U.PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Development of new tools for the identification of plants using chloroplast DNA sequences

Chiara Gabriele Santos

Tese de Doutoramento apresentada à Faculdade de Ciências da Universidade do Porto, Centro Interdisciplinar de Investigação Marinha e Ambiental

Biologia

2018

Development of new tools for the identification of plants using chloroplast DNA sequences

Chiara Gabriele Santos

Doutoramento em Biologia

Departamento de Biologia

2018

Orientador

Doutor Filipe Pereira, CIIMAR

Coorientador

Professor Doutor Vítor Vasconcelos, FCUP

NOTA INTRODUTÓRIA

Dissertação apresentada à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Doutor em Biologia, no âmbito do programa doutoral em Biologia.

Esta tese foi escrita ao abrigo do numero 2 do Artigo 4º do Regulamento Geral dos Terceiros Ciclos de Estudos da Universidade do Porto. É composta por um conjunto coerente de trabalhos de investigação, já publicados ou submetidos para publicações em revistas internacionais com arbitragem científica. A autora contribuiu para a execução técnica do trabalho, a interpretação dos resultados, discussão dos dados e a preparação do manuscrito dos artigos incluídos nesta dissertação.

A doutoranda desenvolveu o trabalho com o suporte de uma bolsa de estudos de doutoramento Pleno no Exterior (BEX 013186/2013-02) atribuída pelo Programa Ciências Sem Fronteiras – Capes Brasil. O Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR) forneceu o suporte logístico para o trabalho experimental.

Publicações

Artigos publicados

1. Design and evaluation of PCR primers for amplification of four chloroplast DNA regions in plants

Artigos submetidos

2. Identification of plant species using variable length chloroplast DNA sequences
3. PlantAligDB: A Database of Nucleotide Sequence Alignment for Plants

Aos meus avós

ACKNOWLEDGMENTS / AGRADECIMENTOS

Ao meu orientador Filipe Pereira. As pessoas que vivem a minha volta sabem da minha admiração e respeito por ele, costumo dizer que devo ter feito muito bem para alguém em alguma vida passada ou em algum momento da minha vida para o universo ter colocado o Filipe no meu caminho. Hoje penso que eu também fiz por tê-lo no meu percurso, eu quem lhe enviei email, que fui buscar saber sobre o trabalho dele e acima de tudo ter aceite o desafio. Que sem sombra de duvidas foi o mais acertado. Não houve uma sequer vez em que estivemos juntos e que ele não me tenha transmitido calma, tranquilidade, segurança, confiança, assertividade e paciência. O que posso dizer, admiro-o como pessoa apesar de conhecer pouco, mas como profissional foi quem mais me ensinou sobre dedicação, sobre gostar do que se faz, sobre fazer aquela magica acontecer, sobre reconforto, sobre a certeza de que tudo vai dar certo, mostrar alternativas, reconhecimento ao esforço, mas ao mesmo tempo sobre criticar e apontar os erros. É desse tipo de pessoas que acredito vir realmente as criticas construtivas, pessoas que já passaram pela mesma situação e que genuinamente a única coisa que querem é o teu bem e o teu sucesso. Obrigada por fazer parte dessa importante etapa da minha vida, obrigada pela paciência, suporte, auxilio e compreensão.

João Carneiro sempre impecável, dedicado e atencioso, obrigada pela parceria.

Joana Damas, Joana Gonçalves e professor António Amorim pela atenção e recetividade.

Micaela Vale, Raquel Castelo Branco, Emília Afonso e todos os membros do CIIMAR por me terem acolhido e facilitado a dinâmica do trabalho dentro do instituto.

As minhas amigas Juliana Garcia, Liliana Remuge, Mónica Antunes e Fátima António pelas inúmeras conversas. Cada uma cumpriu lindamente seu papel nessa jornada que talvez não estivesse sendo concluída sem o suporte e insights que cada uma me deu. Com certeza vocês são uma das preciosidades que Portugal me trouxe.

