DNA Barcoding in Plants: Past, Present, and Future

13

Pradosh Mahadani, Majusha Dasgupta, Joshitha Vijayan, Chandan Sourav Kar, and Soham Ray

Abstract

DNA barcoding is a method of identifying biological specimens and assigning them to their respective species. It involves sequencing of single/multiple short stretch/stretches of previously agreed-upon genomic region which evolves fast enough to allow species-level discrimination. Thus, obtained sequence(s) of unknown samples serve as a molecular identifier which is compared to a reference database of museum samples using specialized algorithms to reveal the identity of the specimen under study. In effect it complements classical taxonomy to quickly identify any newly obtained sample and aid in describing, naming, and classifying it to species. Unlike in animals where DNA barcoding is well standardized utilizing mitochondrial gene *CO1*, DNA barcoding in plants has perpetually been a matter of concern due to low substitution rates of plant mitochondrial genome. Alternatively, plastid genome has been targeted in case of plants for DNA barcoding purpose with some amount of success but ambiguities remain regarding selection of barcode region that can provide best possible resolution. A large number of studies tested the efficiency of seven leading candidate plastid DNA regions (*matK, rbcL, rpoB, rpoC*1 genes and *atpF–atpH*, *psbK–psbI*, *trnH–psbA spacers*) as the standard DNA barcode for plants. Based on universality, sequence quality, and species discrimination rate, a

P. Mahadani
Bioinformatics Centre, National Tea Research Institute, Kolkata, India

M. Dasgupta
Department of Biotechnology, Gauhati University, Jalukbari, Guwahati, Assam, India

J. Vijayan
ICAR-Central Island Agricultural Research Institute, Port Blair, Andaman & Nicobar Islands, India

C. S. Kar · S. Ray (✉)
ICAR-Central Research Institute for Jute & Allied Fibres, Kolkata, West Bengal, India

R. L. Singh et al. (eds.), *Plant Genomics for Sustainable Agriculture*,
https://doi.org/10.1007/978-981-16-6974-3_13

double-locus barcode (*rbcL+ matK*) system is suggested to perform best in plants. However, internal transcribed spacer (ITS) region of plastid gene *ycf1* has recently been suggested as the most promising single-locus plant DNA barcode. On contrary, a recent study argues that with an ever-growing sequence database even double-locus barcode (*rbcL+ matK*) system might become unfit for precise discrimination purpose. Hence, with the availability of next-generation sequencing technologies, partial genome representation-based barcoding, genome skimming based barcoding, full-length multi-barcoding (FLMB), etc. might be the preferred approaches to improve diagnostic power. DNA barcoding in plants not only speeds up writing the encyclopedia of life, but also opens up the possibility of establishing Digital Plant Identification System (DPIS) which works independent of type, age, or developmental stage of the sample under study. Hence, if used properly, DNA barcoding can be an effective and efficient tool for exploring and protecting biodiversity, expedite bioprospecting, and defending against bio-piracy.

Keywords

DNA Barcoding · ITS · *matK* · Species identification

13.1 Introduction

Plant biodiversity is a product of natural evolution and selection pressure that has sustained humankind's security for thousands of years. It is recognized as a valuable gene pool for various traits, which may stand as a potential solution in the face of rising environmental and anthropogenic challenges. Still, a vast amount of biodiversity remains undiscovered to the world. In this aspect, traditional taxonomy has been serving to identify and classify species over many years. As many valuable species and gene pools face the risk of extinction, the first step in preserving biodiversity is assessment. Recent estimates suggest that around 70,000 flowering plant species await to be discovered (Gross 2011). Unfortunately, there are not enough taxonomists to catalog species throughout the world. DNA taxonomy and DNA barcoding are accessory technologies that have helped speed up the process and emerged as a conservation practice tool. By harnessing advances in molecular genetics, sequencing technology, and bioinformatics, DNA barcoding was initially proposed by (Hebert et al. 2003)and has emerged as a vital new tool for taxonomists who take care of inventory and management of our planet's immense and changing biodiversity (Kress and Erickson 2008). DNA barcoding equips the taxonomist with the ability to quickly and cheaply (relatively) provide diagnostic identifications of species present in specific locations with immediate conservation and environment-related implications. Therefore, this diagnostic tool was developed as an aid to the taxonomic identification of species. It uses a standardized DNA region from the genome, which ideally has sufficient sequence variation to discriminate among species (CBOL et al. 2009). It has been advocated as a more efficient approach

than traditional taxonomic practices (Blaxter 2004; Tautz et al. 2003). The classical techniques of plant identification involving the conventional keys are tricky and time-consuming. As it involves micro-and macroscopic characters as well as chemical profiling, which did not evolve successfully. In this aspect, DNA barcoding has rapidly achieved recognition as an essential tool with the power to aid much basic research and applied endeavors in taxonomy and species identification (Hajibabaei et al. 2007; Hebert et al. 2003; Savolainen et al. 2005).

13.2 The Genesis of Concept

In 2003, Paul D. N. Hebert, a professor at the University of Guelph, Ontario, Canada, for the first time proposed the concept of DNA barcoding with an announcement that it would serve as a basic tool for species identification of global biological samples. His announcement is based on his observation and analysis with the class Hexapoda, representing the greatest biodiversity on the planet. The technique involved selective amplification of only 648 bp of mitochondrial *Cytochrome Oxidase SubunitI* (COI) gene near its 5′end. He coined this segment DNA barcode for species-level identification. He justified its universality based on rapid evolution properties of the COI gene and variability properties of A, T, G, and C nucleotides of DNA. He argued that integrating DNA barcodes into traditional taxonomic tools could efficiently reveal unexplored biodiversity more swiftly and more securely in an authenticated way than traditional methods alone. Since the genesis, it has been successfully used for rapid biodiversity assessment studies, bio-monitoring, investigation of the illegal trade of endangered species, feeding ecology studies and for conservation of medicinal and poisonous plants, etc. (Muellner et al. 2011; Hollingsworth et al. 2011). The use of nucleotide sequence variations to investigate biodiversity, however, is not a new concept. It has been long realized that the changes in the four nucleotides A, T, G, and C set the backbone of molecular evolution, leading to discrete variation patterns among organisms. Thus, during the evolution, initial changes accumulate at the molecular level, which in the long-term lead to visible morphological variations. However, even if two organisms are morphologically alike, they may bear substantial variation at the molecular level, and the phenotypic similarity or dissimilarity between organisms is not a true reflection of actual variation. This dilemma often leads to misinterpretation and is a major drawback in biodiversity research where morphological keys are the basis. Several enthusiasts who were inclined to explore variation at the molecular level proposed using nucleotide segments, genes, rDNA, allozymes, etc., as markers to characterize organisms. However, the propositions were mainly suitable for a group of organisms while lacking broad range utility or universal application. DNA barcoding is a comparatively easy, quick, reprove able, universal approach for species identification. The principal requirement for barcoding is judicious locus/loci for DNA barcoding and should be prioritized and standardized so that large databases of sequences for that locus can be generated. Sequences are able to generate without species-specific PCR primers from the taxa

of interest. Three essential principles of DNA barcoding are standardization, minimalism, and extensibility.

