Received: 6 April 2022 | Revised: 2 August 2022 | Accepted: 3 August 2022

DOI: 10.1111/1755-0998.13697

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Accurate identification of taxon-specific molecular markers in plants based on DNA signature sequence

Zhongyi Hua[1] | Chao Jiang[1] | Shuhui Song[2,3] | Dongmei Tian[2,3] | Ziyuan Chen[1] | Yan Jin[1] | Yuyang Zhao[1] | Junhui Zhou[1] | Zhang Zhang[2,3] | Luqi Huang[1] | Yuan Yuan[1]

[1]National Resource Center for Chinese Materia Medica, Chinese Academy of Chinese Medical Sciences (CACMS), Beijing, China

[2]China National Center for Bioinformation, Beijing, China

[3]National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences (CAS), Beijing, China

Correspondence
Yuan Yuan and Luqi Huang, National Resource Center for Chinese Materia Medica, Chinese Academy of Chinese Medical Sciences (CACMS), Beijing 100700, China.
Emails: y_yuan0732@163.com; huangluqi01@126.com

Zhang Zhang, China National Center for Bioinformation, Beijing 100101, China.
Email: zhangzhang@big.ac.cn

Funding information
Key project at central government level for the ability establishment of sustainable use for valuable Chinese medicine resources, Grant/Award Number: 2060302; Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences, Grant/Award Number: CI2021B014/CI2021A041; Special Funds for Basic Resources Investigation Research of the Ministry of Science and Technology, Grant/Award Number: 2018FY10080002; The National Key Research and Development Program of China, Grant/Award Number: 2019YFC1711000; Key project at central government level for the ability establishment of sustainable use for valuable Chinese medicine resources, Grant/Award Number: 2060302; Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences, Grant/Award Number: CI2021B014/CI2021A041; Special Funds for Basic Resources Investigation Research of the Ministry of Science and Technology, Grant/Award Number: 2018FY10080002; The National Key Research and Development Program of China, Grant/Award Number: 2019YFC1711000

Handling Editor: SUHUA SHI

Abstract

Accurate identification of plants remains a significant challenge for taxonomists and is the basis for plant diversity conservation. Although DNA barcoding methods are commonly used for plant identification, these are limited by the low amplification success and low discriminative power of selected genomic regions. In this study, we developed a *k*-mer–based approach, the DNA signature sequence (DSS), to accurately identify plant taxon-specific markers, especially at the species level. DSS is a constant-length nucleotide sequence capable of identifying a taxon and distinguishing it from other taxa. In this study, we performed the first large-scale study of DSS markers in plants. DSS candidates of 3899 angiosperm plant species were calculated based on a chloroplast data set with 4356 assemblies. Using Sanger sequencing of PCR amplicons and high-throughput sequencing, DSSs were validated in four and 165 species, respectively. Based on this, the universality of the DSSs was over 79.38%. Several indicators influencing DSS marker identification and detection have also been evaluated, and common criteria for DSS application in plant identification have been proposed.

KEYWORDS
chloroplast genome, DNA signature sequence, species identification, plants

Zhongyi Hua, Chao Jiang and Shuhui Song: Joint authors

1 | INTRODUCTION

Vascular plants include more than 300,000 known species, playing an irreplaceable role in terrestrial ecosystems and offering benefits, such as food, fuel, and medcine (Chase et al., 2016; Christenhusz & Byng, 2016; Ebert & Engels, 2020). The need for sustainable development and plant biodiversity conservation has prompted international initiatives aimed at developing accurate plant species identification methods. Several methods have been proposed over the past years based on genetic information (CBOL Plant Working Group, 2009; Padial et al., 2010), pollen (Martin & Harvey, 2017), chemistry profiles (Senizza et al., 2019), and using artificial intelligence techniques (Wäldchen et al., 2018). However, the accurate identification of plant species remains a significant challenge because of the tremendous diversity of plants.

Among the existing methods, DNA-based methods have become increasingly popular, and various molecular markers have been developed. Generally, these DNA markers can be classified into two types: taxon-specific and DNA barcode markers. Taxon-specific markers, such as derived cleaved amplified polymorphic sequences (dCAPS) and sequence characterized amplified regions (SCAR), were applied based on the presence–absence variance (PAV) of the marker. In contrast to taxon-specific markers, DNA barcodes use universal primers instead of taxon-specific primers to amplify fragments and identify species based on nucleotide dissimilarity (Hebert et al., 2003). Since the launch of the Barcode of Life Data System (BOLD), DNA barcoding has become a globally accepted method for taxonomists, ecologists, and conservation biologists (Ratnasingham & Hebert, 2007). Two chloroplast DNA barcodes (*rbcL* + *matK*) have been recommended as core plant barcodes (CBOL Plant Working Group, 2009) and the internal transcribed spacer (ITS) and ITS2 with higher discriminatory power have subsequently been proposed as complementary markers (Li et al., 2011). The universal primers of DNA barcodes also enable the combination of DNA barcoding and high-throughput sequencing (HTS), which can identify species from mixtures containing multiple taxa (Piñol et al., 2019; Taberlet et al., 2012). Nonetheless, introgression induced by gene flow (Zhang et al., 2018), incomplete lineage sorting due to retention of ancestral polymorphisms (Goetze et al., 2017), adaptive radiation triggered by natural selection (Guo et al., 2020), and numerous other ongoing processes are common in plant evolution, consequently making DNA barcoding incapable of clustering conspecifics with clear discontinuities from other species (Hollingsworth et al., 2016), and therefore, unable to provide well-defined species boundaries such as taxon-specific markers.

Taxon-specific *k*-mers, also referred to as DNA signature sequences (DSS), are nucleotide sequences with constant length for identifying species by PAV and are promising for plant species identification. Originally, DSSs were used for microbial identification (Tu et al., 2013, 2014) and have become emerging molecular markers for plant identification because of the overwhelming accumulation of chloroplast genomes (Raime et al., 2020; Raime & Remm, 2018). Nevertheless, whether DSS markers can be used over a wide taxonomic range, that is, its universality, remains poorly explored. HTS technologies make it feasible and desirable to identify DSS markers from several publicly available chloroplast genomes. The purpose of this study was to conduct large-scale data analysis to identify DSS markers from chloroplasts, confirm their universality and discriminatory power, and compare them with existing methods.

2 | MATERIALS AND METHODS

2.1 | Collection of plant chloroplast genomes

A total of 4356 chloroplast assemblies covering 3899 species (216 families, 1512 genera) were used in this study. Of these, 3623 assemblies covering 3535 species were obtained from NCBI Organelle Genome Resources (https://www.ncbi.nlm.nih.gov/genome/organelle/), 16 assemblies covering 16 species were downloaded from the Genome Warehouse (GWH) at the National Genomics Data Center (Chen et al., 2021; CNCB-NGDC Members and Partners, 2022), and 717 assemblies covering 434 species were sequenced and assembled in our study and then deposited in GWH. All data from the public databases were accessed on 30 June 2020. To minimize potential errors caused by plant name synonyms, the species name of each assembly collected from the public database was curated in accordance with The Catalogue of Life Checklist 2021 (Bánki et al., 2021). The full list of accession numbers for the plastid genome sequences analysed in this study is summarized in Table S1.