João Miranda, é aquela pessoa pela qual tenho um carinho especial, sem necessidade de explicações. Obrigada por me ouvir, por preocupar-se comigo, por me aconselhar, por acompanhar-me. Indiscutivelmente foste uma pessoa que me acrescentou, e em vários momentos tornou esse percurso mais colorido.

Pedro Rodrigues e David Gomes por serem bons ouvintes nas conversas em monologo.

Dilman Faraj por teres, mesmo que distante, me acalentado e tranquilizado com tua companhia e compreensão.

A minha psicóloga Júlia Machado e ao meu terapeuta André Dourado por me terem ajudado a despertar o potencial que tenho dentro de mim quando eu mesma não o encontrava.

Aos colegas Diogo Aguiar, Patrick Nunes, Paulo Peres e Gonçalo Correia pelas inúmeras palavras e gestos de compreensão e apoio.

As amigas Maria Amorim e Inês Ribeiro, pela companhia entre um gel e outro.

As minhas amigas Jéssica Galhotto, Valle Carmona e Fernanda de Bastian que mesmo distantes sempre me enviaram energias positivas.

A minha família por apesar de não perceber muito bem os meus motivos estiveram sempre ao meu lado sendo compreensivos e resilientes.

Ao "Programa Ciências sem Fronteiras" do Concelho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil pela bolsa de doutoramento BEX: 13186/13-2 e o suporte financeiro.

Finalmente agradeço às seguintes instituições: CIIMAR, I3S e ao Departamento de Biologia da Faculdade de Ciências da Universidade do Porto pelo apoio logístico prestado para a realização desta dissertação.

ABSTRACT

The high genetic diversity of plants is a challenge to those developing new molecular and bioinformatics tools for their characterization. The use of DNA-based methods has facilitated the identification of plants families and species. However, it is clear that efficient methods for the study of most plants are lacking. Although well established in other taxonomic groups (animals and fungi), the DNA barcode concept is not very effective for plants. In this thesis, we started by applying the SPecies Identification by Insertions/Deletions (SPInDel) approach for the identification of plant species. Our method is based on length variation caused by indels polymorphisms in nucleotide sequences. We analysed over 44,000 sequences from 206 plant families. The chloroplast DNA (cpDNA) of plants proved to be particularly suited to the application of our approach. The utility of the SPInDel concept was clear when combining the *atpF-atpH*, *psbA-trnH* and *trnL* (UAA) cpDNA regions. The discriminating power of the selected regions ranged from 5.18% (*trnL* GH) to 42.54% (*trnL* CD), whereas when combined, values greater than 90% were obtained. Low intraspecific diversity was also observed in our dataset, demonstrating the effectiveness of the SPInDel approach in discriminating plant species. In the second part of this thesis, we have developed a set of conserved primers that amplify four informative regions of cpDNA (*atpF-atpH*, *psbA-trnH*, *trnL* CD and *trnL* GH) in the main plant families (Asteraceae, Brassicaceae, Iridaceae, Orchidaceae, Poaceae, Rosaceae and Salicaceae). The correct amplification of the four regions in samples from seven major plant families demonstrated the usefulness of our primers, which were obtained through the alignment of more than 11,000 reference cpDNA sequences. Finally, we have built an online database called PlantAligDB (available at http://plantaligdb.portugene.com), including 514 alignments with more than 66,000 reference sequences, belonging to 223 different families for the main genomic regions used in species identification and phylogenetic studies (*atpF-atpH*, *psbA-trnH*, *trnL*, *rbcL*, *matK* and ITS). The PlantAligDB provides a large source of data that enables the development of molecular markers, to investigate inter and intraspecific variability of genomic regions, among other tools facilitating taxonomic and phylogenetic studies.

