BIOLOGICAL REVIEWS
Cambridge Philosophical Society
Biol. Rev. (2015), **90**, pp. 157–166.
doi: 10.1111/brv.12104

Plant DNA barcoding: from gene to genome

Xiwen Li[1], Yang Yang[1], Robert J. Henry[2], Maurizio Rossetto[3], Yitao Wang[1,*] and Shilin Chen[4,*]

[1] *State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences, University of Macau, Macau, 999078, China*
[2] *Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane, Queensland 4072, Australia*
[3] *National Herbarium of NSW, The Royal Botanic Gardens and Domain Trust, Mrs Macquaries Road, Sydney, New South Wales 2000, Australia*
[4] *Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China*

ABSTRACT

DNA barcoding is currently a widely used and effective tool that enables rapid and accurate identification of plant species; however, none of the available loci work across all species. Because single-locus DNA barcodes lack adequate variations in closely related taxa, recent barcoding studies have placed high emphasis on the use of whole-chloroplast genome sequences which are now more readily available as a consequence of improving sequencing technologies. While chloroplast genome sequencing can already deliver a reliable barcode for accurate plant identification it is not yet resource-effective and does not yet offer the speed of analysis provided by single-locus barcodes to unspecialized laboratory facilities. Here, we review the development of candidate barcodes and discuss the feasibility of using the chloroplast genome as a super-barcode. We advocate a new approach for DNA barcoding that, for selected groups of taxa, combines the best use of single-locus barcodes and super-barcodes for efficient plant identification. Specific barcodes might enhance our ability to distinguish closely related plants at the species and population levels.

Key words: single-locus barcode, universal, plastid-sequencing, super-barcode, specific barcode.

CONTENTS

I. INTRODUCTION

There are an estimated 300000 plant species in the world (IUCN, 2012) but relatively few of these can be identified based on traditional plant identification methods (Hebert *et al.*, 2003; Bickford *et al.*, 2007; Chase & Fay, 2009). Accurate classification and identification of this large number of species remains a significant challenge even for specialist taxonomists. The emergence of DNA barcoding has had a positive impact on biodiversity classification and identification (Gregory, 2005). DNA barcoding is a technique for characterizing species of organisms using a short DNA sequence from a standard and agreed-upon position in the genome (http://barcoding.si.edu/DNABarCoding.htm).

* Address for correspondence (Tel: +853-8397-4691, +86-010-62811448; E-mail: ytwang@umac.mo, slchen@implad.ac.cn).

Since it was first put forward and widely applied in animals (Hebert *et al.*, 2003), DNA barcoding has attracted much attention from taxonomists. DNA barcoding can also be used for a wide range of purposes: to support ownership or intellectual property rights (Stewart, 2005); to reveal cryptic species (Hebert *et al.*, 2004); in forensics to link biological samples to crime scenes (Yoon, 1993; Coyle *et al.*, 2005; Mildenhall, 2006); to support food safety and authenticity of labelling by confirming identity or purity (Galimberti *et al.*, 2012; Huxley-Jones *et al.*, 2012); and in ecological and environmental genomic studies (Valentini *et al.*, 2009).

Global DNA barcoding was initially regarded as a 'big science' programme (Gregory, 2005) and even as the renaissance of taxonomy (Miller, 2007). However, the cytochrome c oxidase 1 (CO1) sequence, which has been developed as a universal barcode in animals, does not discriminate most plants because of a much slower mutation rate (Kress *et al.*, 2005). Although many studies have searched for a universal plant barcode, none of the available loci work across all species (Chase & Fay, 2009; Chen *et al.*, 2010). The Consortium for the Barcode of Life-Plant Working Group (CBOL) recently recommended the two-locus combination of *matK*+ *rbcL* as the best plant barcode with a discriminatory efficiency of only 72% (CBOL Plant Working Group, 2009). Taxonomists have suggested that a multi-locus method may be necessary to discriminate plant species (Hebert *et al.*, 2004; Chase *et al.*, 2007; Kress & Erickson, 2007; Erickson *et al.*, 2008; Kane & Cronk, 2008; Lahaye *et al.*, 2008; Kane *et al.*, 2012). However, CBOL demonstrated that the use of multiple loci did not clearly improve the species-level discriminatory ability of these techniques (CBOL Plant Working Group, 2009).

Researchers have recently proposed the use of the whole-plastid genome sequence in plant identification (Erickson *et al.*, 2008; Sucher & Carles, 2008; Parks, Cronn & Liston, 2009; Nock *et al.*, 2011; Yang *et al.*, 2013). However this concept has not yet been universally accepted. One of the main concerns is the high sequencing cost and difficulties involved in obtaining complete plastid genome sequences in comparison to the use of single-locus barcodes. Hollingsworth, Graham & Little (2011) argued that the full plastid haplotype is not a good marker because it does not always track species boundaries. To date, it is still unclear whether plastid genomes can be regarded as a suitable barcode.

Here we review the history of plant barcode selection and look at future prospects for DNA barcoding in plants (Fig. 1). The feasibility of using the chloroplast genome (cp-genome) as a 'super-barcode' is evaluated, and the concept of a 'specific barcode' derived from the comparison between plastid genome sequences from a target group of taxa is presented as an effective option that might be widely applicable to plant identification studies. Specific barcodes may provide new perspectives in the search for rapid and accurate methods for species discrimination, especially for closely related plants.

II. SINGLE-LOCUS DNA BARCODES

Traditional barcodes have been widely studied but still have significant limitations. Some of these widely used single-locus barcodes are described below.

(1) *MatK*

MatK has a high evolutionary rate, suitable length and obvious interspecific divergence as well as a low transition/transversion rate (Min & Hickey, 2007; Selvaraj, Sarma & Sathishkumar, 2008). Unfortunately, *matK* is difficult to amplify universally using currently available primer sets. The CBOL Plant Working Group (2009) revealed nearly 90% success rate in amplifying angiosperm DNA using a single primer pair. However, the success was limited in gymnosperms (83%) and much worse in cryptogams (10%) even with multiple primer sets. Different primer pairs were required in different taxonomic groups (Chase *et al.*, 2007; Hollingsworth, 2008). Lahaye *et al.* (2008) used specific primers (Cuénoud *et al.*, 2002) to amplify the *matK* gene of 1667 angiosperm plant samples and achieved a success rate of 100%. A further challenge is the different discrimination rates in different taxonomic groups. *Matk* can discriminate more than 90% of species in the Orchidaceae (Kress & Erickson, 2007) but less than 49% in the nutmeg family (Newmaster *et al.*, 2008). Fazekas *et al.* (2008) attempted the identification of 92 species from 32 genera using the *matK* barcode but only achieved a success rate of 56%. These findings demonstrate that the *matK* barcode alone is not a suitable universal barcode.