2.2 | An in silico pipeline for DSS candidate identification

We have developed an in silico pipeline called IdenDSS for DSS candidate identification (https://github.com/Hua-CM/IdenDSS). IdenDSS requires two categories of species as input: the target species whose DSSs need to be identified and the background species from which the target species should be distinguished. In brief, four steps were followed for identifying DSS candidates. The first step was to generate all possible *k*-mers from one of the target species' chloroplast assemblies using the sliding window method by setting a fixed-length (default 40 bp) sliding window with a 1-bp step. Second, after deredundancy, the nonredundant *k*-mers were blasted against other assemblies of the target species to identify *k*-mers conserved in the target species. These conserved non-redundant *k*-mers were then blasted against the chloroplast assemblies of the background species. Lastly, *k*-mers present in the background species assemblies were removed, and the rest were considered as DSS candidates.

2.3 | Validating DSS using HTS data sets

The DSSs were validated according to a previously published procedure with minor modifications (Raime et al., 2020). For a specific

species, HTS data sets used for DSS validation were categorized into two types: conspecific data sets, containing whole-genome sequencing (WGS) reads from conspecific samples, and background data sets, containing WGS reads from other species. The background data set is further divided into two types: the common data set: an HTS data set containing WGS reads from 12 common fruits, vegetables, and crops; and the congeneric data set: an HTS data set containing WGS reads from congeneric species. The common data set was downloaded from the SRA (detailed in Table S2), whereas other data sets were generated in this study according to the methods described in Library Construction and high-throughout sequencing and deposited in the GSA (Wang et al., 2017) at the NGDC (CNCB-NGDC Members and Partners, 2022) under the accession number CRA004065 (detailed in Table S3). Only DSS candidates of 165 species were validated because the tested species must have a conspecific data set and a congeneric data set, but most of the 3899 species did not have the corresponding WGS data set. FastK software was used (Myers, 2020) to determine whether DSS candidates appeared in the HTS data sets. To exclude questionable or incorrect DSS due to sequencing errors, a "detectable" *k*-mer was identified if it could be detected at least twice. If a DSS candidate could only be detected in a conspecific data set but not in the background data set, the DSS candidate was considered as a DSS. For each species, the precision of the DSS candidate was calculated using the following equation:

$$\text{Precision} = \frac{\text{No. of DSSs}}{\text{No. of DSS candidates}}$$

We created random subsets with different numbers of sequenced nucleotides (10^6, 10^7, 10^8, 10^9, and 5×10^9) from the original FASTQ files and validated the DSS using each subset.

2.4 | Evaluating the sampling number of DSS candidates for achieving a DSS

DSS candidates and DSSs validated in 165 species were used to evaluate the sampling number of DSS candidates required to achieve a DSS. Specifically, a set of species (i.e., 10, 20, 30, 40, 50, 60, 70, 80, and 90) were sampled from the 165 species. For each species, different numbers of DSS candidates (i.e., 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100) from different combined DSS candidates were sampled. The proportion of species that possessed at least one DSS among the sampled species was recorded. We performed a bootstrap internal validation procedure with 1000 bootstrap resamples.

2.5 | Validating DSS using PCR amplification and sanger sequencing

Two case studies were conducted to investigate the applicability of validating DSS using polymerase chain reaction (PCR) amplification and Sanger sequencing. The first case study was conducted on *Panax ginseng* and *P. quinquefolius* as these were the predominant and valuable *Panax* species. The global ginseng market is worth more than US$ 2 billion (Baeg & So, 2013). Although multiple DNA markers for *Panax* have been reported (Ji et al., 2019; Nguyen et al., 2017), identification methods using multilocus DNA barcode markers or super DNA barcodes based on complete chloroplast genomes are neither convenient nor economical. The second case study involved *Atractylodes japonica*, *A. macrocephala*, and *A. lancea*. *A. lancea* has a long history of use as an important herb in eastern Asia. However, several congeneric species, such as *A. japonica* and *A. macrocephala*, are frequently found as adulterants in *A. lancea* rhizomes, either deliberately or unintentionally. Currently, there is no established molecular marker for the identification of *A. lancea*.

Plant DNA was extracted using the DNeasy Plant Mini Kit (Qiagen Co. Ltd.). The selected DSSs and corresponding primer details used in the case studies are presented in Table S4. All primers were designed using Primer3 (Untergasser et al., 2012). Briefly, the 25-μl PCR system contained 12.5 μl 2x Taq Master Mix, 1.0 μl PCR forward primer (10 μM), 1.0 μl PCR reverse primer (10 μM), 1.0 μl cDNA, and 9.5 μl dH_2O. PCR amplification was conducted on a Veriti 96 PCR system (Applied Biosystems) under the following conditions: initial denaturation at 95°C for 10 min; 35 cycles of amplification at 94°C for 15 s, 55°C for 10 s, and 72°C for 30 s; and a final extension at 72°C for 5 min. PCR products were evaluated using 1.5% agarose gel electrophoresis. Positive PCR products were purified using a PCR Products Purification Kit (Spin-column) (TransGen) and bidirectionally sequenced using PCR-derived primers. The dideoxy chain termination method was performed at the Beijing Genomics Institute, China.

2.6 | Library construction and high-throughput sequencing

DNA was extracted as described in the previous section. In all HTS experiments performed in our study, sequencing libraries were generated using the NEB Next Ultra DNA Library Prep Kit for Illumina (NEB) following the manufacturer's recommendations, and library quality was assessed on the Agilent Bioanalyser 2100 system (Agilent Technologies). The 150-bp paired-end reads were generated on an Illumina Hiseq 4000 platform (Illumina Inc.).

2.7 | Identification of DNA barcode sequences

The barcode regions and corresponding primers used in this study are listed in Table S5. To identify the target barcodes for each assembly, BLAST (version 2.9.0+) was applied with the following parameters: -task blastn-short -evalue 10. The sequences between the primer target sequences in the assembly were considered as barcode sequences.

2.8 | Evaluation of DNA barcodes' universality and discriminating power

First, all available angiosperm sequences of the barcodes (as of 11 November 2020) were downloaded from the NCBI nucleotide database using query expressions detailed in Table S6. For example, the following query expression was used for *rbcL*: "((rbcL [Title]) AND Magnoliopsida [Organism]) AND 200:4000 [Sequence Length]". Second, the obtained records were filtered to remove sequences whose origin plants were not in the chloroplast data set. Third, the filtered sequences were used as query sequences, and the BLAST+ program (version 2.9.0+) was applied to query the reference database for each sequence with an E-value of less than 1×10^{-5}. The evaluation was only carried out in the barcode region, with records from more than 100 species. Identification was considered successful if the query sequence had the closest match with a conspecific individual in the reference database. The bootstrap method was applied to calculate bias-corrected 95% confidence intervals with 1000 *bootstrap* replicates.

3 | RESULTS

3.1 | Identification of DSS candidates in 3899 angiosperms

The process of identifying DSSs is divided into two parts: identifying DSS candidates and validating the DSSs from these candidates (Figure 1). We do this because, while the bioinformatic pipeline for identifying DSS candidates in silico could be standardized, researchers may weigh and consider a variety of factors when selecting appropriate techniques for validating DSSs, such as economics and technical difficulty. Nonetheless, an important prerequisite to obtaining the DSSs of a specific species is obtaining an adequate number of DSS candidates.

We first introduced a pipeline called IdenDSS for identifying DSS candidates, which overcomes two major issues in the previous method (Raime et al., 2020). (1) Every time the target or background species are altered, the database must be recreated, requiring the creation of 3899 databases to identify DSS candidates for 3899 species; and (2) the maximum length of *k*-mer was restricted to 32 bp in the previous method (Kaplinski et al., 2015).