Keywords: plants, species identification, cpDNA, primers, database

RESUMO

A grande diversidade genética das plantas é desafiante para aqueles que desenvolvem novas ferramentas moleculares e de bioinformática para sua caracterização. O uso de métodos baseados em DNA facilitou a identificação de famílias e espécies de plantas. No entanto, é claro que faltam métodos eficientes para o estudo da maioria das plantas. Embora bem estabelecido em outros grupos taxonómicos (animais e fungos), o conceito de DNA barcoding não é muito eficaz para as plantas. Nesta tese, começamos pela aplicação da abordagem de SPecies Identification by Insertions/Deletions (SPInDel) para identificação de espécies de plantas. Nosso método é baseado na variação do comprimento causada por polimorfismos de indels nas sequências nucleotídicas. Analisamos mais de 44.000 sequências de 206 famílias de plantas. O DNA do cloroplasto (cpDNA) das espécies de plantas revelou-se particularmente adequado para a aplicação da nossa abordagem. A utilidade do conceito SPInDel foi eficiente ao combinar as regiões *atpF-atpH*, *psbA-trnH* e *trnL* (UAA) do cpDNA. O poder de discriminação das regiões selecionadas variou de 5,18% (*trnL* GH) a 42,54% (*trnL* CD), enquanto que quando combinados foram obtidos valores acima de 90%. Uma baixa diversidade intraespecífica também foi observada em nosso conjunto de dados, demonstrando a eficácia da abordagem SPInDel na discriminação das espécies de plantas de forma rápida e fácil. Na segunda parte desta tese, desenvolvemos um conjunto de primers conservados que amplificam quatro regiões informativas de cpDNA (*atpF-atpH*, *psbA-trnH*, *trnL* CD e *trnL* GH) nas principais famílias de plantas (Asteraceae, Brassicaceae, Iridaceae, Orchidaceae, Poaceae, Rosaceae e Salicaceae). A amplificação correta das quatro regiões em amostras de sete importantes famílias de plantas demonstrou a efetividade de nossos primers, que foram obtidos através do alinhamento de mais de 11.000 seqüências de cpDNA de referência. Finalmente, construímos um banco de dados online chamado PlantAligDB (disponível em http://plantaligdb.portugene.com), incluindo 514 alinhamentos com mais de 66.000 seqüências de referência, pertencentes a 223 famílias diferentes para as principais regiões genômicas usadas na identificação de espécies e estudos filogenéticos (*atpF-atpH*, *psbA-trnH*, *trnL*, *rbcL*, *matK* e ITS). A PlantAligDB fornece uma grande fonte de dados que permite o desenvolvimento de marcadores moleculares, investiga a variabilidade inter e intraspecífica das regiões genômicas, entre outras ferramentas que facilitam estudos taxonómicos e filogenéticos.

Palavras-chaves: plantas, identificação de espécies, cpDNA, primers, base de dados

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3'UTR (Three Prime Untranslated Region)

AT (Adenine Thymine)

atpF (ATP synthase subunit CFO I)

atpH (ATP synthase subunit CFO III)

atpI (ATP synthase protein I)

atpA (ATP synthase subunit alpha)

atpB (ATP synthase beta subunit)

atpE (ATP synthase epsilon subunit)

BLAST (Basic Local Alignment Search Tool)

bp (Base Pairs)

CG (Cytosine Guanine)

cpDNA (Chloroplast DNA)

CBOL (Consortion for the Barcode of Life)

CTAB (Cetyltrimethylammonium Bromide Detergent Buffer)

COI (Cytochrome Oxidase I)

ddNTP (Dideoxynucleotides)

GMO (Genetically Modified Organisms)

Indels (Insertions / Deletions)

IR (Inverted Region)

ITS (Internal Transcribed Spacer)

Kbp (Kilobase pairs)

LSC (Large Single Copy)

matK (Maturase K)

Mbp (Mega base pair)

mtDNA (mitochondrial DNA)

NAD (P) H (Nicotinamide Adenine Dinucleotide Phosphate-oxidase)

NCBI (National Center for Biotechnology Information)

ndhF (NADH dehydrogenase F)

NGS (Next-generation sequencing)

nuDNA (Nuclear Genome)