(2) *RbcL*

RbcL is widely used in phylogenetic investigations with over 50000 sequences available in Genbank. The advantages of this gene are that it is easy to amplify, sequence and align in most land plants and is a good DNA barcoding region for plants at the family and genus levels. However, *rbcL* sequences evolve slowly and this locus has by far the lowest divergence of plastid genes in flowering plants (Kress *et al.*, 2005). Consequently, it is not suitable at the species level due to its modest discriminatory power (Fazekas *et al.*, 2008; Lahaye *et al.*, 2008; CBOL Plant Working Group, 2009; Chen *et al.*, 2010). The length of the gene can also be problematic as double-stranded sequencing of the entire gene sequence may require four primers. Despite these limitations, *rbcL* was still suggested as one of the best potential candidate plant barcodes based on the straightforward recovery of the gene sequence, the large amount of easily accessible data and good, but not outstanding, discriminatory power (Blaxter, 2004; CBOL Plant Working Group, 2009; Hollingsworth *et al.*, 2011) even though it was previously rejected as a target for species identification (Gielly & Taberlet, 1994; Renner, 1999; Salazar *et al.*, 2003). Although *rbcL* by itself does not meet the desired attributes of a barcoding locus, it is accepted that *rbcL* in combination with various plastid or nuclear loci can make

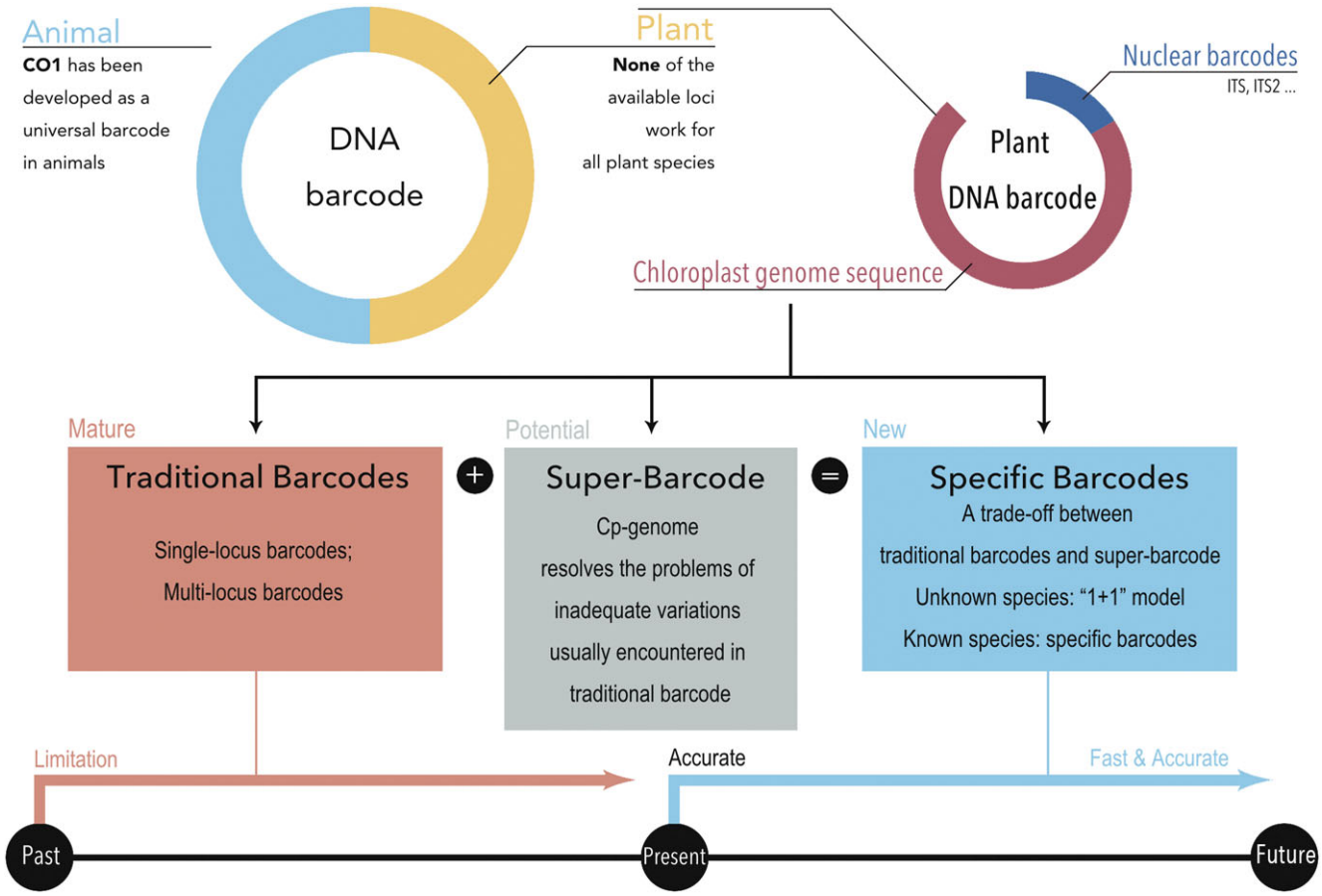

Fig. 1. Schematic timeline of plant barcoding history and possible developments. CO1, cytochrome c oxidase 1; cp, chloroplast; ITS, internal transcribed spacer.

accurate identifications (Newmaster, Fazekas & Ragupathy, 2006; Chase *et al.*, 2007; Kress & Erickson, 2007; CBOL Plant Working Group, 2009; Hollingsworth *et al.*, 2009).

(3) *TrnH-psbA*

TrnH-psbA is currently the most widely used plastid barcode. The presence of highly conserved coding sequences on both sides make the design of universal primers feasible (Shaw *et al.*, 2005), with a single primer pair likely to amplify nearly all angiosperms (Shaw *et al.*, 2007). The non-coding intergenic region exhibits most sequence divergence and has high rates of insertion/deletion (Kress & Erickson, 2007). These attributes make *trnH-psbA* highly suitable as a plant barcode for species discrimination (Kress & Erickson, 2007; Shaw *et al.*, 2007), and extensive barcoding studies demonstrated that in some land plant groups such as *Hydrocotyle*, *Dendrobium* and Pteridophytes (van de Wiel *et al.*, 2009; Yao *et al.*, 2009; Ma *et al.*, 2010) the *trnH-psbA* region could identify nearly all species.

Alignment of the *trnH-psbA* spacer can be highly ambiguous because of its complicated molecular evolution, considerable length variation (Chang *et al.*, 2006), and high rates of insertion/deletion in larger families of angiosperms (Chase *et al.*, 2007). Furthermore, due to the presence of duplicated loci and a pseudogene, the *trnH-psbA* sequence is much longer [>1000 base pairs (bp)] in some conifers and monocots (Chase *et al.*, 2007; Hollingsworth *et al.*, 2009) while it is exceedingly short, less than 300 bp, in other groups (Kress *et al.*, 2005) and shorter than 100 bp in bryophytes (Stech & Quandt, 2010). One of the key problems associated with the use of *trnH-psbA* as a standard barcode is the frequent inversions in some plant lineages, which may lead to large overestimates of genetic divergence and to incorrect phylogenetic assignment (Whitlock, Hale & Groff, 2010). Additionally, because of the premature termination of sequencing reads caused by mononucleotide repeats, longer *trnH-psbA* regions can be difficult to retrieve without taxon-specific internal sequencing primers designed to obtain high-quality bi-directional sequences (Devey, Chase & Clarkson, 2009; Ebihara, Nitta & Ito, 2010). Shorter *trnH-psbA* spacers may not have adequate sequence variation for species discrimination, such as in the genera *Solidago* (Kress *et al.*, 2005). As a consequence, Kress *et al.* (2005) and Chase *et al.* (2007), respectively, proposed that *trnH-psbA* can be used in two-locus or three-locus barcode systems to provide adequate resolution.