The leading DNA barcoding bodies and resources are (1) Consortium for the Barcode of Life (CBOL) http://www.barcodeoflife.org established in 2004. Worldwide DNA barcoding efforts have resulted in the formation of CBOL which promotes DNA barcoding through more than 200 member organizations from 50 countries, (2) International Barcode of Life (iBOL) http://www.ibol.org launched in October 2010, iBOL represents a not -for-profit effort to involve both developing and developed countries in the global barcoding effort, establishing commitments and working groups in 25 countries. It is the largest biodiversity genomics initiative ever undertaken, which maintains a barcode reference library, (3) The Barcode of Life Data systems (BOLD) consists of different institutional nodes from several nations clustered into separate working groups which works coordinately for the development of a specialized repository for DNA barcode sequences and has emerged as a global bio-identification system for species. BOLD is a web-based system for DNA barcodes, combines a barcode repository, analytical tools, an interface for submission of sequences to GenBank, a species identification tool and connectivity for external web developers and bioinformaticians (Ratnasingham and Hebert 2007). As of 2017, BOLD included over 5.9 million DNA barcode sequences from over 542,000 species.

13.3 Technical Know-How of DNA Barcoding

The development of reference data sources for each taxa of the world and thus creation of a reference database is an important step in the realm of barcoding research. It involves either mass participation of renowned taxonomists across the globe for the construction of a sound reference database. Another way by which this is achieved is by focusing on the museum specimens identified by various experts and using their barcode sequences as the standards or references for those taxa. However, all resources are not cataloged in museums and hence, new collections and explorations are also considered vital. As museum specimens maintain some standard data, new collections of specimens were undertaken maintaining some standards records such as collector name, collection date, geographical location, elevation/depth, collection gear, notes on habitat and microhabitat, sex of specimen, life stage, specimen imaging, identifier, etc. Practically cataloguing the total biodiversity of Earth in a museum is not feasible and even if it can be done gradually with time, the specimens get deformed as no fixatives can guarantee total preservation of the samples. Under such circumstances, the specimen's information is maintained in a database. The second part of DNA barcoding involves access of the reference data by enthusiasts for subsequent analysis and interpretation. Additional favorable factors are short length of barcode loci facilitated routine sequencing, even with sub-optimal material, lack of heterozygosity enabling direct polymerase chain reaction followed by sequencing without cloning, ease of alignment that enables the use of character-based data analysis methods, lack of problematic sequence

composition, such as regions with several microsatellites, that reduces sequence quality, universal capability to get amplified/sequenced with standardized primers, easy align ability and capability to get recovered easily from herbarium samples and other degraded DNA samples (Hollingsworth et al. 2009). From the preparation of the data to the final analysis, DNA barcoding technology comprises several practical steps, which will be discussed below briefly.

13.3.1 Sampling

The DNA barcoding is a molecular concept, where focus is on the DNA molecules that remain embedded within each cell. Hence, sampling for DNA barcoding involves both specimen sampling and DNA sampling. Specimen sampling is done from a taxonomist point of view, where a complete coverage of morphological and geographical information is gathered. DNA sampling can be done from any tissue of the organism; however, the areas which bear the key morphological characters for the specimens are always kept intact. The specimen sampling is immediately followed by sampling of a small part of tissue for DNA sampling.

13.3.2 DNA Extraction, PCR Amplification, and Sequencing

The DNA molecule which is the principal component of DNA barcoding concept has to be extracted and preserved. Nowadays there are several technologies involved for isolation of DNA, however, the best technique is adopted which will keep the DNA intact. The isolation is followed by PCR amplification of the targeted DNA barcode segment using available published primers and then sequencing.

13.3.3 Analysis and Interpretation

PCR amplification and sequencing of the barcode segment are followed by analysis of the sequence using bioinformatics tools. The major part of the analysis involves checking the quality of the sequence and its maximum similarity with the reference database.

The prerequisite criteria for any DNA barcode loci are a large amount of sequence variation between species; however, variation should be low enough within species so that a gap between intra- and inter-specific genetic variations can be defined and also known as barcode gap. Besides that, conserved flanking regions for universal primer are required to enable routine amplification across highly divergent taxa. In practice, an unidentified organism's specific standardized portion DNA sequence acts as a repository signal which is compared to the reference sequences databases of known species. The similarity of sequence, i.e., unknown organism to one of the reference sequences leads to a rapid and reproducible identification. Some large group of linkage or association should exist as support for species monophyly and

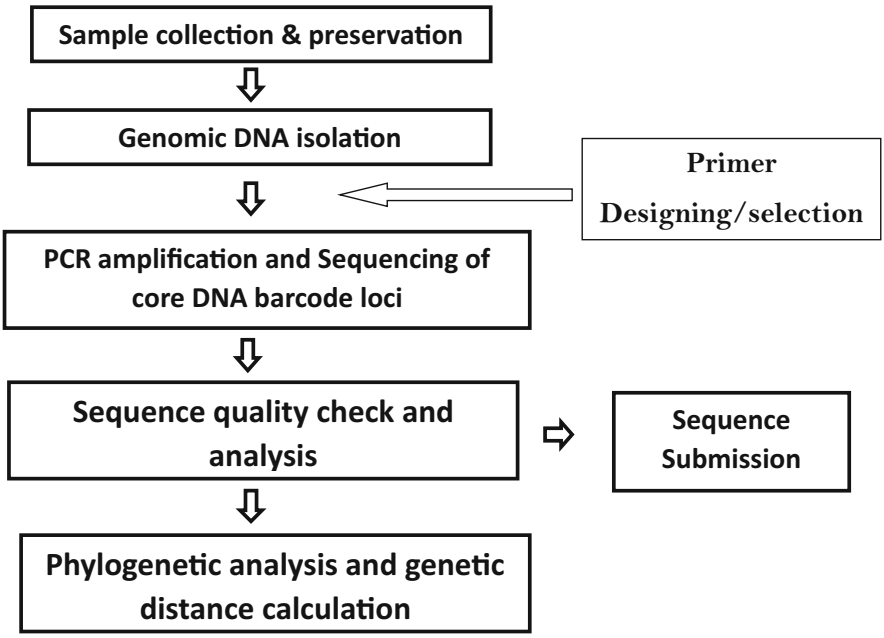

Fig. 13.1 Basic workflow of DNA barcoding approach

the ability of DNA barcoding marker systems to differentiate or distinguish species (Fig. 13.1).