PCR (Polymerase Chain Reaction)

psbA (Photosystem II 32 kDa protein)

psbK (Photosystem II reaction center protein K) psbI (Photosystem II reaction center protein I)

rbcL (Rubisco large subunit)

rpoB (RNA Polymerase beta subunit)

rpoC1 (RNA Polymerase beta subunit-1)

rpoC2 (RNA Polymerase beta-prime chain)

rps16 (Ribosomal Protein S16)

rps7 (Ribosomal Protein S7)

rRNA (Ribosomal RNA)

RuBisCO (Ribulose-1,5-bisphosphate carboxylase/oxygenase)

SNP (Single Nucleotide Polymorphisms)

SPInDel (Species Identification by Insertion Deletion)

SSC (Small Single Copy)

SSR (Simple Sequence Repeat or Microsatellite)

tRNA (Transfer RNA)

trnF (tRNA Phenylalanine gene)

trnH (tRNA-His (GUG))

trnK (tRNA-Lys(UUU))

trnL (tRNA-Leu (UAA))

trnQ (tRNA glutamine-specific)

trnT (tRNA-Thr (UGU))

trnV (tRNA (Val) UAC)

ycf1 (hypothetical chloroplast open reading frame 1)

CHAPTER I

INTRODUCTION

INTRODUCTION

Plants

Plants are extremely variable and dynamic living beings that exhibit an incredible diversity of habits, morphology, anatomy, physiology, reproductive biology, among others. This diversity is a challenge for those who are interested in studying these organisms (Bennetzen 2000, Jansen, Cai et al. 2007, Barolo, Mostacero et al. 2014). Humans obtain food and beverages through the cultivation of plant species (agriculture) (Bommarco, Kleijn et al. 2013, Cassidy, West et al. 2013, Staats, Arulandhu et al. 2016). It is widely recognized that plants have several health benefits. For example, the pharmaceutical industry uses plants to produce several products for therapeutic purposes on a daily basis. The bioactive compounds of the plants are available in the form of tea, medicines, syrups, ointments, oils, sprays, and many others (De Castro, Comparone et al. 2017). Similarly, the cosmetics industry takes advantage of the healing and embellishment properties of plants. Plants are the main source and foundation of all cosmetics like perfumes, shampoos, lotions, creams among others (Aburjai and Natsheh 2003, Mishra, Kumar et al. 2016). Plants are used by architects and landscapers in decoration and ornamentation, and have been used as shelter and protection (Barolo, Mostacero et al. 2014, Lee, Ng et al. 2016). Plants are the primary source of production of living material, and the loss of diversity affects all ecosystems. By this reason, plants are intensively studied by biologists and ecologists working on biodiversity loss, wildlife protection, threat of extinction, targeting traffic and illegal trade (Loreau, Naeem et al. 2001, Teletchea, Maudet et al. 2005, Díaz, Fargione et al. 2006, Karp and Shield 2008, Parker and Helmstetter 2017).

Plant materials can be used in forensic investigations, since it can be associated with the victim, the suspect or the crime scene. Proper identification of plant samples in an archaeological dig helps to understand aspects of ancient human life styles and reconstruct past environments (Coyle, Lee et al. 2005, Kikkawa, Tsuge et al. 2016). Crops plant species are known for be a biological sources to power generation (bioenergy) and liquid transport fuels (Karp and Shield 2008, Yuan, Tiller et al. 2008, Feuillet, Leach et al. 2011). In summary, plants can have various applications and should be studied and protected worldwide.

