Forensic Science International: Genetics 36 (2018) 1–12

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

Identification of plant species using variable length chloroplast DNA sequences

Chiara Santos[a,b], Filipe Pereira[a,*]

[a] Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos s/n 4450-208, Matosinhos, Portugal

[b] Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

ARTICLE INFO

Keywords:
cpDNA
Plants
SPInDel
Species identification
Forensic botany

ABSTRACT

The correct identification of species in the highly divergent group of plants is crucial for several forensic investigations. Previous works had difficulties in the establishment of a rapid and robust method for the identification of plants. For instance, DNA barcoding requires the analysis of two or three different genomic regions to attain reasonable levels of discrimination. Therefore, new methods for the molecular identification of plants are clearly needed. Here we tested the utility of variable-length sequences in the chloroplast DNA (cpDNA) as a way to identify plant species. The SPInDel (Species Identification by Insertions/Deletions) approach targets hypervariable genomic regions that contain multiple insertions/deletions (indels) and length variability, which are found interspersed with highly conserved regions. The combination of fragment lengths defines a unique numeric profile for each species, allowing its identification. We analysed more than 44,000 sequences retrieved from public databases belonging to 206 different plant families. Four target regions were identified as suitable for the SPInDel concept: *atpF-atpH*, *psbA-trnH*, *trnL* CD and *trnL* GH. When considered alone, the discrimination power of each region was low, varying from 5.18% (*trnL* GH) to 42.54% (*trnL* CD). However, the discrimination power reached more than 90% when the length of some of these regions is combined. We also observed low diversity in intraspecific data sets for all target regions, suggesting they can be used for identification purposes. Our results demonstrate the utility of the SPInDel concept for the identification of plants.

1. Introduction

The correct identification of plant species is relevant in forensic investigations where traces of plants can be associated with crimes scenes, in food traceability and quality control, illegal logging and trade, investigations of poisoning with products derived from plants, among others [1–6]. Most molecular methods for species identification are still limited by the need for high amounts of quality DNA, the occurrence of non-specific DNA hybridization, the difficulty of interpreting electrophoretic profiles in mixtures and the high dependence on laboratory conditions [7–9]. Such problems limit the standardization of results for inter and intra-laboratory comparisons.

We have previously developed the SPInDel (Species Identification by Insertions/Deletions) method for molecular species identification [10,11]. Our method uses the size variation of hypervariable regions containing multiple insertion/deletion (indels) polymorphisms that are interspersed with conserved domains. Each species is identified by the combination of the lengths of the hypervariable regions (Fig. 1). The major advantages of the SPInDel method are: a) potential to work in all taxonomic groups; b) simultaneous analysis of multiple loci; c) adaptability to different genotyping platforms with a reduced cost per sample; d) possibility of identifying species without DNA sequencing; e) amenability to multiplexing; f) suitability for identification of species that co-exist in a sample (mixtures); g) possibility of inter-laboratory comparisons, providing a means to standardize methodologies and h) requirement of a conventional laboratory with minimum equipment [10–13].

Our previous works have targeted the mitochondrial DNA (mtDNA) of animals, taking advantage of its relatively high mutation rate [11–13]. However, the mtDNA of plants is not suitable for species identification procedures since it is usually slowly evolving, resulting in the absence of inter-specific variation, has high intra-molecular recombination and pseudogenes [14–16]. Therefore, researches have used the chloroplast DNA (cpDNA) for identification of plant species [17–19]. The analysis of cpDNA sequences have been widely used for species identification and phylogenetic analyses because: a) it has a relative high mutation rate; b) is present at high copy numbers per cell; c) there are thousands of sequences in public databases; d) it has a few

* Corresponding author.
E-mail addresses: chiaragabriele5@gmail.com (C. Santos), fpereirapt@gmail.com (F. Pereira).

https://doi.org/10.1016/j.fsigen.2018.05.009
Received 18 December 2017; Received in revised form 10 May 2018; Accepted 28 May 2018
Available online 29 May 2018
1872-4973/