13.4 Promising Plant DNA Barcoding Loci

Chloroplasts are organelles of prokaryotic origin and house of photosynthetic apparatus and also play a crucial role in sulfur and nitrogen metabolism. Plant DNA barcoding involves sequencing a standard region of the chloroplast genome as a tool for species identification. The chloroplast is a nearly autonomous organelle because it contains the biochemical machinery necessary to replicate and transcribe its own genome and carry out protein synthesis. The DNA of chloroplast is a circular that ranges in size from 120 to 190 kb depending on the species. The chloroplast genome is symbiotic in its origin from both algal and protistan lineages; its gene expression machinery is an assembly of prokaryotic, eukaryotic, and phage-like components, resulting in the acquisition of a significant number of regulatory proteins during evolution. Comparative evaluation indicates gene order and gene content of land plants chloroplast genomes are highly conserved. Traditionally, the plastid genome has been a more readily choice for phylogenetic studies in plants than the nuclear genome. As the chloroplast genome is maternally inherited, no recombination occurs, and, in general, structurally they are more stable with high copy number. Several candidate regions have been proposed as barcoding loci,

including some coding genes (*matK*, *rbcL*, *rpoB*, and *rpoC1*) non-coding region (*psbA-trnH*, *atpF-atpH*, *ycf*) or a combination of several regions.

Maturase K of the chloroplast genome is the most conserved gene in the plant kingdom and is involved in Group II intron splicing. *matK* gene sequence is about 1500 bp long and encodes maturase like protein. Due to the high substitution rates, *matK* is emerging as a potential candidate for DNA barcoding (Hilu and Liang 1997). The *matK*gene has ideal size, large proportion of variation at open reading frame level at first and second codon position. The *matK* gene is rapidly evolving and considered as a good DNA barcode region (Mahadani et al. 2013; Sun et al. 2012). Thus, *matK* sequence plays a vital role in phylogenetic and evolutionary studies. Lahaye et al. (2008) collected more than 1600 plant samples from Mesoamerica and southern Africa, biodiversity hotspot. This was the first large scale study to compare eight potential barcodes in all the samples. As a universal plant DNA barcode, Plastid *matK* gene showed easy amplification, alignment, discrimination power. In addition, analyzing >1000 species of Mesoamerican orchids, DNA barcoding with *matK* alone revealed cryptic species and proved useful in identifying species listed in Convention on International Trade of Endangered Species (CITES) appendixes.

Several researchers proposed *rbcL* as a potential plant barcode region, as large amounts of information are already available in the sequence databases. About 1300 bp long, *rbcL* sequences showed a fair degree of success in discriminating species (Newmaster et al.; 2006). But the *rbcL* marker, which is easy to amplify sequence and align, has a limited discrimination power, especially when among closely related species. The Consortium for the Barcode of Life (Plant Working Group) recognized a combination of *matK* and *rbcL* as the universal plant barcode (CBOL et al. 2009) although the levels of variation are sometimes low and insufficient to recognize species with these two specific markers. In large scale studies, these loci provide a discriminatory efficiency at the species level of 72% and 49.7%, respectively. In some instances, they have failed to differentiate closely related species (Hollingsworth et al. 2009). As a result, other chloroplast regions, e.g., *trnH-psbA, trnL, trnL-F* and the nuclear ribosomal Internal Transcribed Spacer (ITS) are routinely used in combination with *matK* and *rbcL*.

In higher plants, two plastidal RNA polymerases referred as plastid encoding polymerases or PEP (α-, β-, β′-, and β″-subunits) encoded by *rpoA*, *rpoB*, *rpoC1,* and *rpoC2* genes are promising candidates (Serino and Maliga 1998). In the chloroplast genomes, *ndhF* is located at one end of the small single-copy region and encoding the ND5 protein of chloroplast NADH dehydrogenase complex. *ndhF* contains more phylogenetic information than *rbcL* (Kim and Jansen 1995).

Among the non-coding region, *trnH-psbA* has highly conserved PCR priming sites, high numbers of substitutions and is often used as an additional marker, especially when DNA barcoding is applied to closely related plant taxa (Kress and Erickson 2007). However, mononucleotide repeats in the *trnH-psbA* region cause constraints in PCR and sequencing. Although *trnH-psbA* shows high levels of interspecific variation, it has found only limited use in species-level phylogenetic reconstruction due to its short length as well as difficulty of alignments resulting from a

high number of insertion and deletion (indels). The intergenic spacer *trnL-F* has a long history of use in studies on plant phylogenetic and species identification studies (Wallander and Albert, 2000). In some groups this region often contains ploy A and T structures and affects sequence quality (Shaw et al. 2005). In the chloroplast genome, the pseudogene, *ycf*1 is located in the boundary regions between IRb/S and IRa/SSC, respectively. *ycf1* is the first open reading frame coded by Tic214 (part of TIC core complex). The lack of their protein-coding ability is due to partial gene duplication. This gene is related to ATP synthase, and much more closely related to the *rbcL* gene with respect to its genetic structure. Recently, two regions of the plastid gene ycf1, ycf1a, and ycf1b were recognized as most variable loci in plastid. Dong et al. (2015) designed primers for amplification of these regions and analyzed the potential of these regions as DNA barcode in 420 tree species. The study showed ycf1a or ycf1b perform better than any of the *matK*, *rbcL*, and *trnH-psbA* for a large group of plant taxa (Dong et al. 2015) (Fig. 13.2).