Using the IdenDSS pipeline, we assessed and optimized four factors that may affect the identification of DSS candidates (DSS length, assembly numbers, background species number, and related species). At 20-, 30-, 40-, 50-, and 60-bp *k*-mer length, the numbers of species without DSS candidates were 130, 106, 91, 89, and 95, respectively (Figure 2a, Table S7). Based on the above results, a shorter *k*-mer length tends to yield more species without DSSs. We noticed that there were 83 species with no DSS candidates for all five *k*-mer lengths (Table S7). Of the 83, 62 had multiple assemblies. The proportion of species with multiple assemblies was greater than the proportion of species with multiple assemblies in the entire data set (Fisher's test, $p < .01$). We also found that for all five testing lengths, species with only one assembly presented more DSS candidates than species with multiple assemblies (Figure 2b, Table S8). This result suggests that assembly number has a substantial impact on the identification of DSS candidates. For 369 species with multiple assemblies, our results showed that when the *k*-mer length was 40 bp, the number of species with DSS candidates was the highest, and the proportion of species with DSS candidates was 82.38% (Figure 2c). We also used these 369 species to investigate whether the number of DSS candidates decreased with increasing numbers of background species. Our results showed that the number of DSS candidates remained stable when the number of background species was >1000 (Table S9). As molecular markers for closely related species are known to be limited, we further used congeneric species to evaluate their effect on the identification of DSS candidates. Aside from the species that had no sequenced congeneric species, the number of congeneric species had no significant influence on the distribution of species without DSS candidates ($p > .05$, Kolmogorov–Smirnov test, Figure 2d). However, this result has limitations because the species with sequenced congeneric species used in our study represent only a small fraction of plants, and even among these, closely related species cannot be fully portrayed by congeneric species because the degree of species differentiation varies across genera. We will further investigate the influence of closely related species in future case studies. Based on the aforementioned results, we propose two principles for the identification of DSS candidates in plants: (1) Species to identify DSS should have at least two chloroplast assemblies; the more assemblies, the better the results; and (2) the optimal *k*-mer length is 40 bp.

In addition, we found that DSS candidates could appear successively. For example, when the *k*-mer length is set to 40 bp, supposing that there are three DSS candidates identified from positions 1–40, 2–41, and 3–42, these three DSS candidates form a "combined DSS candidate". Thus, we investigated the combined DSSs in the collected genomes (Figure S1-S2) and found that the number of combined DSS candidates was positively correlated with the number of DSS candidates.

3.2 | DSS validation using PCR and sanger sequencing

3.2.1 | Case study 1

Because our data set comprised only one *P. ginseng* assembly and one *P. quinquefolius* assembly, we added an extra *P. ginseng* assembly (accession number: KM067390.1) and a *P. quinquefolius* assembly (accession number: KT028714) in this case study based on the principles proposed above for identifying DSS candidates. A total of 1681 and 2041 DSS candidates (from 64 and 69 combined DSS candidates, respectively) were identified in *P. ginseng* and *P. quinquefolius*, respectively. Taking efficiency and cost into account, five *P. ginseng* and five *P. quinquefolius* DSSs were randomly selected and tested. We found that all five *P. ginseng* DSS candidates and four of the five *P. quinquefolius* DSS candidates were successfully

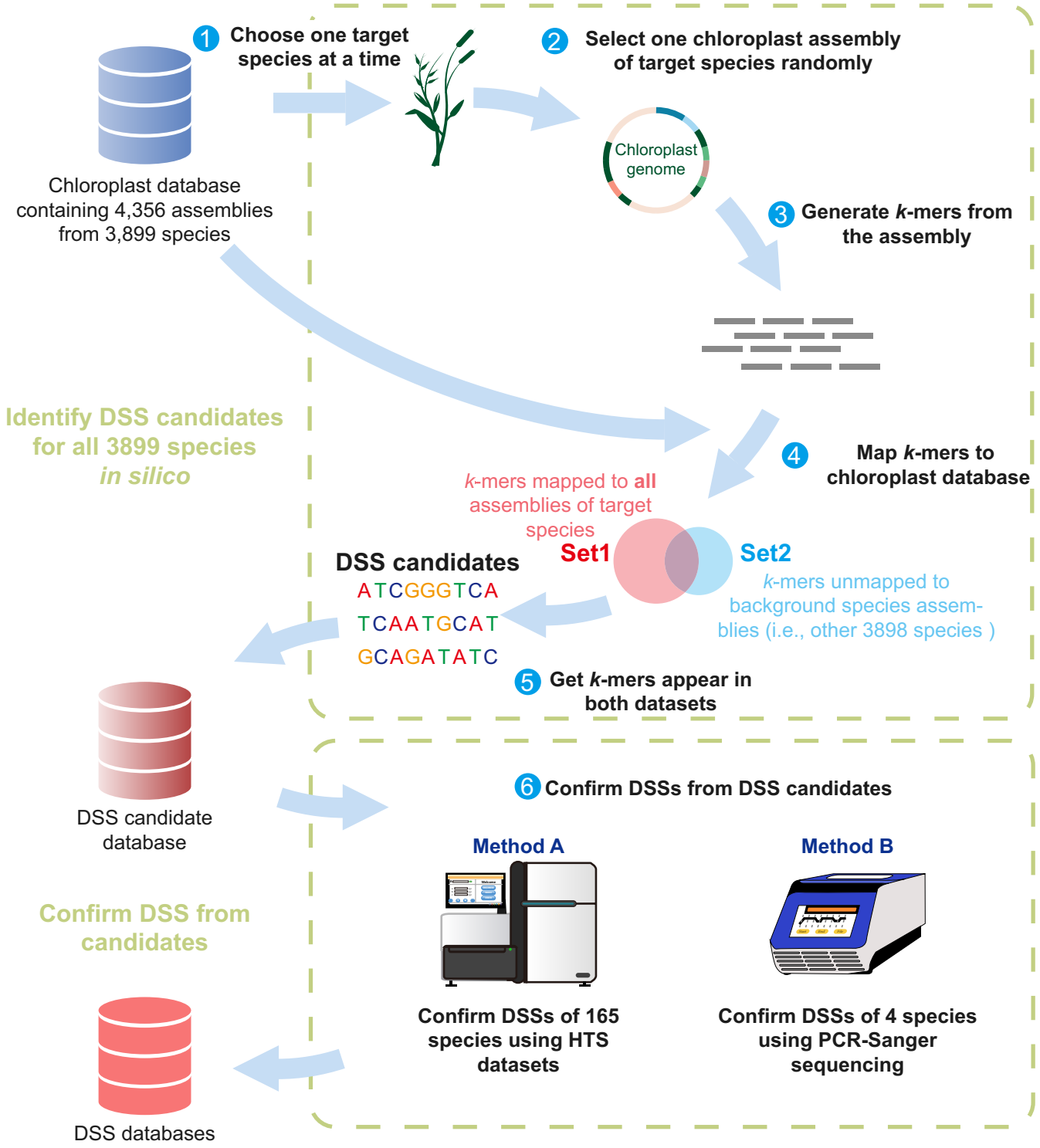

FIGURE 1 Workflow for identifying DNA signature sequence (DSS) in the present study. First, the chloroplast database containing 4356 assemblies from 3899 species is established. Next, an in silico pipeline is applied to identify the DSS candidates of each species. For each species, all possible *k*-mers from one of target species chloroplast assemblies were generated at a fixed *k*-mer length. Then, the nonredundant *k*-mers were blasted against other assemblies in the chloroplast database. The *k*-mers mapped to all assemblies of target species were recorded as Set1. The *k*-mers unmapped to background species (i.e., the other 3898 species in the present study) were recorded as Set2. The *k*-mers appearing in both Set1 and Set2 are considered as DSS candidates. After identifying DSS candidates of all 3899 species, a DSS candidate database is created using all achieved DSS candidates. Then, polymerase chain reaction (PCR)-sanger sequencing and HTS are used to validate the DSSs of four species and 165 species, respectively, using 40 bp DSS candidates in the constructed database.