(4) *ITS*

The *ITS* spacer is a powerful phylogenetic marker at the species level showing high levels of interspecific divergence (Álvarez & Wendel, 2003). The greater discriminatory power of *ITS* over plastid regions at low taxonomic levels has been widely studied leading to it also being suggested as a plant barcode (Stoeckle, 2003; Kress *et al.*, 2005; Sass *et al.*, 2007), especially in parasitic plants which offer less resolution from plastid barcodes (Hollingsworth *et al.*, 2011). However, CBOL has only regarded *ITS* as a supplementary locus (CBOL Plant Working Group, 2009). Some limitations prevent it from being a core barcode: incomplete concerted evolution, fungal contamination and difficulties of amplification and sequencing (Hollingsworth *et al.*, 2011). Fungi contain *ITS* sequences that can be amplified (sometimes preferentially) and confused with plant sequences.

Presenting a different view, the China Plant BOL Group recently argued that when direct sequencing was possible, the *ITS* region should be incorporated into the core barcodes because of higher discriminatory power than plastid barcodes (China Plant BOL Group, 2011). To resolve the difficulties involved in sequencing the entire *ITS*, they suggested *ITS2* as a backup because of its conserved sequence characters which reduce amplification and sequencing problems. It was accepted that *ITS2* could be used as a novel universal barcode for the identification of a broader range of plant taxa (Chen *et al.*, 2010; Gao *et al.*, 2010*a*,*b*; Luo *et al.*, 2010; Pang *et al.*, 2010, 2011) even from herbarium specimens with degraded DNA (Chiou *et al.*, 2007). Although the *ITS2* barcode displays some advantages compared to other candidate loci, including *ITS*, researchers have not given much attention to this region. A major concern is the existence of multiple copies in the genome with high levels of within-species and even within-individual sequence differentiation (Yamaguchi, Kawamura & Horiguchi, 2006), which may lead to inaccurate or misleading results (Álvarez & Wendel, 2003). Song *et al.* (2012) recently showed that the *ITS2* intra-genomic distances were markedly smaller than those of the intra-specific or inter-specific variants in a wide range of plant families. Although the use of *ITS2* circumvents low polymerase chain reaction (PCR) efficiency, more investigations are needed to assess the extent to which the access to fewer characters reduces discrimination power in comparison to the entire *ITS* region (Hollingsworth *et al.*, 2011). For example, the *ITS2* sequences are generally less than 300 bp in *Fritillaria* and do not have adequate interspecific divergence for species resolution.

(5) Other widely used plastid barcodes

At present, DNA barcoding technology relies heavily on chloroplast loci because of their relatively low evolutionary rates compared with nuclear loci (Dong *et al.*, 2012). Beyond the candidate barcodes described above, there are many other widely used plastid barcoding markers, such as *rpoB*, *rpoC*1, *atpF-atpH*, *psbK-psbI*, *ycf5* and *trnL* (P6). Their properties have been discussed in detail by Hollingsworth *et al.* (2011) and Vijayan & Tsou (2010). These chloroplast regions are valuable for phylogenetic analyses and barcoding studies at higher taxonomic levels but are not suitable for plant DNA barcoding at lower taxonomic levels because of insufficient variation.

Molecular evolution of cp-genome sequences also shows both lineage-specific and nonrandom spatial patterns of substitution (Gruenheit *et al.*, 2008; Zhong *et al.*, 2011; Dong *et al.*, 2012; Ahmed *et al.*, 2013). For example, Dong *et al.* (2012) demonstrated that the region of *ycf1* located in the IRb region is conservative while the two regions located in the SSC region are extremely variable. Such substitution patterns in chloroplast genomes indicate complex processes of mutation that are asymmetric, and lack independence between sequence positions. Thus, the patterns of substitution are not well described by currently used substitution models, particularly with respect to deeper phylogenetic divergences (Lockhart & Steel, 2005). Chloroplast sequence evolution can be inconsistent across lineages, and phylogenetic incongruence between different chloroplast gene loci is possible (Lockhart & Steel, 2005; Magee *et al.*, 2010; Wu *et al.*, 2011; Dong *et al.*, 2012). Therefore it can be problematic to find an ideal universal barcode applicable at various taxonomic levels.

III. CANDIDATE MULTI-LOCUS DNA BARCODES

Despite extensive efforts to identify a universal plant barcode comparable to CO1 in animals, the task has proved difficult due to the lack of adequate variation within single loci (Kress *et al.*, 2005; Newmaster *et al.*, 2006; Chase *et al.*, 2007; Kress & Erickson, 2007; Sass *et al.*, 2007; Fazekas *et al.*, 2008; Lahaye *et al.*, 2008). Many researchers have suggested that a multi-locus method will be required to obtain adequate species discrimination (Hebert *et al.*, 2004; Kress & Erickson, 2007; Erickson *et al.*, 2008; Kane & Cronk, 2008; Lahaye *et al.*, 2008; CBOL Plant Working Group, 2009; Chase & Fay, 2009). Various combinations of plastid loci have been proposed including *rbcL* + *trnH-psbA* (Kress & Erickson, 2007), *rpoC1* + *rpoB* + *matK* or *rpoC1* + *matK* + *trnH-psbA* (Chase *et al.*, 2007) and *matK* + *atpF-atpH* + *psbK-psbI* or *matK* + *atpF-atpH* + *trnH-psbA* (Pennisi, 2007). These combined barcodes exhibit higher species discrimination than single-locus approaches. Different research groups have tested different combinations using different taxa while attempting to select a universal barcode, however universal agreement is yet to be reached. Fazekas *et al.* (2008) compared these barcode combinations using the same large-scale taxonomic samples, but none could identify more than 70% of tested species.

The CBOL Plant Working Group recently recommended *matK* + *rbcL* as the universal barcode combination due to the straightforward recovery of the *rbcL* region and the discriminatory power of the *matK* sequence (CBOL Plant Working Group, 2009). Although the choice of *rbcL* + *matK*

offered slightly higher identification efficiency than other combinations, the *rbcL* + *matK* barcode still failed to meet the original goal of a universal DNA barcode. Firstly, the combination of *rbcL* + *matK* cannot avoid the low PCR efficiency of *matK* and secondly, the success of *rbcL* + *matK* in discriminating plants is typically lower than that of CO1 in animals. Combined barcodes increase analytical difficulties compared to single-locus markers, especially when one of the target loci does not amplify. What's more, CBOL demonstrated that the use of seven candidate loci did not significantly improve species-level discriminatory ability compared to *rbcL* + *matK*. Some authors considered that the failure of multiple-locus barcodes to increase species discrimination was not simply due to the lack of variation; rather it reflected the discrepancies between the plastid gene tree and species boundaries (Fazekas *et al.*, 2009; Hollingsworth *et al.*, 2011). Thus, the combinations of candidate loci cannot eliminate the inherent deficiencies of current DNA barcoding of plants.