In case of plants, mitochondrial genes are poor candidates for species-level discrimination due to low rate of sequence change. As the plastid genome evolves very slowly relative to other genomes and shows intra-molecular recombination, more than one barcode is necessary to provide enough to work. Although it is widely accepted that a single ("universal") set of barcode regions should be adopted to establish a reference barcoding database for all plants. The seven plastid regions *rpo*C1, *rpo*B, *rbc*L, *mat*K, *trn*H-*psb*A, *atp*F-*atp*H, and *psb*K-*psb*I were evaluated in three divergent groups of land plants (Newmasteret al., 2006. Hollingsworth et al. 2009). Study reports that 92% to 96% of plant specimens can be distinguished by combining the two core barcode markers *rbcL* and *matK*. In general, the genes used in angiosperms are *matK*, *rpo*C1, *rpo*B, *acc*D, *YCF5*, and *ndh*J, whereas in non-angiosperms *mat*K, *rpo*C1, *rpo*B, *acc*D, and *ndh*J. The Plant researcher from Consortium for the Barcode of Life (CBOL) proposed additional combinations of non-coding and coding plastid. In plant systematics for phylogenetic, *rbcL* sequenced most commonly, followed by the *trnL-F* intergenic spacer, *matK*, *ndhF*, and *atpB-rbcL* has been suggested as a candidate for plant barcoding.

Internal transcribed spacer regions (*ITS*) of nuclear ribosomal DNA are often highly variable in angiosperms at the generic and species level and divergent copies are often present within single individuals. About 400–800 bp long ITS regions are the most commonly sequenced region among the nuclear ribosomal cistron regions (18S-5.8S-26S), across the plants defined barcode gap between inter- and intra-specific variations (Group C P B et al. 2011). Chen et al. 2010 reported that The ITS1 and ITS2 (each $<$300 bp) adjoining the 5.8 S locus have a higher degree of variation than the rRNA genes (Chen et al. 2010). These genes contain enough phylogenetic signal for discrimination of both plants and animals. The ITS2 in comparison with ITS1 is more suitable for amplification and sequencing due to its shorter length of the target region which is referred to as a mini-barcode. The ITS of nuclear DNA has been used as a target for analyzing fungal diversity in environmental samples, and has been selected as the standard marker for fungal DNA barcoding (Schoch et al., 2012).

Fig. 13.2 Schematic diagram of Chloroplast genome. The graphical map *Camellia sinensis* var. *assamica* (Accession No: JQ975030) was drawn using OGDRAW (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html)

13.5 Utility of Plant DNA Barcoding

The additional power of the DNA barcode speeds up writing the encyclopedia of life. It opens up the way to develop an electronic field guide which workes at all stages of life. It can deal with fragments, unmask the look-alikes, reduce ambiguity, democratize access, and thus sprouts a new leaves on the tree of life. Forensic investigators have also applied these plant DNA barcodes in the regulatory areas of traffic in endangered species and monitoring commercial products, such as foods and herbal supplements. Categories of use include species-level taxonomy(Mahadani et al. 2013), biodiversity inventories (Lahaye et al. 2008; Hollingsworth 2008), phylogenetic evaluation (Hajibabaei et al. 2007), conservation assessment and

environmental preservation(Hollingsworth et al. 2011), species interactions and ecological networks (Erickson et al. 2008), cryptic diversity information (Fazekas et al. 2008), ecological forensics(Mishra et al. 2016), community assembly, traffic in endangered species, and monitoring of commercial products (Stoeckle et al. 2011; Mahadani and Ghosh 2013).

DNA barcoding assists in identification by expanding its power to detect species by including all life history stages of life, like pollen, seed, seedling and unstructured plant parts, etc. Kool et al. (2012) tested the functionality, efficiency, and accuracy of the use of barcoding for identifying 110 medicinal plant roots combining *rpc1*, *trnH-psbA*, and *ITS*. These three candidates identified the majority of samples up to the genus level. DNA barcoding helps to find out undiscovered species that are potentially new to encyclopedia of life (Kool et al. 2012).Over the last decade, four plant DNA barcode markers, viz. *rbcL*, *matK*, *trnH-psbA*, and ITS2, have been tested, and used to address many questions in systematics, ecology, evolutionary biology, and conservation. Mahadani et al. (2013) examined the sequences of core DNA barcode *matK* and *trnH-psbA* for differentiation of ethnomedicinal plants of family Apocynaceae from North east India. Among the selected medicinal Rauvolfioideae species, ~758 bp *matK* sequence showed easy amplification, alignment, and high level of discrimination value in comparison to the *trnH-psbA* spacer sequences. The partial *matK* sequences exhibited 3 indels in multiple of 3 at 5′ ends, but clustered cohesively, with their conspecific Genbank sequences. The possessing indels in multiple of 3 could be utilized as molecular markers in further studies both at the intra-specific and inter-specific levels (Mahadani et al. 2013).

Reliable identification of plant material by regulatory authorities is often of vital essence. Their domain includes identification of pests, pathogens and invasive species, illegal trades, identifying food or herbal medicine labeling errors/fraud. In this aspect, DNA barcoding approaches to assess the plant components of herbal medicines and dietary supplements, and evidence of market adulteration have been reported from many findings. A large array of commercial tea products were first time authenticated through *rbcL* and *matK* barcode sequences. Matching DNA identification to listed ingredients was limited by incomplete databases for the two markers, shared or nearly identical barcode among some species and lack of standard names for the plant species. About 1/3 of herbal tea generated DNA identification were not found on levels. This study demonstrated the importance of plant barcoding (Stoeckle et al. 2011). Six *Sabia* species and their seven adulterants were investigated DNA barcodes (*trnH-psbA*, *rbcL-a*, *matK*). Based on sequence alignments, they concluded that not only *trnH-psbA* spacer sequence distinguished *S. parviflora* from other *Sabia* species, but the *matK*+*rbcL-a* sequence also differentiated it from the substitute and adulterants. The three candidate barcodes identified *S. parviflora* and distinguished it from common substitutes or adulterants (Sui et al. 2011).

In traditional taxonomy-based identification, as seedlings, roots, seeds, and pollen and other gametophytes of many species appear similar, it is difficult to identify species from individual tissue types/juvenile life stages. Thus, with paleo barcoding, even barcode datasets with imperfect species resolution can still provide knowledge

gains. Moreover, the field of pollen barcoding is growing rapidly, and even modest increases in discriminatory power beyond morphological identification holds great promise to enhance understanding of the dynamics and consequences of pollination and pollen movement (Bell et al. 2016).