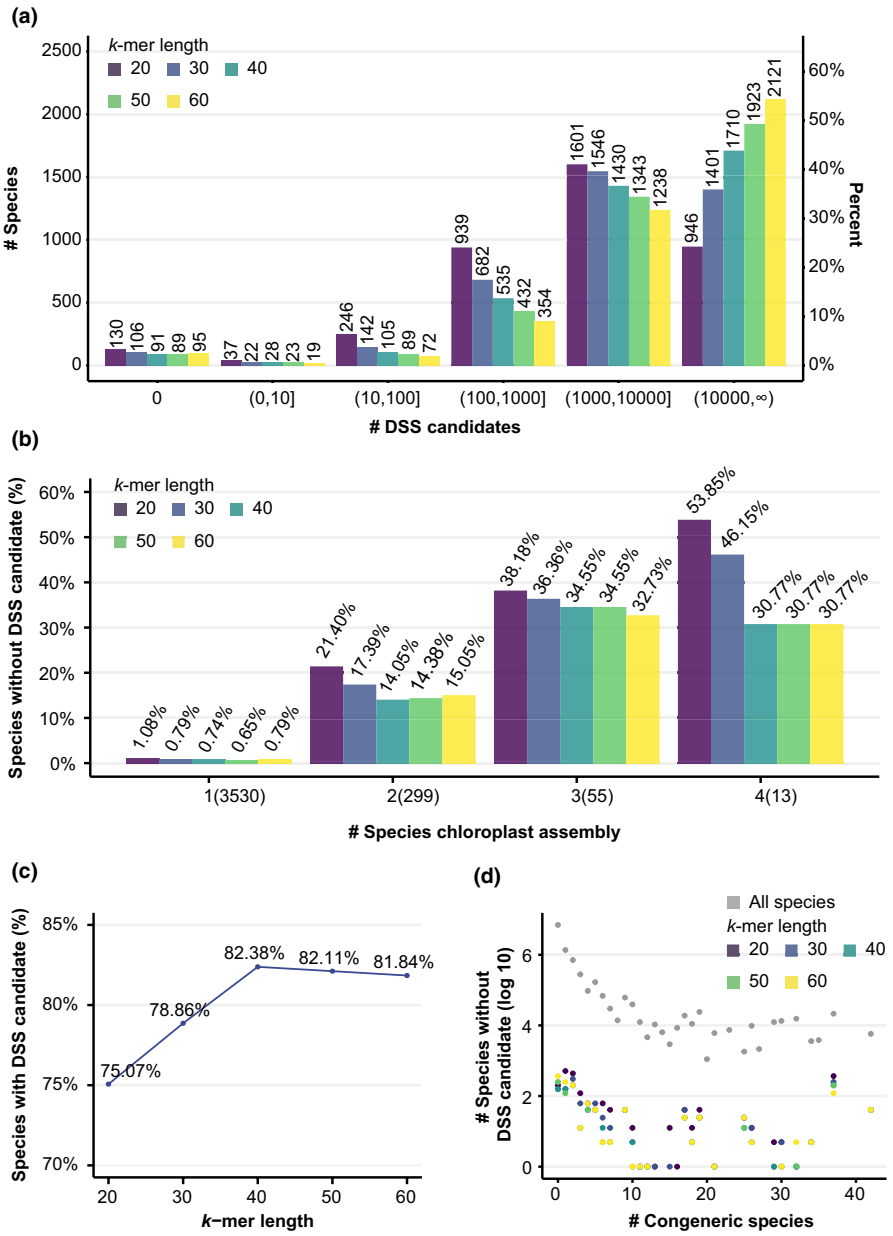

FIGURE 2 The DNA signature sequence (DSS) candidates in angiosperms. (a) Number of species by considering different number of DSS candidates. (b) Proportion of species without DSS candidates by considering different number of species chloroplast assemblies. Numbers in parentheses indicate the number of species with corresponding number of assemblies in 3899 species. Two species possessing five assemblies not shown. (c) Proportion of species with DSS candidate under different *k*-mer lengths in 369 species with multiple assemblies. (d) Number of species without DSS candidate by considering different number of congeneric species.

distinguished from *each other* (Figure 3a), indicating that these DSS candidates were DSSs.

3.2.2 | Case study 2

Five *A. lancea* DSS candidates and five *A. macrocephala* DSS candidates were randomly selected from the 646 and 1327 DSS candidates identified in this study (from 30 and 53 combined DSS candidates, respectively). The results (Figure 3b) showed that two of the five *A. macrocephala* DSS candidates could distinguish *A. macrocephala* from the other two species. In contrast to *A. macrocephala*, none of the five *A. lancea* DSS candidates could distinguish *A. lancea* from *A. japonica*. These results indicate that two *A. macrocephala* DSS candidates, but none of *the A. lancea* DSS candidates, were DSSs.

We hypothesize that none of the selected *A. lancea* DSS candidates could separate *A. lancea* from *A. japonica* due to a bias induced by the lack of background species because *A. japonica* chloroplasts were not included in our 3899-species data set. To validate this hypothesis, one sample of *A. japonica* was collected and sequenced, and its chloroplast sequence (accession no. GWHAZVW01000000) was added to the background data set, based on which *A. lancea* DSS candidates were reidentified. Only one DSS candidate was identified (Figure 3c) and further validated as a DSS.

3.3 | HTS-based DSS validation

Large-scale validation of DSS candidates was performed using WGS data generated from HTS. According to the principles

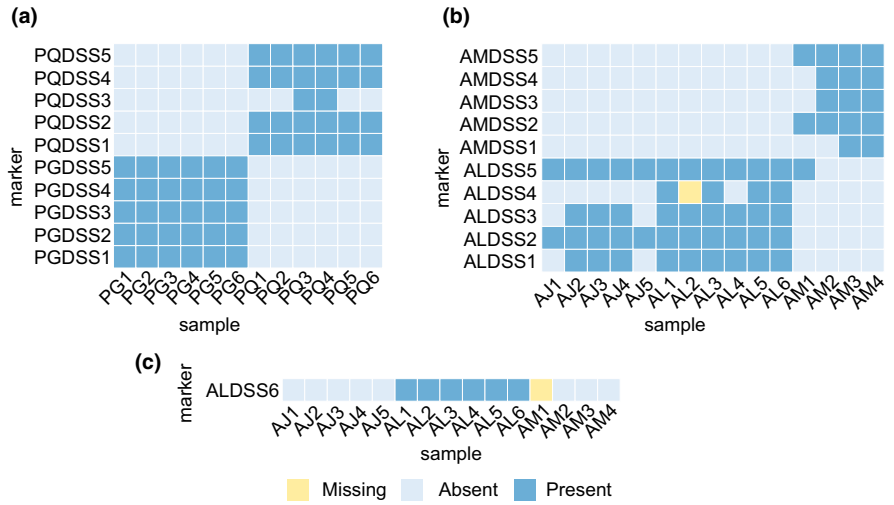

FIGURE 3 The validation of DNA signature sequences (DSSs) using polymerase chain reaction (PCR) and sanger sequencing. (a) the PAV of *P. ginseng* and *P. quinquefolius* DSS candidates in *panax* samples. PG1–PG6 are six *P. ginseng* samples, PQ1–PQ6 are six *P. quinquefolius* samples. PGDSS1–PGDSS5 are five *P. ginseng* DSS candidates, PQDSS1–PQDSS5 are five *P. quinquefolius* DSS candidates. (b) the PAV of *A. lancea* and *A. macrocephala* DSS candidates in *Atractylodes* samples. AJ1–AJ4, AL1–AL5, and AM1–AM5 are four *A. japonica*, five *A. lancea*, and five *A. macrocephala* samples, respectively. ALDSS1–ALDSS5 are five *A. lancea* DSS candidates, AMDSS1–AMDSS5 are five *A. macrocephala* DSS candidates. (c) the PAV of the *A. lancea* DSS candidates after adding *A. japonica* into background species. Missing, the failure of PCR amplification; absent, the DNA sequence in the amplification product was not identical to the DSS candidate; present, the DNA sequence in the amplification product was identical to the DSS candidate.

mentioned above, we set the *k*-mer length at 40 bp to identify DSS candidates, and a DSS candidate was regarded as such if it could be detected exclusively in the conspecific data set but not in the background data set. In total, WGS was performed on 165 species and their DSSs were validated using the HTS method (Tables S3 and S8).