IV. SUPER-BARCODING: A NEW WAY FOR PLANT DISCRIMINATION

Because of the inherent limitations of single-locus DNA barcodes, a new method is needed to identify closely related plant species (Heinze, 2007). It has recently been pointed out that the complete cp-genome contained as much variation as the CO1 locus in animals and may be used as a plant barcode (Kane & Cronk, 2008). The complete cp-genome has a conserved sequence ranging from 110 to 160 kbp, greatly exceeding the length of commonly used DNA barcodes and providing more variation to discriminate closely related plants. The cp-genome has been used as a versatile tool for phylogenetics. It can greatly increase resolution at lower taxonomic levels in plant phylogenetic, phylogeographic and population genetic analyses, facilitating the recovery of lineages as monophyletic, and was therefore proposed as a species-level DNA barcode (Parks *et al.*, 2009). Using the cp-genome as a marker circumvents possible issues with gene deletion and low PCR efficiency (Huang *et al.*, 2005). The analysis of this super-barcode also resolves the problems of sequence retrieval usually encountered in traditional barcoding studies. Compared with the nuclear genome, the cp-genome is small in size and has a higher interspecific and lower intraspecific divergence, which makes it more suitable as a genome-based barcode. Species identification can also be performed according to whether a gene exists in either of two species, which is regarded as the simplest test of species identification based on barcoding approaches (Hebert *et al.*, 2004). This is because super-barcoding is more efficient in detecting gene loss and defining gene order than traditional barcoding (Luo *et al.*, 2008, 2009).

Although sequences from single or multiple chloroplast and nuclear genes have been useful for differentiating species, the cp-genome has been used efficiently to distinguish between closely related species (Parks *et al.*, 2009; Nock *et al.*, 2011), populations (Doorduin *et al.*, 2011) and individuals (Kane *et al.*, 2012; McPherson *et al.*, 2013). This approach is still relatively controversial, Hollingsworth *et al.* (2011) suggested that often the plastid genome could not completely track species boundaries. However their conclusion was largely based on an individual case study (discussed by Fazekas *et al.*, 2009) rather than on large-scale comparative analyses. In comparison, Joly, McLenachan & Lockhart (2009) have provided a promising method based on the use of minimum genetic distances to distinguish between hybridization and incomplete lineage sorting. Software implementing this method (Joly, 2012), termed 'JML' was recently used to analyse chloroplast gene sequences and identify a hybrid and geographically isolated lineage of *Pachycladon* persisting in the Southern alps of New Zealand (Becker *et al.*, 2013). In this example, the power and resolution of JML was greatly improved by analysing concatenated chloroplast loci. JML seems particularly appropriate for evaluating the issue of species boundaries using part (e.g. 5–10 k) or entire cp-genomes as a plant super-barcode. Nevertheless plastid-genome-based species classification and identification have been progressively more accepted by taxonomists (Shendure & Ji, 2008; Kumar *et al.*, 2009; Wu *et al.*, 2010; Bayly *et al.*, 2013; Yang *et al.*, 2013). The main challenges of super-barcoding are the establishment of a rich cp-genome database and the reduction of sequencing cost, as well as obtaining a higher quality and quantity of DNA (Kane *et al.*, 2012). The first cp-genome was sequenced in 1986 (Shinozaki *et al.*, 1986); by 2012 there were 254 complete plant cp-genomes within public databases, which only accounts for less than 0.01% of total plant species and is still a small number for widespread species identification. With the development of next-generation sequencing (NGS), the number of cp-genomes sequenced has increased rapidly (Fig. 2). The number of new cp-genomes published in 2012 greatly exceeded the total number sequenced in each of the previous 20 years.

Sample preparation has been regarded as the key factor in multiplex sequencing (sequencing of multiple tagged samples together in one lane) of plastid genomes (Parks *et al.*, 2009). Low-quality DNA templates such as contaminated DNA samples generate noise which require labour-intensive evaluations during sequence assembly. NGS requires a much larger amount of more-purified DNA than PCR-based sequencing techniques but standard methods of cp-genome extraction have been established (Diekmann *et al.*, 2008; Shi *et al.*, 2012). Although it was not initially straightforward (Hollingsworth *et al.*, 2011), researchers recently provided standardized protocols for extracting pure chloroplast DNA using fresh leaves, assisting plastid sequencing and sequence assembly. Targeted enrichment protocols are being trialed (Stull *et al.*, 2013), but recent procedures can use total DNA as a template for cp-genome sequencing not only solving the problem of extracting chloroplast DNA from dried and even degraded materials but also simplifying the whole process (Nock *et al.*, 2011). A recent comparative study demonstrated that deriving bio-informatically the entire cp-genome from

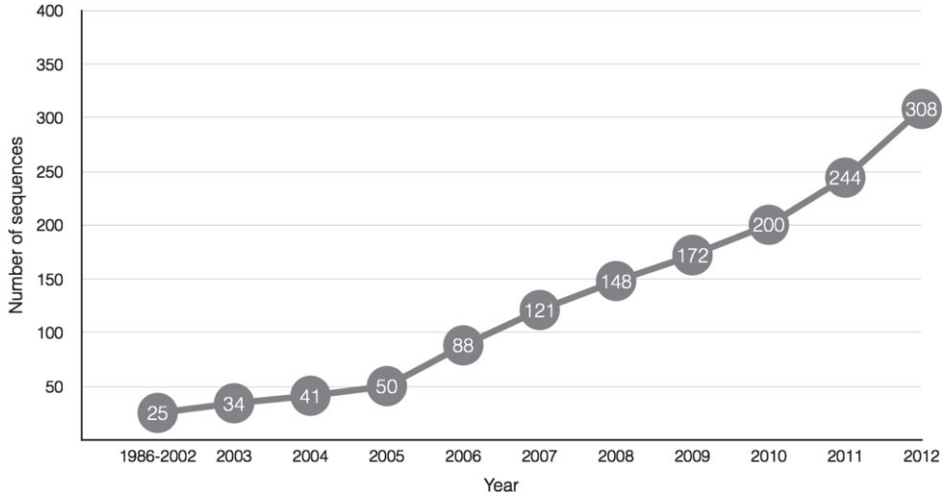

Fig. 2. The total number of complete chloroplast genome sequences submitted to Genbank from 1986 to 2012.

whole-DNA shotgun sequence data without the need for a reference genome, is as accurate but considerably less resource intensive that obtaining it from purified chloroplast DNA (McPherson *et al.*, 2013). Thus neither extraction methods nor sequencing capacity can any longer be considered as limiting factors for obtaining cp-genome data, as NGS can generate many individual super-barcodes (Doorduin *et al.*, 2011). NGS along with multiplex identifiers (MID) technology and other multiplexing tools can allow for the sequencing of 100 or more complete cp-genomes in a single run. McPherson *et al.* (2013) showed that it is possible to obtain the full cp-genome from less than 1 GB of whole-DNA shotgun data. Although assembling short sequence reads into cp-genomes in the absence of a reference genome may require some data inspection and interpretation, a closely related reference is not absolutely needed for sequence assembly (Straub *et al.*, 2011), and dedicated pipelines are being developed (McPherson *et al.*, 2013). As sequencing read length continues to increase, assembling plastid genomes without a reference genome will become increasingly popular for a broad range of applications, particularly as in-house library-making (Rohland & Reich, 2012) and multiplexing will reduce costs to well below $100 per cp-genome, especially with the potential to sequence 100 or more samples in a single lane.

Although sequencing cost has substantially decreased (Kane *et al.*, 2012), current costs for whole cp-genome sequencing still exceed that of obtaining single-locus barcodes by Sanger sequencing, particularly when primer and PCR optimization are not required for the latter approach. Even excluding these factors if plastid-based identifications are reliant on a fully annotated cp-sequence, the necessary analyses can be complex and difficult to standardize.