13.6 Challenges of Plant Barcodes

Several factors can potentially contribute towards a lack of unique species identification with DNA barcodes. To successfully implement DNA barcoding, sufficient time since speciation is required for point mutations or genetic drift. It leads to developing of a set of genetic characters that "group" or outgroup conspecific individuals are together unique from other species. In phylogenetic evaluation, barcode sequences are shared among related taxa or species in clades where speciation has been very recent. These problematic scenarios arise mainly in groups like woody species with long generation times and/or slow mutation rates and groups with evidence of recent radiation. Composition of monophyletic species is more in animal (>90%) than plants (~70%) using barcode markers. Both animals and plant systems have barcode gaps based on intra- and inter-specific genetic distances. However, animal species showed larger barcode gaps than plants. However, overall fine scale species discrimination in plants is relatively more difficult than animals because species boundaries are less well defined (Fazekas et al. 2008). Polyploid speciation may cause divergence between barcode sequences and taxon concepts where multiple allopolyploid species share a common parent species. In such cases, they may show similarity in plastid sequences, whereas independent origins of allopolyploid species can lead to taxa treated as conspecifics possessing divergent haplotypes. At least initially, plastid haplotype(s) with a diploid progenitor will be shared by the species that have originated by allo or autoploidy (Wang et al. 2018). The complexity of taxonomically complex groups (TCGs) cannot be solved using one or few markers, as these groups result from recurrent ecotypic origin of taxa, or recurrent ploidy transitions, apomixes, or recent hybrid speciation. Species discrimination success can be predicted by its dispersal ability and in that case an inverse correlation between intra- and inter-specific gene flow may rise. In case of *Solanum* sect. *Petota* (wild potatoes), *ITS*, *trnH-psbA*, *matK* regions showed too much intraspecific variation and lacked sufficient polymorphism in plastid markers (Spooner 2009). The universal barcode concept in plants is not working in Indian *Berberis* and two other genera, *Ficus* and *Gossypium*. Even the most promising plant DNA barcode loci (one from nuclear genome—*ITS*, and three from plastid genome—*trnH-psbA, rbcL*, and *matK*) failed to resolve species identification in these plant groups (Roy et al. 2010). Mahadani and Ghosh (2014) provide an alternative approach to identify the species using indel polymorphism as a species-level marker in Citrus.

13.7 Prospect of DNA Barcoding

The major challenge for DNA barcoding in plants is to achieve the proportion of unique species identifications. The selection of markers often depends on the nature of the application or research queries. For instance, single specimen-based studies tend to use a blend of the traditional DNA markers, while to recover a higher number of taxa from degraded or mixed DNA samples, metabarcoding approaches are taken which aims for shorter, easy to amplify fragments. The criteria of using multiple loci or multi-tiered system increase sample handling, preparation time, and costs in plant barcoding. Various limitations of traditional plant DNA barcoding has also been overcome by the advancement of high-throughput sequencing technologies which expedited the progress of plant genomics, particularly chloroplast genomics. Recently complete chloroplast genomes have also been shown to discriminate closely related species successfully. Until now, most DNA barcoding methods follow a traditional PCR-based approach followed by dideoxy chain termination (Sanger)-based sequencing. Alternatively, next-generation sequencing (NGS) technologies which decrease the cost of sequencing solve the problem partially by sequencing large portions of genomes (genome skimming) or whole genomes (organellar or otherwise "genome skimming" approaches (Coissac et al. 2016; Li et al. 2015). Short universally primed amplicon is ideal for sequence characterization through new parallelized high-throughput sequencing technologies, allowing inexpensive but comprehensive studies of biodiversity to be a realistic goal. These methods generate millions of sequence reads in a single run, they are still expensive for many research groups with regard to consumables, informatics, computational power, and data storage. In this aspect, though traditional Sanger-based sequencing technology is more expensive than next-generation sequencing (NGS) and is generally hampered by the need for relatively high target amplicon yield, coamplification of nuclear mitochondrial pseudogenes, confusions with sequences from intracellular endosymbiotic bacteria and instances of intracellular variability (i.e., heteroplasmy). Due to all these limitations, the high-throughput technology of next-generation sequence-based DNA barcoding has recently showed promising outcome for the elucidation of plant genetic diversity and its conservation.

13.8 Next-Generation Sequencing and DNA Barcoding

NGS technology allows for the sequencing of millions of DNA fragments from thousands of DNA templates in parallel and produces millions of short reads. NGS is a term loosely applied to the set of technologies used for genome-scale sequencing, viz. Roche 454, Illumina, Ion Torrent, SMRT, etc. It finds vast implementation, because of its protocol simplicity, reduced cost per read, high throughout and added information, sequencing sensitivity and accuracy by enabling the simultaneous detection of co-amplified products such as homologues, prologues, and contaminants. In this aspect, 454 pyro sequencing was the first NGS platform that came into the market. It permits the analysis of mixtures of DNA fragments that are

co-amplified during PCR or obtained by pooling different PCR products. Parallel sequencing of PCR amplicons is most effective when limited sequence data are targeted per specimen. NGS is also a powerful tool to detect numerous DNA sequence polymorphism based markers within a short timeframe and triggered numerous ground-breaking discoveries from many organisms (Van et al. 2013). The information that has emerged serves as a strong molecular tool for species exploration, progression, transformation studies, and the conservation of biodiversity.

13.9 DNA Barcode-Based High-Resolution Melting Curve Analysis (Bar-HRM)

In combination, high-resolution melting (HRM) analysis and DNA barcoding has emerged as a potential molecular method for plant species authentication, commonly known as Bar-HRM approach. The Bar-HRM has proven to be a reliable method for detection of contamination of different plant mixtures, particularly at the early stages of production like industrial quality control procedures for herbal medicines, etc. (Fernandes et al. 2020; Lee et al. 2019; Madesis et al. 2012). It is a novel DNA-based, cost-effective, and reliable quick identification assay that detects single base changes between samples. DNA dissociation ("melting") kinetics is monitored to detect the point mutations, indels, and methylated DNA.