Because the library size influences the detection of *k*-mer in HTS, we compared the proportion of DSS validated from DSS candidates, that is, the precision of DSS candidates, under five different library sizes to obtain a suitable library size. The higher the precision of the DSS candidates, the more reliable they are. Except for six species that do not exhibit any DSS at any sequencing depth, the highest precision of DSS candidates was achieved in 132 of the 159 species when the library size was 1 Gb or 100 Mb (Figure 4a, Table S10).

We then investigated the influence of three factors on the precision of DSS candidates using their precision under 10^9 sequenced nucleotides. First, regarding the number of assemblies, we compared the precision of DSS candidates identified from all assemblies (Table S10) with those identified from individual assemblies (Table S11) using 94 species with multiple assemblies in 165 species. The results demonstrated that the precision of the DSS candidates based on all assemblies was relatively high (Figure 4b). Second, we found that the precision of DSS candidates and congeneric species numbers was weakly correlated ($r = 0.262$, $p < .01$; Spearman's test, Figure 4c). For the third factor, the influence of the number of DSSs was assessed. Biologically, DSSs from the same combined DSS are the same markers. Thus, the precision of DSS candidates in the top and last 20% of species ranked by the number of combined DSS candidates, rather than the number of DSSs, were compared. The precision of the DSS candidates in the last 20% was not lower than that in the top 20% (Figure 4d). Our results suggest that for a given species, the number of DSS candidates and closely related species has no effect on the precision of DSS candidates, but the presence of multiple assemblies did. To increase the precision of the DSS candidates, we suggest using multiple assemblies to identify DSS candidates for each species.

Moreover, benefitting from the DSSs of 165 species validated by the HTS method, we investigated another question: How many DSS candidates need to be tested to achieve a DSS? Is testing five DSS candidates sufficient for most species to achieve at least one DSS, such as *P. ginseng*? This question is critical if DSSs must be validated using a low-throughput PCR-based method. To this end, we evaluated the probability of at least one DSS being present under different sampling numbers of DSS candidates. Because the precision of DSS candidates is related to the chloroplast assembly number, we separately evaluated the sampling number of DSS candidates in 71 species with one assembly and 94 species with multiple assemblies to eliminate bias. In addition, to remove sampling bias, we sampled different numbers of species (ranging from 10 to 40), and similar results proved that there was no species bias (Figure 4e, Figure S2). For 94 species with multiple assemblies, more than 90% of the species achieved at least one DSS when the number of DSS candidates sampled was over 40 (Figure 4e, Table S12), whereas for species with one assembly, more than 70 sampling DSS candidates were needed to achieve the same effect (Figure 4e, Table S12).

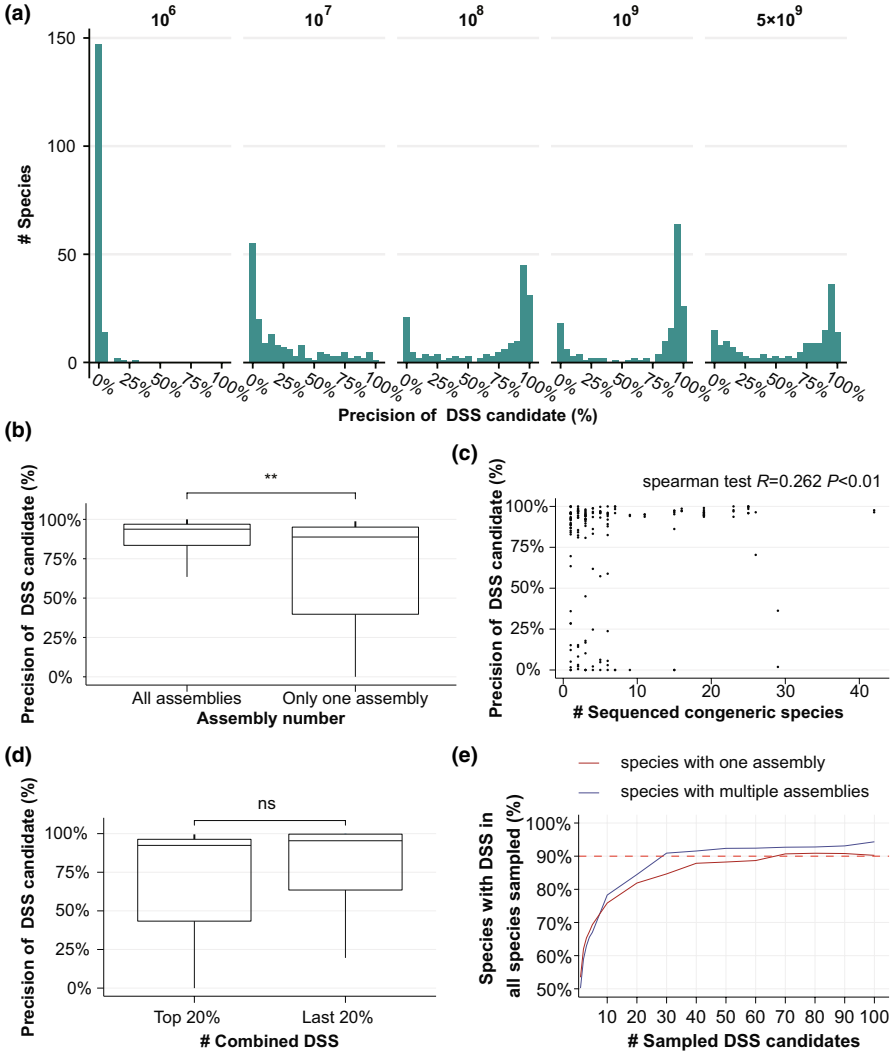

FIGURE 4 The validation of DSSs using HTS and factors affecting the precision of DSS candidates. (a) Histogram of the number of species over different precision of DSS candidates. The DSSs were validated on data sets of five different library size (10^6, 10^7, 10^8, 10^9, and 5×10^9 bp sequenced nucleotides). (b) the precision of DSS candidates using different number of assemblies. (c) Correlation between the precision of DSS candidates and the number of sequenced congeneric species. (d) Precision of DSS candidates in species with different numbers of DSSs. (e) the probability that one DSS is present under different sampling number of DSS candidates when the number of sampling species is 40. $^{**}p < .01$; ns, not significant.

3.4 | Comparison of DSS-based methods and DNA barcoding

To compare the discriminatory power and universality of DSSs with DNA barcoding, we collected 29 DNA barcoding regions from the chloroplast data set (see Methods), resulting in 103,307 barcode sequences (Table 1). Consistent with previous studies (CBOL Plant Working Group, 2009; Li et al., 2011), the discriminatory power of most chloroplast barcodes at the species level was approximately 66%, with a descending order with the top five barcodes being *atpF–atpH* (66.67%), *rps16* (66.66%), *trnL–trnF* (66.50%), *matK* (62.50%), and *trnH–psbA* (60.00%) (Table 1). It is not surprising that DSS, a species-specific marker, has a lower universality than most DNA barcode markers (Figure 5) (the universality of four primary barcodes were: *trnL–trnF* [98.23%], *trnH–psbA* [98.08%], *rbcL* [85.66%], *matK* [87.46%]).