Continuing advances in NGS technologies have provided new options for obtaining chloroplast sequences. The Roche/454 sequencing platform currently provides the longest sequence reads and is a good but relatively expensive choice for *de novo* sequencing if there are no closely related plastid sequences in public databases. The Illumina platform has provided a cheaper alternative. Further advances in these platforms are likely in the near future reducing the costs of chloroplast sequencing by increasing sequencing data volumes and providing increased opportunities for combining samples for sequencing in the same run. For example, it is expected that samples of total rice DNA might be multiplexed (e.g. 96-fold combining samples from 96 samples) and sequenced in a single run to obtain enough coverage (of the order of 1000-fold for the chloroplast of each genotype) to allow *de novo* cp-genome assembly and analysis. Preliminary studies across multiple species are showing that *de novo* cp-genome assembly from shotgun data is efficient and informative even without a reference genome or any knowledge of genome size (van der Merwe *et al.*, 2013). These advances will reduce the cost to be almost equal to that of a single-locus barcode per cp-genome. As sequencing technology and bioinformatics continue to improve rapidly, complete plastome sequencing will become more popular and may eventually replace Sanger-based DNA barcoding. The chloroplast provides a barcode that can also be successfully tailored to the study of relationships in specific plant groups (Bayly *et al.*, 2013; Yang *et al.*, 2013).

V. SPECIFIC BARCODE: A TRADE-OFF BETWEEN SINGLE-LOCUS BARCODES AND SUPER-BARCODES

Single-locus barcodes lack adequate variations while fully annotated super-barcodes currently can be costly and may be overly complicated for laboratories that lack the necessary experience. To resolve this current challenge, we put forward the concept of using 'specific barcodes' which involve a trade-off between single-locus barcodes and super-barcodes (Fig. 1). A specific barcode is a fragment of DNA sequence that has a sufficiently high mutation rate to enable species identification within a given taxonomic group. Because

Table 1. DNA markers tested for their suitability for barcoding in given plant groups

Genera	Barcode markers used	Success rate of unique identification (%)	References
Lemnaceae	*atpF-atpH*	92.85	Wang *et al.* (2010)
Asteraceae	*ITS2*	97.4	Gao *et al.* (2010*b*)
Fabaceae	*matK*	96	Gao *et al.* (2011)
Rutaceae	*ITS2*	100	Luo *et al.* (2010)
Orchid	*matK*	90	Lahaye *et al.* (2008)
Hydrocotyle	*trnH-psbA*	100	van de Wiel *et al.* (2009)
Dendrobium	*psbA-trnH*	100	Yao *et al.* (2009)
Medicinal plants	*ITS2*	99.8	Chen *et al.* (2010)
Cycas	*ITS*	91.7	Sass *et al.* (2007)
Macrozamia	*ITS*	100	Sass *et al.* (2007)
Aspalathus	*trnT-trnL*	100	Edwards *et al.* (2008)
Swartzia	*ITS2*	97.4	Gao *et al.* (2010*a*)
Taxus	*trnL-F/ITS*	100	Liu *et al.* (2011)
Pteridophytes	*psbA-trnH*	90.2	Ma *et al.* (2010)
Solanum	*trnS-trnG/ndhF*	100	Zhang *et al.* (2013)

specific barcodes are chosen directly from the plastid genome sequences of target families or genera, universal primers can be easily designed for the group of interest. This avoids the problem of low PCR efficiency in amplification and extensive optimizations that can be time and resource intensive. Furthermore, species from a given group are likely to share genes and gene orders which will simplify sequence acquisition across multiple target taxa. In addition, specific barcodes could be controlled to a suitable length, which avoids the risk of ambiguous alignment caused by variable sequence length (Chase *et al.*, 2007).

This approach is simpler than obtaining super-barcodes for each sample, and many options are available to choose from for informative markers, such as genes, intergenic spacers, partial gene sequences, partial intergenic spacers and even sequences including partial gene sequences and partial intergenic spacers. Although there are over 300000 plant species (IUCN, 2012) if one particular barcode is selected per study group (a specific clade or genus for example), the total number of barcodes needed across all plants is likely to be accessible. In fact, specific DNA barcodes are likely to be shared at higher taxonomic levels making this approach even more appealing (Table 1).

Currently, when selecting plant barcodes for species-specific identification four main choices are available: evaluate candidate plastid markers proposed by CBOL (Kumar *et al.*, 2009; Wang *et al.*, 2010); choose commonly used markers in a given group (Zhang *et al.*, 2013); search mutational hotspots and loci by investigating the distribution of oligonucleotide repeat sequences and the relationships between repeats, indels and substitutions in a single representative plastid genome (Ahmed *et al.*, 2013); or use plastid-comparative analyses to select a suitable locus displaying adequate species-level divergence (Kuang *et al.*, 2011; Dong *et al.*, 2012). Specific barcodes focus on the latter method of finding barcodes for complete species-level resolution. A specific barcode may include one of the single-locus barcodes (e.g. *matK* or *PsbA-trnH*) or could be based on new markers that have never been used before.

The initial goal of DNA barcodes was to find a universal locus for the identification of all plants. However, there is no such universal barcode locus for land plants, especially in the chloroplast where lineage-specific evolution and non-random spatial patterns of substitution can occur (Ahmed *et al.*, 2013). That is why the specific-barcode approach relies on the use of dedicated cp-regions for each target group of species. In addition to genes and intergenic spacers, any DNA fragment with adequate variations (and not duplicated within the chloroplast to avoid analytical issues stemming from paralogy) can be used as a marker. While markers used in single-locus DNA barcodes such as the *rbcL* region can provide resolution at a higher taxonomic rank (e.g. family or genus), specific barcodes can assist species-level identifications, which is what we now typically require. Although some methods can address the issue of species boundaries in some particular plant groups (Joly, 2012; Becker *et al.*, 2013), the cp-genome sequence may not always suggest the same boundaries between species as those currently recognized by taxonomists. The availability of improved approaches to cp-genome analysis as proposed here will provide tools that should allow these issues to be explored more fully. This may not resolve these questions but should allow these taxonomic challenges to be more widely known and hopefully better understood.

The wide application of specific barcodes has two prerequisites: a rich database of cp-genome sequences (however these do not need to represent the fully annotated genome of the target taxa) and another database including primers for each plant group derived from the exploration of these cp-genomes. Known species could be distinguished by using the corresponding specific barcodes from the primer database. As for unknown species, two steps will be needed. First, unknown species are classified using single-locus barcodes (e.g. *rbcL*) at the family or genus levels. Second, the corresponding specific barcodes are chosen from cp-genome datasets to achieve discrimination at the species level (Fig. 1). This '1+1' model is different from the tiered approach (Newmaster *et al.*, 2006), especially in its second step. Although both the approaches include two steps potentially relying on two barcode loci, specific barcoding screens new markers in the second stage by comparing plastid genomes while the tiered approach relies on commonly used markers. The flexibility in choice of a specific barcode would have enormous advantages given the variation observed in substitution rates. In this respect, we may obtain a range of barcodes of similar value to CO1 in animals.

Obtaining sufficient plastid genome sequence is a critical step in identifying a suitable specific barcode from an alignment. Dong *et al.* (2012) scanned 12 entire cp-genomes to search for mutationally active regions to be used for

barcoding at the genus level. Ahmed *et al.* (2012) compared six plastid genomes from five genera to investigate the extent of genome-wide association between inverted repeats, indels, and substitutions in aroid cp-genomes. Yang *et al.* (2013) performed population-level phylogenomic analyses using eight cp-genome sequences from five *Cymbidium* species. We suggest 8–10 closely related plastid genome sequences from different species for alignment to search for specific barcodes. A specific barcode can then be selected at a genome-scale level for a certain group or specific lineage. If a close reference sequence is necessary, obtaining one plastid genome should be enough to support *de novo* sequence assembly.