Identification of plant species

Despite the wide applicability of plants, it is often difficult to attribute the correct species belonging to a particular individual. This difficulty is due to the lack of a universally accepted species concept, the different criteria and methods used in plant classifications and to the high diversity of existing plant species (Joly, Goëau et al. 2016, Parker and Helmstetter 2017). Food, archaeological, herbal or forensic samples are often damaged, fragmented and/or transformed, thus preventing proper identification through morphological characteristics (Kikkawa, Tsuge et al. 2016). The traditional taxonomy system for species identification is limited by this reason. Visual search using different morphological characteristics has lower performance, efficacy and normalization of processes. Using morphology for plant identification is practical and inexpensive, but can be difficult and unsustainable for the diversity of existing plant species (Wang, Wu et al. 2010). Moreover, it has important limitations, such as phenotypic plasticity of some plants, effective morphological keys limited to a specific stage of life or genus, impossibility of identification in samples of mixture and the requirement of a high level of specialization on the part of the taxonomists (Hebert, Cywinska et al. 2003, Li, Ye et al. 2015, Zeng, Zhou et al. 2017).

Given the limitations inherent in using morphological characteristics, it becomes important to develop techniques that ensure their correct identification. Molecular diagnostic approaches and DNA-based methods have become a powerful tool for identifying plant species (Derocles, Evans et al. 2015, Li, Yang et al. 2015). DNA is more stable, resistant and thermostable than proteins are, less affected by external conditions and could potentially be retrieved from any substrate because it is present in almost all cells of an organism (Bustin 2005, Lee, Ng et al. 2016, Mishra, Kumar et al. 2016). DNA markers are independent of morphology, stage of development and environmental factors besides being particularly useful for distinguishing morphologically similar species and having a long stage of vegetative growth. In addition, molecular evolution and phylogenetics have shown that, because of the degeneracy of the genome and the presence of many non-coding regions, DNA can provides more information than proteins do (Teletchea, Maudet et al. 2005, Lin, Lin et al. 2015).

DNA-based techniques have enabled researchers to identify and authenticate several species in a simple, fast and inexpensive way. However, a universally accepted approach to solving all the problems associated with identifying of plants species is not available (Scriver, Marinich et al. 2015). Currently available techniques have different degrees of resolution, information generation and applicability, depending on the taxonomic level (Nam, Lee et al. 2015, Thomsen and Willerslev 2015). Almost all traditional DNA-based methods rely on the Polymerase Chain Reaction (PCR) to produce multiple copies of the genome region of interest (Hwang, Kim et al. 2015). The PCR technique and its variants can be used to explore specific variations in the DNA sequence, in the identification of species and to track food origins, successfully applied in the detection of Genetically Modified Organisms (GMOs) and pathogens in food products, among others (Moon, Kim et al. 2016). The multiplex PCR is a derivation approach of PCR technique that allows the simultaneously amplification of two or more different DNA sequences in a single reaction by using a combination of different primers. The advantages of this technique are the flexibility, the speed and the reduced cost. The main challenges are the possible inhibition between primer sets, the false amplification and the lack of efficacy in different sample matrixes. The design of primers for multiplex PCR is an important step in the procedure since it is necessary to amplify different targets with the same reaction conditions (Ronning, Rudi et al. 2005, Moon, Kim et al. 2016).

In order to supplement and ensure the data obtained by PCR and electrophoresis, DNA sequencing is the most widely used technique because of its high productivity and accuracy, generating a large amount of data quickly and inexpensively (Feuillet, Leach et al. 2011, Yang, Li et al. 2014, Sarwat and Yamdagni 2016). The main drawback of DNA sequencing approaches is to obtain clear sequences of large regions, particularly difficult in samples with low quality and/or low amounts of DNA (Pereira, Carneiro et al. 2008). The next-generation sequencing technologies (NGS), or high throughput sequencing, handle millions of small DNA fragments in parallel. Despite the production of a large number of sequences at low cost, they require a more purified DNA and the quality of the sequences produced are generally of lower quality and shorter than those obtained by the Sanger sequencing (Cheng, Guo et al. 2003, Feuillet, Leach et al. 2011).