Furthermore, DNA barcodes for multiple regions of many species have been reported for the first time. For example, 3752 *rbcL* sequences were obtained from 3340 species in our study, of which 1093 had no prior *rbcL* records in GenBank. This was also observed in regions of less concern such as *atpI-atpH*, *ndhJ-trnL*, and *petL-psbE*. These results also indicated that our identified DNA barcodes from the existing plastid data set could be a supplementary source of DNA barcodes.

4 | DISCUSSION

4.1 | The best practice for identifying DSS

We split the process of identifying DSSs into two: identifying DSS candidates and validating DSSs from these candidates. First, for identifying DSS candidates, we noticed that DSS candidates of species with multiple assemblies had higher precision (Figure 4b). We speculate that this is because some false-positive interspecies variations can be eliminated by intraspecies variations reflected in multiple assemblies. This also explains why species with multiple assemblies are more likely to possess no DSS candidates than those with only one assembly (Figure 2b). Furthermore, there should be an optimal *k*-mer length for identifying DSS candidates using multiple assemblies because both intraspecies and interspecies variations increase with *k*-mer length simultaneously.

TABLE 1 The discriminatory power and universality of DNA barcodes

Barcode	Number of sequences	Number of species (without prior records)	Family discrimination success (95% confidence interval)	Genus discrimination success (95% confidence interval)	Species discrimination success (95% confidence interval)
accD	3679	3254 (2962)	47.96%–48.00%	31.00%–31.31%	19.00%–19.19%
atpB	1317	1186 (950)	87.88%–88.00%	81.82%–82.00%	63.00%–63.64%
atpF-atpH	4327	3825 (3557)	99.00%–100.00%	93.94%–93.94%	66.67%–66.67%
atpI-atpH	4412	3797 (3769)	-	-	-
matK	3408	3022 (908)	96.00%–96.97%	91.00%–91.00%	62.00%–63.00%
ndhF-rpl32	1302	1160 (1133)	-	-	-
ndhJ	2388	2075 (1997)	-	-	-
ndhJ-trnL	4193	3741 (3740)	-	-	-
petL-psbE	4298	3832 (3820)	-	-	-
psaI-accD	3526	3121 (3092)	-	-	-
psbB-psbH	2367	2091 (2025)	-	-	-
psbD-trnT	3645	3211 (3192)	-	-	-
psbJ-petA	3937	3546 (3496)	-	-	-
psbK-psbI	4199	3743 (3671)	-	-	-
rbcL	3752	3340 (1093)	95.00%–95.88%	84.85%–85.00%	56.00%–56.57%
rpl14-rpl36	4142	3664 (3648)	-	-	-
rpl32-trnL	3765	3341 (3274)	-	-	-
rpoB	4210	3765 (3271)	67.78%–68.54%	57.83%–58.33%	34.12%–34.88%
rpoC1	4318	3847 (3078)	88.78%–89.47%	77.55%–78.00%	45.83%–46.39%
rps12-rpl20	2464	2217 (2197)	-	-	-
rps16	3749	3364 (2374)	96.97%–96.97%	90.72%–90.91%	65.31%–66.00%
rps16-trnK	3772	3321 (3293)	-	-	-
trnC-trnD	4110	3670 (3652)	-	-	-
trnD-trnT	3865	3468 (3446)	-	-	-
trnH-psbA	4314	3824 (2787)	98.00%–98.99%	92.00%–93.00%	60.00%–60.00%
trnL-trnF	4290	3830 (2693)	100.00%–100.00%	93.00%–93.00%	66.00%–67.00%
trnV-ndhC	4266	3671 (3647)	-	-	-
ycf1	2486	2068 (1851)	97.40%–97.44%	90.54%–90.79%	60.76%–61.33%

According to subsequent analysis, the optimal length was 40 bp (Figure 2c).

Although multiple assemblies can improve the precision of DSS candidates, there may still be false-positive DSS candidates due to sequencing errors, random mutations, etc.; thus, validating DSS is necessary. A potential barrier preventing researchers from using PCR-based methods is the possibility that too many candidates must be tested to achieve a DSS. Our results showed that 90% of the species acquired DSS after testing 40 DSS candidates (Figure 4e, Table S12). In addition, the number of DSS candidates is not linearly correlated with the probability of DSS presence, implying that most species do not have to go through a test of 40 DSS candidates to acquire a DSS. For species with multiple assemblies, five DSS candidates achieved a 67% probability of finding at least one DSS and 10 DSS candidates, corresponding to 78% (Figure 4e).

The HTS method can rapidly screen the DSSs from DSS candidates. *Based on the HTS-based validation of 165 species, we found that* the optimal sequencing amount for DSS validation was 1 Gb. One possible explanation for the 1 Gb overwhelming the 5 Gb data set is that the probability of reads containing DSS sequences generated from the nuclear genome rather than the chloroplast genome increases with sequencing depth. Therefore, 100 Mb–1 Gb sequencing data should be optimal for validation of DSS candidates in most plants based on the trade-off between reducing bias from the background species nuclear genome and increasing the likelihood of detecting DSS candidates from target species.

Apart from the factors discussed above, background species are another critical, yet challenging issue. In this case study, we obtained 30 combined DSS candidates from a background species data set that did not include *A. lancea*. However, only one of these has been proven to be a DSS marker that can distinguish *A. lancea* from *A. japonica*. In contrast, *A. macrocephala* DSS candidates identified using the same data set did not exhibit this deficiency. Although it is difficult to estimate how frequently this problem is likely to occur,

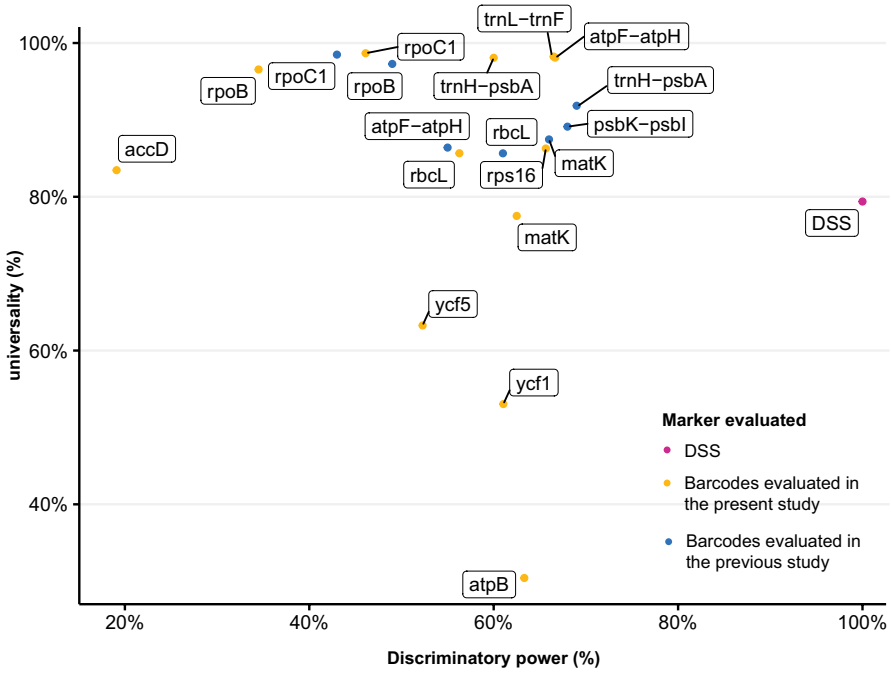

FIGURE 5 The universality and discriminatory power of different markers. The data of barcodes evaluated in a previous study was obtained from CBOL working group (2009).

increasing the background species may address this issue to some extent because the number of DSS candidates remains stable when the number of background species is over 1000. Nonetheless, it is preferable for researchers to sequence the chloroplast genome of closely related species that need to be distinguished from the target species but are not available in public databases. In the case of genus *Atractylodes*, we speculate that the capability of DSS to distinguish species not included in the background data set may be affected by affinity because *A. japonica* is more closely related to *A. lancea* than *A. macrocephala* (Wang et al., 2021).