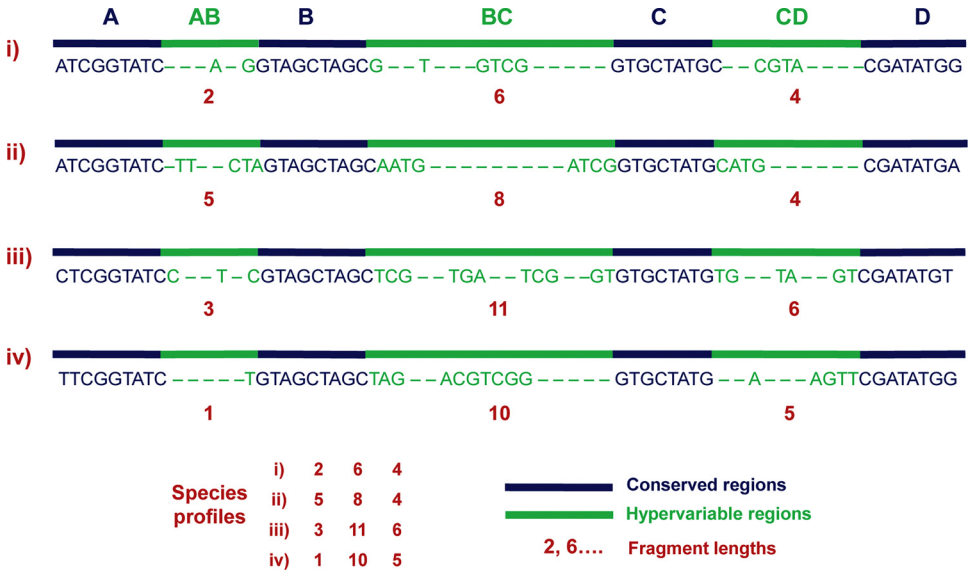

Fig. 1. Schematic illustration of the strategy used in the species identification by the insertions/deletions method (SPInDel). Illustration of the sequence alignment for four hypothetical species (i to iv). Four conserved regions (blue) define three hypervariable domains (green). Each species is identified by a numeric profile resulting from the combination of lengths in hypervariable regions (red numeric codes). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

highly conserved regions suitable for the design of 'universal' primers and e) it is usually uniparentally inherited, and non-recombinant, making it effectively haploid [20–22].

The cpDNA of plants is particularly suitable for the application of the SPInDel concept by having several coding regions (usually conserved) interspersed with large non-coding domains such as introns or intergenic spacers (usually rich in indels). Here we tested the use of the SPInDel concept for the identification of plants using data collected from public databases. Our results suggest that the identification of plants species can be obtained through analysis of DNA regions with variable lengths.

2. Materials and methods

2.1. Nucleotide sequences

We retrieved from the NCBI Entrez Nucleotide database (http://www.ncbi.nlm.nih.gov) all available cpDNA sequences from three different genomic regions suitable for the SPInDel concept (hypervariable regions interspersed with conserved domains): *atpF-atpH* (*ATPase I subunit – ATPase III subunit*), *psbA-trnH* (*PSII 32 kDa protein – tRNA-His* (GUG)) and *trnL* (*tRNA-Leu* (UAA)). We removed all redundant sequences belonging to the same species (duplicates) and those without a clear species assignment. Moreover, we also reverse complement some sequences that were found in the opposite direction. The DNA sequences of the three selected cpDNA regions were organized by family, according to the NCBI taxonomy (Table 1). The sequences in each family were aligned using the default parameters of the MUSCLE software [23] running in the Geneious version 5.5.8 [24]. The sequence alignments were repeated after excluding those sequences that do not cover the entire region of interest. We only used alignments with ten or more species per family for the SPInDel calculations. The multiple sequence alignments can be found in our public database named PlantAligDB (http://plantaligdb.portugene.com).

2.2. Selection of SPInDel conserved regions

We obtained a consensus sequence from each sequence alignment that represents the most frequent nucleotides in each position (i.e. family). The consensus sequences of each family were then aligned in order to allow the identification of SPInDel conserved regions, i.e., regions with no or small variability at the sequence level that can be used as primer-binding sites (Fig. 1). The SPInDel conserved regions were selected according to the criteria previously described [10,11]. In the case of *trnL* (UAA), we used as conserved regions those named "C", "D", "G" and "H" by Taberlet et al. [25]. The complete *trnL* (UAA) region defined by the regions C and D (*trnL* CD) and a shorter segment located inside CD defined by regions G and H (*trnL* GH) were analysed (Supplementary Fig. S1).

2.3. SPInDel analyses

The sequence alignments of each family for the four different cpDNA regions (*atpF-atpH*, *psbA-trnH*, *trnL* CD and *trnL* GH) were submitted to the SPInDel workbench [10] in order to perform diverse calculations. Supplementary Table S1 summarizes the SPInDel terminology. For the assessment of intra-species diversity, we selected four species for the *trnL* CD, *trnL* GH and *psbA-trnH* regions by considering those with the largest number of available sequences and representing different families (Supplementary Table S2). In the case of *atpF-atpH*, only two species with more than ten individuals were found. The sequences from each species were aligned as previously described. The alignments were analysed in the SPInDel workbench using the same conserved regions defined previously for the family of each species.

The SPInDel concept is based on the combination of sequence lengths from different genomic regions. Therefore, we concatenated the

Table 1
Number of sequences, families, SPInDel conserved and hypervariable regions retrieved from GenBank.

Region	Total number of sequences recovered from GenBank	Total number of filtered sequences[a]	Number of families	Number of families with N ≥ 10	Number of conserved regions	Number of hypervariable regions
atpF-atpH	2360	1317	156	29	2	1
psbA-trnH	14550	5632	327	79	2	1
trnL CD	4083	2714	117	44	4	3
trnL GH	54494	35198	351	173	2	1

[a] Filtered – one per species, with complete taxonomy and covering the region of interest.