Although the increased availability of published cp-genomes will facilitate the design of specific barcodes, current advances of NGS provide further opportunities for this approach. In cases where low diversity is expected (for example a recently radiated clade), one single NGS run of multiplexed DNA can be enough to identify phylogenetically informative sites. A study on over 80 rainforest tree species is currently exploring this approach (H. McPherson, personal communication).

VI. CONCLUSIONS

(1) DNA barcoding aims to find a single sequence to identify all species. Yet, no single-locus barcode can achieve the goal. In addition to inadequate variation and low PCR efficiency (often due to sequence variation in the primer binding regions), gene deletion is an important limiting factor for single loci preventing their use as a universal DNA barcode. For example, algae do not contain the *matK* sequence.

(2) Multi-locus markers have been assumed to be more successful in species identification, but studies to date demonstrated that these are also inadequate for universal plant identification. Despite significant recent effort, the development of single-locus barcodes has stalled, placing plant DNA barcoding at a crossroads. Fortunately, developments in DNA sequencing allowing cost-efficient plastid sequencing are driving plant identification into a post-barcode era.

(3) Whole-plastid-based barcodes have shown great potential in species discrimination, especially for closely related taxa. Continuing advances in sequencing technology may make these super-barcodes the method of choice for plant identification. Although routine technology is not yet established in many taxonomic laboratories, a choice is already possible between cost efficiency and practicality. Well-equipped laboratories can rely on in-house technical advances to reduce costs per base pair of sequence. Traditional laboratories can outsource NGS techniques at a higher cost but with the advantage of only having to provide plant material and follow-on bio-informatic analyses.

(4) The ultimate goal of DNA barcoding is to distinguish species rather than find a universal marker. Specific barcodes for each plant group suitable for application in traditional laboratories may be defined based upon the analysis of whole-chloroplast data. Specific barcoding is expected to become more widely used, providing fast and accurate molecular identifications at the species and population levels.

VII. ACKNOWLEDGEMENTS

We would like to thank Meijun Aoli for helping to prepare the figures and thank Yuanye Dang for her critical reading of the manuscript. This work was supported by Research Fund of University of Macau (077/2011/A3, 074/2012/A3) and Macao Science and Technology Development Fund (UL016/09Y4/CMS/WYT01/ICMS, MYRG208 (Y3-L4)-ICMS11-WYT).

VIII. REFERENCES

Ahmed, I., Biggs, P. J., Matthews, P. J., Collins, L. J., Hendy, M. D. & Lockhart, P. J. (2012). Mutational dynamics of aroid chloroplast genomes. *Genome Biology and Evolution* **4**, 1316–1323.

Ahmed, I., Matthews, P. J., Biggs, P. J., Naeem, M., McLenachan, P. A. & Lockhart, P. J. (2013). Identification of chloroplast genome loci suitable for high-resolution phylogeographic studies of *Colocasia esculenta* (L.) Schott (Araceae) and closely related taxa. *Molecular Ecology Resources* **13**, 929–937.

Álvarez, I. & Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* **29**, 417–434.

Bayly, M. J., Rigault, P., Spokevicius, A., Ladiges, P. Y., Ades, P. K., Anderson, C., Bossinger, G., Merchant, A., Udovicic, F. & Woodrow, I. E. (2013). Chloroplast genome analysis of Australian eucalypts– Eucalyptus, Corymbia, Angophora, Allosyncarpia and Stockwellia (Myrtaceae). *Molecular Phylogenetics and Evolution* **69**, 704–716.

Becker, M., Gruenheit, N., Steel, M., Voelckel, C., Deusch, O., Heenan, P. B., McLenachan, P. A., Kardailsky, O., Leigh, J. W. & Lockhart, P. J. (2013). Hybridization may facilitate in situ survival of endemic species through periods of climate change. *Nature Climate Change* **3**, 1039–1043.

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., Ingram, K. K. & Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution* **22**, 148–155.

Blaxter, M. L. (2004). The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **359**, 669–679.

CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 12794–12797.

China Plant BOL Group (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 19641–19646.

Chang, C. C., Lin, H. C., Lin, I. P., Chow, T. Y., Chen, H. H., Chen, W. H., Cheng, C. H., Lin, C. Y., Liu, S. M. & Chaw, S. M. (2006). The chloroplast genome of Phalaenopsis aphrodite (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. *Molecular Biology and Evolution* **23**, 279–291.

Chase, M. W., Cowan, R. S., Hollingsworth, P. M., van den Berg, C., Madrinan, S., Petersen, G., Seberg, O., Jorgsensen, T., Cameron, K. M., Carine, M., Pedersen, N., Hedderson, T. A. J., Conrad, F., Salazar, G. A., Richardson, J. E., Hollingsworth, M. L., Barraclough, T. G., Kelly, L. & Wilkinson, M. (2007). A proposal for a standardised protocol to barcode all land plants. *Taxon* **56**, 295–299.

Chase, M. W. & Fay, M. F. (2009). Barcoding of plants and fungi. *Science* **325**, 682–683.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y. & Leon, C. (2010). Validation

of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE* **5**, e8613.

Chiou, S. J., Yen, J. H., Fang, C. L., Chen, H. L. & Lin, T. Y. (2007). Authentication of medicinal herbs using PCR-amplified ITS2 with specific primers. *Planta Medica* **73**, 1421–1426.

Coyle, H. M., Lee, C., Lin, W., Lee, H. C. & Palmbach, T. M. (2005). Forensic botany: using plant evidence to aid in forensic death investigation. *Croatian Medical Journal* **46**, 606–612.

Cuénoud, P., Savolainen, V., Chatrou, L. W., Powell, M., Grayer, R. J. & Chase, M. W. (2002). Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences. *American Journal of Botany* **89**, 132–144.

Devey, D. S., Chase, M. W. & Clarkson, J. J. (2009). A stuttering start to plant DNA barcoding: microsatellites present a previously overlooked problem in non-coding plastid regions. *Taxon* **58**, 7–15.

Diekmann, K., Hodkinson, T. R., Fricke, E. & Barth, S. (2008). An optimized chloroplast DNA extraction protocol for grasses (Poaceae) proves suitable for whole plastid genome sequencing and SNP detection. *PLoS ONE* **3**, e2813.

Dong, W., Liu, J., Yu, J., Wang, L. & Zhou, S. (2012). Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS ONE* **7**, e35071.

Doorduin, L., Gravendeel, B., Lammers, Y., Ariyurek, Y., Chin-A-Woeng, T. & Vrieling, K. (2011). The complete chloroplast genome of 17 individuals of pest species jacobaea vulgaris: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. *DNA Research* **18**, 93–105.

Ebihara, A., Nitta, J. H. & Ito, M. (2010). Molecular species identification with rich floristic sampling: DNA barcoding the pteridophyte flora of Japan. *PLoS ONE* **5**, e15136.

Edwards, D., Horn, A., Taylor, D., Savolainen, V. & Hawkins, J. A. (2008). DNA barcoding of a large genus, Aspalathus L. (Fabaceae). *Taxon* **57**, 1317–1327.