Based on our results, we propose five principles for identifying DSS candidates: (1) species to identify DSS should have at least two chloroplast assemblies, and the more assemblies, the better the results; (2) the optimal *k*-mer length for DSS is 40 bp; (3) closely related species that need to be distinguished should be included in background species; (4) when using the HTS method to validate DSS, the optimal sequencing amount is 1 Gb; and (5) when using low-throughput methods to validate DSS, testing 40 DSS candidates should be sufficient.

4.2 | Do all species possess DSSs?

To evaluate a molecular marker, two touchstones used in DNA barcoding can be consulted (CBOL Plant Working Group, 2009; Kress & Erickson, 2008): (1) Universality: which loci can be routinely amplified and sequenced across land plants? (2) Discrimination: Which loci enable most species to be distinguished? DSSs are constant-length nucleotide sequences that can detect the presence of a taxon and distinguish it from background species; thus, the discriminatory power of DSS was 100%. Therefore, the most critical question for DSS application is the universality of DSS (i.e., do all species possess DSSs?). Because we split the process of identifying DSSs into two processes, the universality of DSS can be calculated by the following formula:

$$\text{The universality of DSS} = \frac{\text{Species with DSS candidates}}{\text{All species}} \times \frac{\text{Species with DSS}}{\text{Species with DSS candidates}} \times 100\%$$

To identify DSS candidates for a specific target species, a key issue is defining what the "background species" in the DSS definition is. Theoretically, in terms of plant identification, the background species comprise all other plant species, except for the target species. However, in practical applications, it is neither possible nor necessary to include all other plants because only a few of them have been sequenced. In this study, we use other 3898 species as background species when the DSS of each species was identified. In compliance with the principles proposed above, we used the results of species with multiple assemblies and HTS-based validation results using 1 Gb sequencing data to evaluate the universality of DSS. As a result, 304 of 369 species (82.38%) with multiple assemblies were DSS candidates, and the HTS-based validation showed that DSSs could be found in 159 of 165 species (96.36%) with DSS candidates. Therefore, the universality of the DSSs was 79.38% (82.38%×96.36%). A potential concern may be that for a specific species, the number of DSSs may decrease with an increase in background species, especially for closely related species. Our results demonstrated that when the number of background species was over 1000, the number of DSS candidates did not decrease significantly with increasing numbers of background species. Our results also indicated that the number of closely related species did not influence the occurrence of DSS candidates (Figure 2d) or the precision of DSS candidates (Figure 4c).

4.3 | Future prospects for DSS

Efficient plant species identification methods are the basis of plant biodiversity conservation; however, there is still a long way to go. DSS markers can be used as powerful supplements to address the issue that DNA barcoding does not perform well in certain taxa to some extent. *Gentiana* spp. are derived from multiple lineages and undergo radiative speciation (Zhang et al., 2009). A previous study reported that the discriminatory power of single-locus DNA barcodes ranged from 60% to 74.42% in *Gentiana* and the discriminatory power of multi-locus barcodes ranged from 71.43% to 88.24% (Liu et al., 2016). In our study, the DSSs of four *Gentiana* spp. were validated, showing high precision of DSS candidates at 99.8%, 93.6%, 96.3%, and 94.3% for *G. rigescens*, *G. scabra*, *G. trifloral*, and *G. manshurica*, respectively (Table S2).

Another valuable aspect of DSS is that it is PCR-free. The detection of plant components unidentifiable by morphology from mixtures (e.g., food, herbal products, and environmental samples) can portray the biodiversity of environmental samples and improve the oversight of trade in endangered species (Manzanilla et al., 2022; Taberlet et al., 2012). Metabarcoding is a powerful tool for detecting plant components in mixtures (de Boer et al., 2017), but it is hampered by low amplification efficiency because DNA in complex mixtures always degrades due to heating, pH changes, and other factors (Lo & Shaw, 2018; Paula et al., 2015; Sakaridis et al., 2013). In contrast, DSS is PCR-free when combined with HTS because it is simple to obtain reads longer than 100 bp using HTS. This strength makes DSS a promising tool for detecting components of plant origin in complex mixtures.

AUTHOR CONTRIBUTIONS

Zhang Zhang, Luqi Huang and Yuan Yuan supervised and coordinated the study. Zhongyi Hua, Chao Jiang, and Yuan Yuan Designed and conceived the study. Zhongyi Hua performed wet-laboratory work. Ziyuan Chen, Chao Jiang, Yan Jin, Yuyang Zhao, and Junhui Zhou provided samples, material and reagents. Shuhui Song, Dongmei Tian and Zhongyi Hua analysed data. Zhongyi Hua and Zhang Zhang interpreted the data. Zhongyi Hua, Chao Jiang, Shuhui Song, and Zhang Zhang wrote the article with input from all coauthors.

ACKNOWLEDGEMENTS

This work was supported by Special Funds for Basic Resources Investigation Research of the Ministry of Science and Technology (grant no. 2018FY10080002), Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (grant no. CI2021B014/CI2021A041), Key project at central government level for the ability establishment of sustainable use for valuable Chinese medicine resources (grant no. 2060302), The National Key Research and Development Program of China (grant no. 2019YFC1711000).

CONFILCT OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

IdenDSS has been made available under the MIT Licence on GitHub at https://github.com/Hua-CM/IdenDSS. The raw sequence data generated in this study have been deposited in the GSA in NGDC, China National Centre for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number CRA004065 that are publicly accessible at https://ngdc.cncb.ac.cn/gsa. Other data needed to reproduce the results of this study are all presented in the Supporting Information Data.

ORCID

Zhongyi Hua https://orcid.org/0000-0002-6659-9824

REFERENCES

Baeg, I.-H., & So, S.-H. (2013). The world ginseng market and the ginseng (Korea). *Journal of Ginseng Research*, *37*(1), 1–7.

Bánki, O., Roskov, Y., Vandepitte, L., DeWalt, R. E., Remsen, D., Schalk, P., Orrell, T., Keping, M., Miller, J., Aalbu, R., Adlard, R., Adriaenssens, E., Aedo, C., Aescht, E., Akkari, N., Alonso-Zarazaga, M. A., Alvarez, B., Alvarez, F., Anderson, G., … von Konrat, M. (2021). *Catalogue of life checklist (annual checklist 2021)*. Catalogue of Life. https://doi.org/10.48580/d4sb

CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, *106*(31), 12794–12797.

Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., Mabberley, D. J., Sennikov, A. N., Soltis, P. S., & Stevens, P. F. (2016). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, *181*(1), 1–20.

Chen, M., Ma, Y., Wu, S., Zheng, X., Kang, H., Sang, J., Xu, X., Hao, L., Li, Z., Gong, Z., Xiao, J., Zhang, Z., Zhao, W., & Bao, Y. (2021). Genome warehouse: A public repository housing genome-scale data. *Genomics, Proteomics & Bioinformatics.*, *19*, 584–589.