The denaturation thermodynamics of individual double-stranded DNA to single strands is based on individual nucleotide pairs' binding affinities. Moreover, melting patterns will vary due to variations in product sizes, GC contents, and nucleotide composition, which vary due to indels, mutations, and methylations, inferred in terms of varying melting temperatures (T_m). In HRM in addition to standard PCR equipment and reagents, it requires a generic DNA intercalation fluorescent dye which is added to the previously amplified PCR products. As the double-stranded DNA samples dissociate with increasing temperature, the dye is progressively released and fluorescence diminishes. These differences of fluorescent measurements collected at standard temperature increments, which are plotted as a melting curve. So, variations in length, GC content, and base sequence will alter the melting profile defined by a plot. This plotted curve is generated between melting temperature and fluorescent level due to the release of intercalating SYBR Green I dye in a real-time PCR system. This melting curve's shape and peak are characteristic for individual specimen sample, allowing for comparison and discrimination among samples.

The HRM analysis has many advantages (1) As the sequencing is not required for Bar-HRM, which is generated by combining DNA barcode with HRM (called Bar-HRM), limitations of DNA barcoding technique could be minimized, (2) HRM analysis method is quite sensitive detecting 0.1%–1% presence of adulterated sample, (3) It is a high-throughput technology that is capable of analyzing multiple samples at the same time, (4) Post-PCR processes are not needed thus cross-contamination could be avoided, (5) The sample genotype can be traced

by evaluating HRM curve analysis. Bar-HRM has thus been proven to be a powerful tool for species identification capable of identifying species and quantitatively detecting adulterants from mixtures of samples of different specimens.

To optimize HRM conditions, care should be taken in terms of primer designing, PCR reagents, and cycling conditions since small differences in melting curves can arise from sources other than the nucleotide sequence. Factors like genomic DNA (gDNA) quality, amplicon length, primer design, dye selection, and PCR conditions are all predicted to affect the melting behavior (Montgomery et al., 2007; van der Stoep et al., 2009).

13.10 High-Throughput Plant DNA Barcoding Using Microfluidic Enrichment Barcoding *(ME Barcoding)*

ME Barcoding is a precious tool for DNA metabarcoding. It is a cost-effective method for high-throughput DNA barcoding that uses microfluidic PCR-based target enrichment (Gostel et al. 2020), for species-level phylogenetic reconstruction. Microfluidic PCR-based barcoding might be preferable to molecular phylogenetics because of its efficiency, with minimal starting template size. There was a very low amount of template and reagents needed for PCR reactions (0.033 mL in the Access Array™ System). Nowadays, Fluidigm Access Array and Illumina MiSeq are used in M.E barcoding. The barcode can be generated from 96 or even more samples for each of the four primary DNA barcode loci in plants: *rbcL*, *matK*, *trnH-psbA*, and ITS. Fluidigm Access Array simultaneously amplifies targeted regions for 48 DNA samples and thus hundreds of PCR primer pairs (producing up to 23,040 PCR products) during a single thermal cycling protocol. This technique is emerging as an alternative to traditional PCR and Sanger sequencing to generate large amounts of plant DNA barcodes and build more comprehensive barcode databases. Microfluidic PCR amplification followed by high-throughput sequencing can produce by locus sequence with minimal resource investment. However, there are two limitations of this approach, viz. (1) A high initial equipment cost, (2) lower sequencing success compared to Sanger methods (Uribe-Convers et al. 2016).

Alternative HTS platforms (e.g., Pacbio SMRT) could be better suited to build plant DNA barcode libraries due to the *matK* region's length. The single molecule, real-time (SMRT) sequencing implemented on the SEQUEL platform to sequence barcode sequence libraries for COI. The instrument had capacity to sequences from more than 5 million DNA extracts a year (Hebert et al. 2018). Combining Pacbio with ME Barcoding could help determine whether the longer sequence read length provided by this single molecule, real-time (SMRT) sequencing approach (up to 60 kb) can improve the recovery success of all four plant DNA barcode loci (Fig. 13.3).

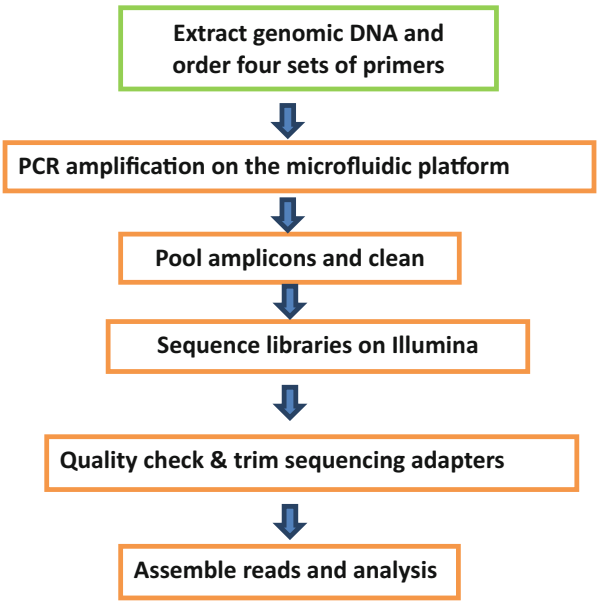

Fig. 13.3 Steps of microfluidic enrichment barcoding

13.11 Full-Length Multi-Barcoding (FLMB)

FLMB is a superior but feasible approach for identifying complex biological mixtures, which shows perfect interpretation for DNA barcoding that could lead to its application in multi-species mixtures. This full-length multi-barcoding (FLMB) via long-read sequencing is employed to identify biological compositions inadequate and well-controlled studies. For instance, in recent years, using various science, engineering, and biotechnology tools, the market foods are modified to improve their taste, color, and flavor, making them commercially more attractive. In this aspect, FLMB can detect most commercially processed foods and herbal mixtures for quantitative analysis of unknown fruit mixtures. It can also determine the composition of mixed spices, flavored teas, vegetable stock cubes, curry, deep-frozen vegetables, food supplements, and health drinks, as well as comprehensive identification of biological origin for herbs (Zhang et al. 2019). Overall, this tool has the potential to provide novel insights into biodiversity analysis in many research areas.

To test the efficacy of FLMB based DNA metabarcoding, DNA is extracted from individual and mixed biological samples and then individually quantified using qPCR, followed by library preparation and SMRT based Sequencing. Bioinformatics analysis is done for proper authentication of individual samples from the contaminated or mixture of samples. The working principle of FLMB is depicted in Fig. 13.4.