Erickson, D. L., Spouge, J., Resch, A., Weigt, L. A. & Kress, J. W. (2008). DNA barcoding in land plants: developing standards to quantify and maximize success. *Taxon* **57**, 1304–1316.

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., Percy, D. M., Hajibabaei, M. & Barrett, S. C. H. (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE* **3**, e2802.

Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., Percy, D. M., Graham, S. W., Barrett, S. C. H., Newmaster, S. G., Hajibabaei, M. & Husband, B. C. (2009). Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources* **9**, 130–139.

Galimberti, A., De Mattia, F., Losa, A., Bruni, I., Federici, S., Casiraghi, M., Martellos, S. & Labra, M. (2012). DNA barcoding as a new tool for food traceability. *Food Research International* **50**, 55–63.

Gao, T., Sun, Z., Yao, H., Song, J., Zhu, Y., Ma, X. & Chen, S. (2011). Identification of fabaceae plants using the DNA barcode matK. *Planta Medica* **77**, 92–94.

Gao, T., Yao, H., Song, J., Liu, C., Zhu, Y., Ma, X., Pang, X., Xu, H. & Chen, S. (2010a). Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. *Journal of Ethnopharmacology* **130**, 116–121.

Gao, T., Yao, H., Song, J., Zhu, Y., Liu, C. & Chen, S. (2010b). Evaluating the feasibility of using candidate DNA barcodes in discriminating species of the large Asteraceae family. *BMC Evolutionary Biology* **10**, 324.

Gielly, L. & Taberlet, P. (1994). The use of chloroplast DNA to resolve plant phylogenies: noncoding versus rbcL sequences. *Molecular Biology and Evolution* **11**, 769–777.

Gregory, T. R. (2005). DNA barcoding does not compete with taxonomy. *Nature* **434**, 1067.

Gruenheit, N., Lockhart, P. J., Steel, M. & Martin, W. (2008). Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. *Molecular Biology and Evolution* **25**, 1512–1520.

Hebert, P. D. N., Cywinska, A., Ball, S. L. & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, Series B: Biological Sciences* **270**, 313–321.

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 14812–14817.

Heinze, B. (2007). A database of PCR primers for the chloroplast genomes of higher plants. *Plant Methods* **3**, 4–10.

Hollingsworth, P. M. (2008). DNA barcoding plants in biodiversity hot spots: progress and outstanding questions. *Heredity* **101**, 1–2.

Hollingsworth, M. L., Andra Clark, A., Forrest, L. L., Richardson, J., Pennington, R. T., Long, D. G., Cowan, R., Chase, M. W., Gaudeul, M. & Hollingsworth, P. M. (2009). Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Molecular Ecology Resources* **9**, 439–457.

Hollingsworth, P. M., Graham, S. W. & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE* **6**, e19254.

Huang, C. Y., Grünheit, N., Ahmadinejad, N., Timmis, J. N. & Martin, W. (2005). Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant Physiology* **138**, 1723–1733.

Huxley-Jones, E., Shaw, J. L., Fletcher, C., Parnell, J. & Watts, P. C. (2012). Use of DNA barcoding to reveal species composition of convenience seafood. *Conservation Biology* **26**, 367–371.

IUCN (2012). The IUCN red list of threatened species, Version 2012.2. Available at http://www.iucnredlist.org Accessed 25.03.2013.

Joly, S. (2012). JML: testing hybridization from species trees. *Molecular Ecology Resources* **12**, 179–184.

Joly, S., McLenachan, P. A. & Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *American Naturalist* **174**, E54–E70.

Kane, N. C. & Cronk, Q. (2008). Botany without borders: barcoding in focus. *Molecular Ecology* **17**, 5175–5176.

Kane, N., Sveinsson, S., Dempewolf, H., Yang, J. Y., Zhang, D., Engels, J. M. M. & Cronk, Q. (2012). Ultra-barcoding in cacao (theobroma spp.; malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* **99**, 320–329.

Kress, W. J. & Erickson, D. L. (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**, e508.

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A. & Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 8369–8374.

Kuang, D. Y., Wu, H., Wang, Y. L., Gao, L. M., Zhang, S. Z. & Lu, L. (2011). Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics. *Genome* **54**, 663–673.

Kumar, S., Hahn, F. M., McMahan, C. M., Cornish, K. & Whalen, M. C. (2009). Comparative analysis of the complete sequence of the plastid genome of Parthenium argentatum and identification of DNA barcodes to differentiate Parthenium species and lines. *BMC Plant Biology* **9**, 131–142.

Lahaye, R., Van Der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T. G. & Savolainen, V. (2008). DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 2923–2928.

Liu, J., Möller, M., Gao, L. M., Zhang, D. Q. & Li, D. Z. (2011). DNA barcoding for the discrimination of Eurasian yews (Taxus L., Taxaceae) and the discovery of cryptic species. *Molecular Ecology Resources* **11**, 89–100.

Lockhart, P. & Steel, M. (2005). A tale of two processes. *Systematic Biology* **54**, 948–951.

Luo, K., Chen, S. L., Chen, K. L., Song, J. Y., Yao, H., Ma, X., Zhu, Y. J., Pang, X. H., Yu, H., Li, X. W. & Liu, Z. (2010). Assessment of candidate plant DNA barcodes using the Rutaceae family. *Science China Life Sciences* **53**, 701–708.

Luo, H., Shi, J., Arndt, W., Tang, J. & Friedman, R. (2008). Gene order phylogeny of the genus Prochlorococcus. *PLoS ONE* **3**, e3837.

Luo, H., Sun, Z., Arndt, W., Shi, J., Friedman, R. & Tang, J. (2009). Gene order phylogeny and the evolution of methanogens. *PLoS ONE* **4**, e6069.

Ma, X. Y., Xie, C. X., Liu, C., Song, J. Y., Yao, H., Luo, K., Zhu, Y. J., Gao, T., Pang, X. H., Qian, J. & Chen, S. L. (2010). Species identification of medicinal pteridophytes by a DNA barcode marker, the chloroplast psbA-trnH intergenic region. *Biological and Pharmaceutical Bulletin* **33**, 1919–1924.

Magee, A. M., Aspinall, S., Rice, D. W., Cusack, B. P., Semon, M., Perry, A. S., Stefanovic, S., Milbourne, D., Barth, S., Palmer, J. D., Gray, J. C., Kavanagh, T. A. & Wolfe, K. H. (2010). Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research* **20**, 1700–1710.

McPherson, H., van der Merwe, M., Delaney, S. K., Edwards, M. A., Henry, R. J., McIntosh, E., Rymer, P. D., Milner, M. L., Siow, J. & Rossetto, M. (2013). Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. *BMC Ecology* **13**, 8.

van der Merwe, M., McPherson, H., Siow, J. & Rossetto, M. (2013). Next-Gen phylogeography of rainforest trees: exploring landscape-level cpDNA variation from whole-genome sequencing. *Molecular Ecology Resources* **14**, 199–208.

Mildenhall, D. (2006). Hypericum pollen determines the presence of burglars at the scene of a crime: an example of forensic palynology. *Forensic Science International* **163**, 231–235.

Miller, S. E. (2007). DNA barcoding and the renaissance of taxonomy. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 4775–4776.

Min, X. J. & Hickey, D. A. (2007). Assessing the effect of varying sequence length on DNA barcoding of fungi. *Molecular Ecology Notes* **7**, 365–373.