Christenhusz, M. J. M., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, *261*(3), 201–217.

CNCB-NGDC Members and Partners. (2022). Database resources of the National Genomics Data Center, China National Center for bioinformation in 2022. *Nucleic Acids Research*, *50*(D1), D27–D38.

de Boer, H. J., Ghorbani, A., Manzanilla, V., Raclariu, A. C., Kreziou, A., Ounjai, S., Osathanunkul, M., & Gravendeel, B. (2017). DNA metabarcoding of orchid-derived products reveals widespread illegal orchid trade. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1863), 20171182.

Ebert, A. W., & Engels, J. M. M. (2020). Plant biodiversity and genetic resources matter! *Plants*, *9*(12), 1706.

Goetze, M., Zanella, C. M., Palma-Silva, C., Büttow, M. V., & Bered, F. (2017). Incomplete lineage sorting and hybridization in the evolutionary history of closely related, endemic yellow-flowered species of subgenus (Bromeliaceae). *American Journal of Botany*, *104*(7), 1073–1087.

Guo, C., Ma, P. F., Yang, G. Q., Ye, X. Y., Guo, Y., Liu, J. X., Liu, Y. L., Eaton, D. A. R., Guo, Z. H., & Li, D. Z. (2020). Parallel ddRAD and genome skimming analyses reveal a radiative and reticulate evolutionary history of the temperate bamboos. *Systematic Biology*, *70*(4), 756–773.

Hebert, P. D., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321.

Hollingsworth, P. M., Li, D. Z., van der Bank, M, & Twyford, A. D. (2016). Telling plant species apart with DNA: From barcodes to genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150338.

Ji, Y., Liu, C., Yang, Z., Yang, L., He, Z., Wang, H., Yang, J., & Yi, T. (2019). Testing and using complete plastomes and ribosomal DNA sequences as the next generation DNA barcodes in *panax* (Araliaceae). *Molecular Ecology Resources*, *19*(5), 1333–1345.

Kaplinski, L., Lepamets, M., & Remm, M. (2015). GenomeTester4: A toolkit for performing basic set operations-union, intersection and complement on k-mer lists. *Gigascience*, *4*(1), s13742-015-0097-y.

Kress, W. J., & Erickson, D. L. (2008). DNA barcodes: Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences*, *105*(8), 2761–2762. https://doi.org/10.1073/pnas.0800476105

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., Yang, J. B., Fu, C. X., Zeng, C. X., Yan, H. F., Zhu, Y. J., Sun, Y. S., Chen, S. Y., Zhao, L., Wang, K., … Duan, G. W. (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*(49), 19641–19646.

Liu, J., Yan, H.-F., & Ge, X.-J. (2016). The use of DNA barcoding on recently diverged species in the genus Gentiana (Gentianaceae) in China. *PLoS One*, *11*(4), e0153008.

Lo, Y. T., & Shaw, P. C. (2018). DNA-based techniques for authentication of processed food and food supplements. *Food Chemistry*, *240*, 767–774.

Manzanilla, V., Teixidor-Toneu, I., Martin, G. J., Hollingsworth, P. M., de Boer, H. J., & Kool, A. (2022). Using target capture to address conservation challenges: Population-level tracking of a globally-traded herbal medicine. *Molecular Ecology Resources*, *22*(1), 212–224.

Martin, A. C., & Harvey, W. J. (2017). The global pollen project: A new tool for pollen identification and the dissemination of physical reference collections. *Methods in Ecology and Evolution*, *8*(7), 892–897.

Myers, G. (2020). FastK: A K-mer counter (for HQ assembly data sets). Github. https://github.com/thegenemyers/FASTK

Nguyen, V. B., Park, H. S., Lee, S. C., Lee, J., Park, J. Y., & Yang, T. J. (2017). Authentication markers for five major panax species developed via comparative analysis of complete chloroplast genome sequences. *Journal of Agricultural and Food Chemistry*, *65*(30), 6298–6306.

Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, *7*(1), 16.

Paula, D. P., Linard, B., Andow, D. A., Sujii, E. R., Pires, C. S. S., & Vogler, A. P. (2015). Detection and decay rates of prey and prey symbionts in the gut of a predator through metagenomics. *Molecular Ecology Resources*, *15*(4), 880–892.

Piñol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology*, *28*(2), 407–419.

Raime, K., Krjutskov, K., & Remm, M. (2020). Method for the identification of plant DNA in food using alignment-free analysis of sequencing reads: A case study on lupin. *Frontiers in Plant Science*, *11*, 646.

Raime, K., & Remm, M. (2018). Method for the identification of taxon-specific k-mers from chloroplast genome: A case study on tomato plant (Solanum lycopersicum). *Frontiers in Plant Science*, *9*, 6.

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364.

Sakaridis, I., Ganopoulos, I., Argiriou, A., & Tsaftaris, A. (2013). A fast and accurate method for controlling the correct labeling of products containing buffalo meat using high resolution melting (HRM) analysis. *Meat Science*, *94*(1), 84–88.

Senizza, B., Rocchetti, G., Ghisoni, S., Busconi, M., De Los Mozos Pascual, M., Fernandez, J. A., Lucini, L., & Trevisan, M. (2019). Identification of phenolic markers for saffron authenticity and origin: An untargeted metabolomics approach. *Food Research International (Ottawa, Ont.)*, *126*, 108584.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050.

Tu, Q., He, Z., Deng, Y., & Zhou, J. (2013). Strain/species-specific probe design for microbial identification microarrays. *Applied and Environmental Microbiology*, *79*(16), 5085–5088.

Tu, Q., He, Z., & Zhou, J. (2014). Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Research*, *42*(8), e67.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3–new capabilities and interfaces. *Nucleic Acids Research*, *40*(15), e115.

Wäldchen, J., Rzanny, M., Seeland, M., & Mäder, P. (2018). Automated plant species identification-trends and future directions. *PLoS Computational Biology*, *14*(4), e1005993.

Wang, Y., Song, F., Zhu, J., Zhang, S., Yang, Y., Chen, T., Tang, B., Dong, L., Ding, N., Zhang, Q., Bai, Z., Dong, X., Chen, H., Sun, M., Zhai, S., Sun, Y., Yu, L., Lan, L., Xiao, J., … Zhao, W. (2017). GSA: Genome sequence archive. *Genomics, Proteomics & Bioinformatics*, *15*(1), 14–18.

Wang, Y., Wang, S., Liu, Y., Yuan, Q., Sun, J., & Guo, L. (2021). Chloroplast genome variation and phylogenetic relationships of *Atractylodes* species. *BMC Genomics*, *22*(1), 103.

Zhang, N., Ma, Y., Folk, R. A., Yu, J., Pan, Y., & Gong, X. (2018). Maintenance of species boundaries in three sympatric Ligularia (Senecioneae, Asteraceae) species. *Journal of Integrative Plant Biology*, *60*(10), 986–999.

Zhang, X.-L., Wang, Y.-J., Ge, X.-J., Yuan, Y.-M., Yang, H.-L., & Liu, J.-Q. (2009). Molecular phylogeny and biogeography of Gentiana sect. Cruciata (Gentianaceae) based on four chloroplast DNA datasets. *Taxon*, *58*(3), 862–870.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hua, Z., Jiang, C., Song, S., Tian, D., Chen, Z., Jin, Y., Zhao, Y., Zhou, J., Zhang, Z., Huang, L., & Yuan, Y. (2022). Accurate identification of taxon-specific molecular markers in plants based on DNA signature sequence. *Molecular Ecology Resources*, *00*, 1–12. https://doi.org/10.1111/1755-0998.13697