/10.1016/j.foodchem.2019.125884

62. Kochzius M, Seidel C, Antoniou A et al (2010) Identifying fishes through DNA barcodes and microarrays. PLoS ONE 5:e12620. https://doi.org/10.1371/journal.pone.0012620

63. Teletchea F, Bernillon J, Duffraisse M et al (2008) Molecular identification of vertebrate species by oligonucleotide microarray in food and forensic samples. J Appl Ecol 45:967–975. https://doi.org/10.1111/j.1365-2664.2007.01415.x

64. Liao Y-C, Cheng H-W, Wu H-C et al (2019) Completing circular bacterial genomes with assembly complexity by using a sampling strategy from a single MinION run with barcoding. Front Microbiol 10:2068. https://doi.org/10.3389/fmicb.2019.02068

65. Wang WY, Srivathsan A, Foo M et al (2018) Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: validating a reverse workflow for specimen processing. Mol Ecol Resour 18:490–501. https://doi.org/10.1111/1755-0998.12751

66. Li R, Xie M, Dong N et al (2018) Efficient generation of complete sequences of MDR-encoding plasmids by rapid assembly of MinION barcoding sequencing data. GigaScience. https://doi.org/10.1093/gigascience/gix132

67. Parker J, Helmstetter AJ, Devey D et al (2017) Field-based species identification of closely-related plants using real-time nanopore sequencing. Sci Rep 7:8345. https://doi.org/10.1038/s41598-017-08461-5

68. Srivathsan A, Baloğlu B, Wang W et al (2018) A MinION™-based pipeline for fast and cost-effective DNA barcoding. Mol Ecol Resour. https://doi.org/10.1111/1755-0998.12890

69. Burton AS, Stahl SE, John KK et al (2020) Off earth identification of bacterial populations using 16S rDNA nanopore sequencing. Genes 11:76. https://doi.org/10.3390/genes11010076

70. Zhang H, Xu H, Liu H et al (2020) PacBio single molecule long-read sequencing provides insight into the complexity and diversity of the Pinctada fucata martensii transcriptome. BMC Genomics 21:481. https://doi.org/10.1186/s12864-020-06894-3

71. Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13:278–289. https://doi.org/10.1016/j.gpb.2015.08.002

72. Tedersoo L, Tooming-Klunderud A, Anslan S (2018) PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytol 217:1370–1385. https://doi.org/10.1111/nph.14776

73. Yssouf A, Almeras L, Raoult D, Parola P (2016) Emerging tools for identification of arthropod vectors. Fut Microbiol 11:549–566. https://doi.org/10.2217/fmb.16.5

74. Dridi B, Raoult D, Drancourt M (2012) Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry identification of Archaea: towards the universal identification of living organisms. APMIS 120:85–91. https://doi.org/10.1111/j.1600-0463.2011.02833.x

75. Munir S, Ahmed S, Ibrahim M et al (2020) A spellbinding interplay between biological barcoding and nanotechnology. Front Bioeng Biotechnol 8:883. https://doi.org/10.3389/fbioe.2020.00883

76. Yin H, Jia M, Yang S et al (2012) A nanoparticle-based bio-barcode assay for ultrasensitive detection of ricin toxin. Toxicon 59:12–16. https://doi.org/10.1016/j.toxicon.2011.10.003

77. Amini B, Kamali M, Salouti M, Yaghmaei P (2017) Fluorescence bio-barcode DNA assay based on gold and magnetic nanoparticles for detection of Exotoxin A gene sequence. Biosens Bioelectron 92:679–686. https://doi.org/10.1016/j.bios.2016.10.030

78. Ding S, Wang L, He Z et al (2021) Identifying exogenous DNA in liquid foods by gold nanoparticles: potential applications in traceability. ACS Food Sci Technol 1:605–613. https://doi.org/10.1021/acsfoodscitech.1c00048

79. Valentini P, Galimberti A, Mezzasalma V et al (2017) DNA barcoding meets nanotechnology: development of a universal colorimetric test for food authentication. Angew Chem Int Ed 56:8094–8098. https://doi.org/10.1002/anie.201702120

80. Taboada L, Sánchez A, Pérez-Martín RI, Sotelo CG (2017) A new method for the rapid detection of Atlantic cod (*Gadus morhua*), Pacific cod (*Gadus macrocephalus*), Alaska pollock (*Gadus chalcogrammus*) and ling (*Molva molva*) using a lateral flow dipstick assay. Food Chem 233:182–189. https://doi.org/10.1016/j.foodchem.2017.04.087

81. Schlenker C, Brooks J, Oostra K, McLaughlin R (2017) whole sample next-generation DNA sequencing method: an alternative to DNA barcoding. In: FoodSafetyTech. https://foodsafetytech.com/feature_article/whole-sample-next-generation-dna-sequencing-method-alternative-dna-barcoding/. Accessed 19 Jul 2022