The DNA barcoding concept seeks to identify biological specimens and assign them to a specific species using a standardized genomic region called DNA barcode, which corresponds to small part (<1000bp) that can be easily obtained from the genome (Lahaye 2008, Staats, Arulandhu et al. 2016). The concept of DNA barcode was first proposed by Hebert, Cywinska et al. (2003) and has attracted the attention of the world's scientific community. In order to achieve a good discriminating power, the locus barcode must be sufficiently informative and variable to unequivocally differentiate neighbouring species in its taxonomic group but conserved sufficiently in the same species to define a clear threshold between intra and inter-specific diversity. The sequence variation of a barcode marker between species must be high enough to distinguish them, whereas the distance within the species must be sufficiently small. This difference in distances is known as the “DNA barcode gap". An effective barcode becomes weak when interspecific and intraspecific distances overlap. Another desirable feature for an ideal barcode locus is to have highly conserved connection sites that allow the reliable amplification and bidirectional sequencing with a single pair of primers. This is particularly important in a mixture of samples so that several species can be identified at the same time (Wang, Wu et al. 2010, Vassou, Kusuma et al. 2015, Kikkawa, Tsuge et al. 2016, Mishra, Kumar et al. 2016). The DNA barcode is a simplified solution for a complex problem that is difficult to apply to all species (Mishra, Kumar et al. 2016). However, this approach has been applied in the authentication of medicinal plants marketed, food safety, monitoring of biodiversity and conservation, control of illicit trafficking of protected species, forensic botany, among others (Hajibabaei, Singer et al. 2007, Zaiko, Martinez et al. 2015, Mishra, Kumar et al. 2016).

The combination of NGS with the DNA barcoding is known as metabarcoding. The *meta* prefix refers to multiple species identified simultaneously from complex samples such as faeces, soil, seawater and environmental mass samples (Zaiko, Martinez et al. 2015, Valentini, Taberlet et al. 2016). The main limitations associated with the use of DNA metabarcoding are the unavailability of a truly universal extraction method; the discriminatory power of the bioinformatics methods used in the analyses; the PCR bias caused by different primers; the high dependence of a large reference databases with quality sequences based on good taxonomy and coverage; the reduced sequencing costs and the achievement of sufficiently long quality sequences. The approach of DNA barcoding was tested using different genomic regions (Ratnasingham and Hebert 2007, Bhargava and Sharma 2013, Staats, Arulandhu et al. 2016).

The mitochondrial gene cytochrome oxidase I (COI) is considered a universal barcode DNA for animals because the genetic variation of this locus is sufficient to study the process that occurs in relatively short and recent intervals of time, the same way that it has conserved regions that allow the design of primers (Lahaye 2008, Geller, Meyer et al. 2013). However, mitochondrial genes in plants are rarely used for species identification due to their slow evolution, low nucleotide substitution and limited divergence (Staats, Arulandhu et al. 2016, Gualberto and Newton 2017). Nuclear DNA (nuDNA) genes can be used, but their applicability is limited by the absence of conserved primers for their amplification. However, because of sufficient intra-species conservation and interspecies specificity, the nuclear rRNA genes are successfully used as targets for identification of fungal (Wang, Fu et al. 2014) and bacterial species (Marsh, O'Sullivan et al. 2014).

Therefore, the search for a region barcode for plants has been moved to the chloroplast genome (Chen, Yao et al. 2010, Thomsen and Willerslev 2015). The chloroplast DNA (cpDNA) presents valuable regions for phylogenetic analyses of high taxonomic levels. However, methods targeting a single cpDNA locus provided insufficient variability for species identifications (Li, Yang et al. 2015, Staats, Arulandhu et al. 2016). After considerable effort to find a sufficiently informative locus comparable to the COI used in animals, some researchers have suggested a multi-locus approach, where combined barcodes could present increased discrimination of species (Saddhe, Jamdade et al. 2017).