Newmaster, S. G., Fazekas, A. J. & Ragupathy, S. (2006). DNA barcoding in land plants: evaluation of rbcL in a multigene tiered approach. *Canadian Journal of Botany* **84**, 335–341.

Newmaster, S. G., Fazekas, A. J., Steeves, R. A. D. & Janovec, J. (2008). Testing candidate plant barcode regions in the Myristicaceae. *Molecular Ecology Resources* **8**, 480–490.

Nock, C. J., Waters, D. L., Edwards, M. A., Bowen, S. G., Rice, N., Cordeiro, G. M. & Henry, R. J. (2011). Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal* **9**, 328–333.
Pang, X., Song, J., Zhu, Y., Xie, C. & Chen, S. (2010). Using DNA barcoding to identify species within euphorbiaceae. *Planta Medica* **76**, 1784–1786.
Pang, X., Song, J., Zhu, Y., Xu, H., Huang, L. & Chen, S. (2011). Applying plant DNA barcodes for Rosaceae species identification. *Cladistics* **27**, 165–170.
Parks, M., Cronn, R. & Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* **7**, 84–100.
Pennisi, E. (2007). Wanted: a barcode for plants. *Science* **318**, 190–191.
Renner, S. S. (1999). Circumscription and phylogeny of the Laurales: evidence from molecular and morphological data. *American Journal of Botany* **86**, 1301–1315.
Rohland, N. & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research* **22**, 939–946.
Salazar, G. A., Chase, M. W., Soto Arenas, M. A. & Ingrouille, M. (2003). Phylogenetics of Cranichideae with emphasis on Spiranthinae (Orchidaceae, Orchidoideae): evidence from plastid and nuclear DNA sequences. *American Journal of Botany* **90**, 777–795.
Sass, C., Little, D. P., Stevenson, D. W. & Specht, C. D. (2007). DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. *PLoS ONE* **2**, e1154.
Selvaraj, D., Sarma, R. K. & Sathishkumar, R. (2008). Phylogenetic analysis of chloroplast matK gene from Zingiberaceae for plant DNA barcoding. *Bioinformation* **3**, 24–27.
Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., Liu, W., Miller, J., Siripun, K. C., Winder, C. T., Schilling, E. E. & Small, R. L. (2005). The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* **92**, 142–166.
Shaw, J., Lickey, E. B., Schilling, E. E. & Small, R. L. (2007). Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* **94**, 275–288.
Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology* **26**, 1135–1145.
Shi, C., Hu, N., Huang, H., Gao, J., Zhao, Y. J. & Gao, L. Z. (2012). An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLoS ONE* **7**, e31468.
Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayshida, N., Matsubayasha, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B. Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kata, A., Tohdoh, N., Shimada, H. & Sugiura, M. (1986). The complete nucleotide sequence of the tobacco chloroplast genome. *Plant Molecular Biology Reporter* **4**, 111–148.
Song, J., Shi, L., Li, D., Sun, Y., Niu, Y., Chen, Z., Luo, H., Pang, X., Sun, Z., Liu, C., Lv, A., Deng, Y., Larson-Rabin, Z., Wilkinson, M. & Chen, S. (2012). Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal DNA. *PLoS ONE* **7**, e43971.
Stech, M. & Quandt, D. (2010). 20,000 species and five key markers: the status of molecular bryophyte phylogenetics. *Phytotaxa* **9**, 196–228.
Stewart, C. N. Jr. (2005). Monitoring the presence and expression of transgenes in living plants. *Trends in Plant Science* **10**, 390–396.
Stoeckle, M. (2003). Taxonomy, DNA, and the bar code of life. *BioScience* **53**, 796–797.
Straub, S., Fishbein, M., Livshultz, T., Foster, Z., Parks, M., Weitemier, K., Cronn, R. & Liston, A. (2011). Building a model: developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. *BMC Genomics* **12**, 211–232.
Stull, G. W., Moore, M. J., Mandala, V. S., Douglas, N. A., Kates, H.-R., Qi, X., Brockington, S. F., Soltis, P. S., Soltis, D. E. & Gitzendanner, M. A. (2013). A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences* **1**, 1200497.
Sucher, N. J. & Carles, M. C. (2008). Genome-based approaches to the authentication of medicinal plants. *Planta Medica* **74**, 603–623.
Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J. E. & Taberlet, P. (2009). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular Ecology Resources* **9**, 51–60.
Vijayan, K. & Tsou, C. (2010). DNA barcoding in plants: taxonomy in a new perspective. *Current Science (Bangalore)* **99**, 1530–1541.
Wang, W., Wu, Y., Yan, Y., Ermakova, M., Kerstetter, R. & Messing, J. (2010). DNA barcoding of the Lemnaceae, a family of aquatic monocots. *BMC Plant Biology* **10**, 205–215.
Whitlock, B. A., Hale, A. M. & Groff, P. A. (2010). Intraspecific inversions pose a challenge for the trnH-psbA plant DNA barcode. *PLoS ONE* **5**, e11533.
van de Wiel, C. C. M., van der Schoot, J., van Valkenburg, J. L. C. H., Duistermaat, H. & Smulders, M. J. M. (2009). DNA barcoding discriminates the noxious invasive plant species, floating pennywort (Hydrocotyle ranunculoides L.f.), from non-invasive relatives. *Molecular Ecology Resources* **9**, 1086–1091.
Wu, F. H., Chan, M. T., Liao, D. C., Hsu, C. T., Lee, Y. W., Daniell, H., Duvall, M. R. & Lin, C. S. (2010). Complete chloroplast genome of Oncidium Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. *BMC Plant Biology* **10**, 68–79.
Wu, C. S., Wang, Y. N., Hsu, C. Y., Lin, C. P. & Chaw, S. M. (2011). Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. *Genome Biology and Evolution* **3**, 1284–1295.
Yamaguchi, A., Kawamura, H. & Horiguchi, T. (2006). A further phylogenetic study of the heterotrophic dinoflagellate genus, Protoperidinium (Dinophyceae) based on small and large subunit ribosomal RNA gene sequences. *Phycological Research* **54**, 317–329.
Yang, J. B., Tang, M., Li, H. T., Zhang, Z. R. & Li, D. Z. (2013). Complete chloroplast genome of the genus Cymbidium: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evolutionary Biology* **13**, 84.
Yao, H., Song, J. Y., Ma, X. Y., Liu, C., Li, Y., Xu, H. X., Han, J. P., Duan, L. S. & Chen, S. L. (2009). Identification of dendrobium species by a candidate DNA barcode sequence: the chloroplast psbA-trnH intergenic region. *Planta Medica* **75**, 667–669.
Yoon, C. K. (1993). Forensic science. Botanical witness for the prosecution. *Science* **260**, 894–895.
Zhang, W., Fan, X., Zhu, S., Zhao, H. & Fu, L. (2013). Species-specific identification from incomplete sampling: applying DNA barcodes to monitoring invasive solanum plants. *PLoS ONE* **8**, e55927.
Zhong, B., Deusch, O., Goremykin, V. V., Penny, D., Biggs, P. J., Atherton, R. A., Nikiforova, S. V. & Lockhart, P. J. (2011). Systematic error in seed plant phylogenomics. *Genome Biology and Evolution* **3**, 1340–1348.

(*Received 5 June 2013; revised 5 February 2014; accepted 27 February 2014; published online 25 March 2014*)