82. Crampton-Platt A, Yu DW, Zhou X, Vogler AP (2016) Mitochondrial metagenomics: letting the genes out of the bottle. GigaScience. https://doi.org/10.1186/s13742-016-0120-y

83. Jiang M, Xu S-F, Tang T-S et al (2022) Development and evaluation of a meat mitochondrial metagenomic (3MG) method for

composition determination of meat from fifteen mammalian and avian species. BMC Genomics 23:36. https://doi.org/10.1186/s12864-021-08263-0

84. Ji Y, Huotari T, Roslin T et al (2020) SPIKEPIPE: a metagenomic pipeline for the accurate quantification of eukaryotic species occurrences and intraspecific abundance change using DNA barcodes or mitogenomes. Mol Ecol Resour 20:256–267. https://doi.org/10.1111/1755-0998.13057

85. Yang Z, Rannala B (2016) Species identification by bayesian fingerprinting: a powerful alternative to DNA barcoding. 041608

86. Obringer R, Rachunok B, Maia-Silva D et al (2021) The overlooked environmental footprint of increasing Internet use. Resour Conserv Recycl 167:105389. https://doi.org/10.1016/j.resconrec.2020.105389

87. Liu Y, Xu C, Sun Y et al (2021) Method for quick DNA barcode reference library construction. Ecol Evol 11:11627–11638. https://doi.org/10.1002/ece3.7788

88. Guo J, Cheng T, Xu H et al (2019) An efficient and cost-effective method for primer-induced nucleotide labeling for massive sequencing on next-generation sequencing platforms. Sci Rep 9:3125. https://doi.org/10.1038/s41598-019-38996-8

89. Marine RL, Magaña LC, Castro CJ et al (2020) Comparison of Illumina MiSeq and the Ion Torrent PGM and S5 platforms for whole-genome sequencing of picornaviruses and caliciviruses. J Virol Methods 280:113865. https://doi.org/10.1016/j.jviromet.2020.113865

90. Speranskaya AS, Khafizov K, Ayginin AA et al (2018) Comparative analysis of Illumina and Ion Torrent high-throughput sequencing platforms for identification of plant components in herbal teas. Food Control 93:315–324. https://doi.org/10.1016/j.foodcont.2018.04.040

91. Janzen DH, Hallwachs W, Pereira G et al (2020) Using DNA-barcoded Malaise trap samples to measure impact of a geothermal energy project on the biodiversity of a Costa Rican old-growth rain forest. Genome 63:407–436. https://doi.org/10.1139/gen-2020-0002

92. Martínez-Arce A, De Jesús-Navarrete A, Leasi F (2020) DNA barcoding for delimitation of putative mexican marine nematodes species. Diversity 12:107. https://doi.org/10.3390/d12030107

93. Hou F, Wen L, Peng C, Guo J (2018) Identification of marine traditional Chinese medicine dried seahorses in the traditional Chinese medicine market using DNA barcoding. Mitochondrial DNA Part A 29:107–112. https://doi.org/10.1080/24701394.2016.1248430

94. Minoudi S, Karaiskou N, Avgeris M et al (2020) Seafood mislabeling in Greek market using DNA barcoding. Food Control 113:107213. https://doi.org/10.1016/j.foodcont.2020.107213

95. Burrell AS, Jolly CJ, Tosi AJ, Disotell TR (2009) Mitochondrial evidence for the hybrid origin of the kipunji, *Rungwecebus kipunji* (Primates: Papionini). Mol Phylogenet Evol 51:340–348. https://doi.org/10.1016/j.ympev.2009.02.004

96. Osathanunkul M, Suwannapoom C, Osathanunkul K et al (2016) Evaluation of DNA barcoding coupled high resolution melting for discrimination of closely related species in phytopharmaceuticals. Phytomedicine 23:156–165. https://doi.org/10.1016/j.phymed.2015.11.018

97. Carbone I, Kohn LM (1999) A method for designing primer sets for speciation studies in filamentous ascomycetes. Mycologia 91:553–556. https://doi.org/10.1080/00275514.1999.12061051

98. Rehner SA, Buckley E (2005) A Beauveria phylogeny inferred from nuclear ITS and EF1-α sequences: evidence for cryptic diversification and links to Cordyceps teleomorphs. Mycologia 97:84–98. https://doi.org/10.1080/15572536.2006.11832842

99. Siddiqui ZH, Abbas ZK, Hakeem KR et al (2020) A molecular assessment of red algae with reference to the utility of DNA barcoding. In: Trivedi S, Rehman H, Saggu S et al (eds) DNA barcoding and molecular phylogeny. Springer, Cham, pp 103–118