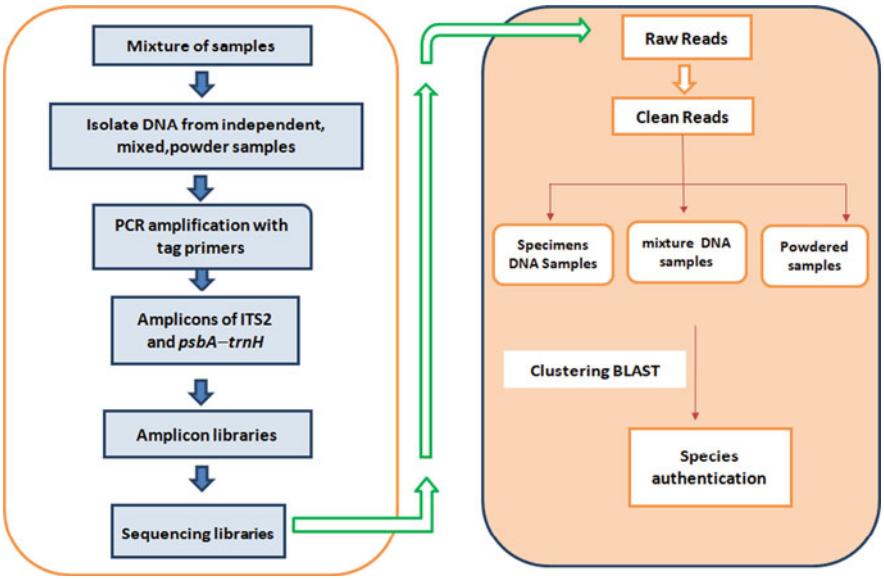

Fig. 13.4 Steps of full-length multi-barcoding

13.12 Genome Skimming Based Barcoding

One approach which offers a relatively straightforward mechanism to improve and extend DNA barcodes is genome skimming (Dodsworth 2015). As a genomic DNA extract typically contains a mix of both nuclear and organellar DNA (plastid and mitochondria), NGS generates data across the three genomes. Therefore, genome skimming deals with the ultimate goal of assembling organellar reference genomes. Through genome skimming, there is also potential to make a highly fragmented nuclear genome assembly. Overall, genome skimming is scalable, cost-effective, and can be used effectively with degraded DNAs from herbarium specimens or highly fragmented nuclear genome assembly (Nevill et al. 2020). This approach recovers simultaneously all of the different "standard" plant barcoding regions and provides a direct link with all other phylogenetically informative genomic regions. The second benefit of genome skimming is that it is compatible with the standard plant barcodes and genome sequencing. Genome skimming can recover plastid barcode loci and ITS, thus adding to the standard barcoding loci's growing reference database. Many genome skimming studies only assemble the organellar and ribosomal DNA, excluding the nuclear reads.

The key challenges to widespread adoption of genome skimming as an extended barcode will be dependent on the efficacy of its implementation at a vast scale, cost implications for library preparation (which is also time-consuming), consumables, computational power, and data storage. Another major challenge of genome

skimming DNA barcoding is how to use nuclear data effectively (Coissac et al. 2016) because 99% genome sequence data are discarded. Coissac et al. (2016) proposed DNA mark pipeline which will enable future DNA-based identification.

13.13 Restriction Site-Associated DNA Sequencing (RAD)

Restriction site-**A**ssociated **D**NA sequencing (RADseq) and its derivative methods have been applied mostly for assessing population structure, hybridization, demographic history, phylogeography of organisms (Baird et al. 2008). The reduced representation of genome scale has the potential to be implemented as an alternative species identification tool. This method accesses large numbers of sequence variations adjacent to restriction-enzyme cut sites and sequence the homologous regions across hundreds of individuals, without genome sequence information. RAD sequencing is one promising approach which has already been used to authenticate complex species. RAD showed huge phylogenetic resolution among temperate bamboo species which has less molecular variation due to their recent origin (Wang et al. 2017).

13.14 Conclusion

The primary aim of DNA barcoding is to identify known specimens and to help flag possible new species, thereby making taxonomy more useful for science and society. Thus, it is based on conventional and inexpensive protocols for DNA extraction, amplification, and sequencing. DNA barcoding is an approach to developing a global, open-access library of standardized DNA barcode sequences, which would help non-expert identify specimens up to species level. Certain limitations (low PCR efficiency, inadequate variation in single-locus barcode) restrict achievement of a universal DNA barcode system for land plants. However, multi-locus DNA barcoding approach is still one of the most effective strategies for barcoding some complex plants groups. With the advancement of next-generation sequencing technologies, genome skimming RAD seq, etc. were evolved to sampling variation throughout the genome and help identify the complex plant species with better species resolution. Integrations of genome skimming RAD seq, HRM, ME Barcoding, FLMB approaches have further paved the way in overcoming the present limitations of plant DNA barcoding which would play a vital role towards the development of Digital Plant Identification System.

References

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3(10):e3376

Bell KL, De Vere N, Keller A, Richardson RT, Gous A, Burgess KS, Brosi BJ (2016) Pollen DNA barcoding: current applications and future prospects. Genome 59(9):629–640
Blaxter ML (2004) The promise of a DNA taxonomy. Philos Trans R Soc Lond B Biol Sci 359(1444):669–679
CBOL, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL (2009) A DNA barcode for land plants. Proc Natl Acad Sci 106(31):12,794–12,797
Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS One 5(1): e8613
Coissac E, Hollingsworth PM, Lavergne S, Taberlet P (2016) From barcodes to genomes: extending the concept of DNA barcoding. Wiley Online Library
Dodsworth S (2015) Genome skimming for next-generation biodiversity analysis. Trends Plant Sci 20(9):525–527
Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S (2015) ycf1, the most promising plastid DNA barcode of land plants. Sci Rep 5(1):1–5
Erickson DL, Spouge J, Resch A, Weigt LA, Kress JW (2008) DNA barcoding in land plants: developing standards to quantify and maximize success. Taxon 57(4):1304–1316
Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SC (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS One 3(7):e2802
Fernandes TJ, Amaral JS, Mafra I (2020) DNA barcode markers applied to seafood authentication: an updated review. Crit Rev Food Sci Nutr:1–32
Gostel MR, Zúñiga JD, Kress WJ, Funk VA, Puente-Lelievre C (2020) Microfluidic Enrichment Barcoding (MEBarcoding): a new method for high throughput plant DNA barcoding. Sci Rep 10(1):1–13
Gross M (2011) Herbaria source of new plant species. Curr Biol 21:R6–R7
Group C P B, Li D-Z, Gao L-M, Li H-T, Wang H, Ge X-J, Liu J-Q, Chen Z-D, Zhou S-L, Chen S-L (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proc Natl Acad Sci 108(49):19,641–19,646
Hajibabaei M, Singer GA, Hebert PD, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet 23(4):167–172
Hebert PD, Cywinska A, Ball SL, Dewaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc Lond Ser B Biol Sci 270(1512):313–321
Hebert PD, Braukmann TW, Prosser SW, Ratnasingham S, DeWaard JR, Ivanova NV, Janzen DH, Hallwachs W, Naik S, Sones JE (2018) A Sequel to Sanger: amplicon sequencing that scales. BMC Genomics 19(1):1–14
Hilu KW, Liang G (1997) The matK gene: sequence variation and application in plant systematics. Am J Bot 84(6):830–839
Hollingsworth P (2008) DNA barcoding plants in biodiversity hot spots: progress and outstanding questions. Nature Publishing Group
Hollingsworth ML, Andra Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Mol Ecol Resour 9(2):439–457
Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PLoS One 6(5):e19254
Kim K-J, Jansen RK (1995) ndhF sequence evolution and the major clades in the sunflower family. Proc Natl Acad Sci 92(22):10,379–10,383
Kool A, de Boer HJ, Krüger Å, Rydberg A, Abbad A, Björk L, Martin G (2012) Molecular identification of commercialized medicinal plants in southern Morocco. PLoS One 7(6):e39459

Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS One 2(6):e508
Kress WJ, Erickson DL (2008) DNA barcodes: genes, genomics, and bioinformatics. Proc Natl Acad Sci 105(8):2761–2762
Lahaye R, Van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V (2008) DNA barcoding the floras of biodiversity hotspots. Proc Natl Acad Sci 105(8):2923–2928
Lee SY, Lamasudin DU, Mohamed R (2019) Rapid detection of several endangered agarwood-producing Aquilaria species and their potential adulterants using plant DNA barcodes coupled with high-resolution melting (Bar-HRM) analysis. Holzforschung 73(5):435–444
Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S (2015) Plant DNA barcoding: from gene to genome. Biol Rev 90(1):157–166
Madesis P, Ganopoulos I, Anagnostis A, Tsaftaris A (2012) The application of Bar-HRM (Barcode DNA-High Resolution Melting) analysis for authenticity testing and quantitative detection of bean crops (Leguminosae) without prior DNA purification. Food Control 25(2):576–582
Mahadani P, Ghosh SK (2013) DNA barcoding: a tool for species identification from herbal juices. DNA Barcodes 1:35–38
Mahadani P, Sharma GD, Ghosh SK (2013) Identification of ethnomedicinal plants (Rauvolfioideae: Apocynaceae) through DNA barcoding from Northeast India. Pharmacogn Mag 9(35):255
Mishra P, Kumar A, Nagireddy A, Mani DN, Shukla AK, Tiwari R, Sundaresan V (2016) DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. Plant Biotechnol J 14(1):8–21
Muellner A, Schaefer H, Lahaye R (2011) Evaluation of candidate DNA barcoding loci for economically important timber species of the mahogany family (Meliaceae). Mol Ecol Resour 11(3):450–460
Nevill PG, Zhong X, Tonti-Filippini J, Byrne M, Hislop M, Thiele K, Van Leeuwen S, Boykin LM, Small I (2020) Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. Plant Methods 16(1):1–8
Ratnasingham S, Hebert PD (2007) BOLD: the barcode of life data system. Mol Ecol Notes 7(3):355–364. (http://www.barcodinglife.org)
Roy S, Tyagi A, Shukla V, Kumar A, Singh UM, Chaudhary LB, Datt B, Bag SK, Singh PK, Nair NK (2010) Universal plant DNA barcode loci may not work in complex groups: a case study with Indian Berberis species. PLoS One 5(10):e13674
Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Philos Trans R Soc B Biol Sci 360(1462):1805–1811
Serino G, Maliga P (1998) RNA polymerase subunits encoded by the plastid rpoGenes are not shared with the nucleus-encoded plastid enzyme. Plant Physiol 117(4):1165–1170
Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. Am J Bot 92(1):142–166
Spooner DM (2009) DNA barcoding will frequently fail in complicated groups: an example in wild potatoes. Am J Bot 96(6):1177–1189
Stoeckle MY, Gamble CC, Kirpekar R, Young G, Ahmed S, Little DP (2011) Commercial teas highlight plant DNA barcode identification successes and obstacles. Sci Rep 1(1):1–7
Sui X-Y, Huang Y, Tan Y, Guo Y, Long C-l (2011) Molecular authentication of the ethnomedicinal plant Sabia parviflora and its adulterants by DNA barcoding technique. Planta Med 77(05):492–496
Sun X-Q, Zhu Y-J, Guo J-L, Peng B, Bai M-M, Hang Y-Y (2012) DNA barcoding the Dioscorea in China, a vital group in the evolution of monocotyledon: use of mat K gene for species discrimination. PLoS One 7(2):e32057

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy. Trends Ecol Evol 18(2):70–74

Uribe-Convers S, Settles ML, Tank DC (2016) A phylogenomic approach based on PCR target enrichment and high throughput sequencing: resolving the diversity within the south American species of Bartsia L. (Orobanchaceae). PLoS One 11(2):e0148203

Van K, Rastogi K, Kim K, Lee S (2013) Next-generation sequencing technology for crop improvement. SABRAO J Breeding Genet 45(1):84–99

Wang X, Ye X, Zhao L, Li D, Guo Z, Zhuang H (2017) Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (Poaceae: Bambusoideae). Sci Rep 7(1):1–11

Wang X, Gussarova G, Ruhsam M, de Vere N, Metherell C, Hollingsworth PM, Twyford AD (2018) DNA barcoding a taxonomically complex hemiparasitic genus reveals deep divergence between ploidy levels but lack of species-level resolution. AoB Plants 10(3). https://doi.org/10.1093/aobpla/ply026

Zhang P, Liu C, Zheng X, Wu L, Liu Z, Liao B, Shi Y, Li X, Xu J, Chen S (2019) Full-length multi-barcoding: DNA barcoding from single ingredient to complex mixtures. Genes (Basel) 10(5):343. https://doi.org/10.3390/genes10050343