Several groups have tested different combinations, the Plant Working Group (PWG) for the Consortion for the Barcode of Life (CBOL) examined the suitability of seven candidates (*matK*, *rbcL*, *trnH-psbA*, *atpF-atpH*, *rpoB*, *rpoC1* and *psbK- psbI*) and proposed the *matK* and *rbcL* regions as core barcode for plants. This combination has been suggested because of its universality, easy recovery of *rbcL* and the good discriminatory power of *matK* sequences, but it cannot avoid the low effectiveness of *matK* in PCR due to lack of universal primers and low power discrimination of *rbcL*. The combination offers only a slightly high identification efficiency compared to previous methods. Some researchers suggested the use of the ITS nuclear locus (nrITS) and the *psbA-trnH* intergenic space as additional loci. The CBOL recognizes that any combination is far from ideal (Chen, Yao et al. 2010, Wang, Wu et al. 2010, Vassou, Kusuma et al. 2015, Staats, Arulandhu et al. 2016). An approach based on nuclear and organelle genomes could be more satisfactory because uniparental inheritance can never show the plant complex (Yao, Song et al. 2010).

The multiple locus strategy has opened new avenues for species identification. However, the combination of barcodes increases the difficulties of analysis with respect to the single locus. The failure of the barcode approach is not simply due to the lack of variation but also reflects the differences between the genetic trees of the plastid genes and the species boundary. The combination of loci does not eliminate the inherent deficiencies derived from the plant DNA barcoding. Barcode markers have been proposed to identify hotspots of biodiversity in distant organisms, but few studies have developed barcodes for identification in family, genus or between close relatives. The discriminatory potential of the DNA barcode varies from one family to another (Wang, Wu et al. 2010, Vassou, Kusuma et al. 2015, Saddhe, Jamdade et al. 2017).

The use of the complete cpDNA as a single marker circumvents possible problems such as altered gene order, low PCR efficiency and relatively short DNA sequences (Hajibabaei, Singer et al. 2007, Nock, Waters et al. 2011). The problems associated with the complete sequencing of cpDNA are the high costs and difficulties associated with obtaining complete sequences. For instance, the complete cpDNA of *Salvia miltiorrhiza* is 151,328 bp in length (Qian, Song et al. 2013), the *Theobroma cacao* have a chloroplast genome of 160,546 bp (Kane, Sveinsson et al. 2012), the *Lactuca sativa* chloroplast DNA is 152,772 bp in length (Timme, Kuehl et al. 2007). However, for lineages that radiate rapidly, the use of a single genome remains ineffective. Until now, it is not clear whether the complete plastid genome can be considered as an adequate barcode, but the results show that it can contribute to the identification of plant species. Although the cost of sequencing has decreased considerably, current costs for the complete cpDNA sequencing are even greater than those of a single locus barcode by traditional sequencing. Even excluding these factors if plastid identification depends on a fully annotated chloroplast sequence, the necessary analyses can be complex and difficult to normalize (Petit, Duminil et al. 2005, Zeng, Zhou et al. 2017).

The DNA barcoding had a positive impact on biodiversity rankings and identification of plants species. This approach benefits with the development of NGS but is still far from being completely viable to the identification of species, especially at deeper levels. However, despite all the contributions and progress made in species identification techniques, it is expensive and impractical with respect to gel-based DNA markers and is still possible to develop new methods that will help overcome the inherent limitations encountered in this area of science (Pereira, Carneiro et al. 2008, Parker and Helmstetter 2017).

The presence of insertions/deletions (indels) is responsible for length variation of a DNA sequence when comparing samples (Taberlet, Gielly et al. 1991). The study of indels proved helpful in species identification (Jin, Jin et al. 2014, Mahadani and Ghosh 2014). High levels of species identification have been achieved in different taxa (animals, fungi and bacteria) through the determination of the length variation of the sequences caused by the indels (Carneiro, Pereira et al. 2012, Gonçalves, Marks et al. 2015, Hwang, Kim et al. 2015, Alves, Pereira et al. 2017). The use of indels polymorphisms for the identification of species may be advantageous if the intra-species variability is lower than that of SNPs. Indels are less prone to recurrent mutations (i.e. identical insertions or deletions occurring in independent lineages), which means that there is a low probability that similar sequences originated by convergence (homoplasy). The insertion of a nucleotide that restores a previous deletion at the same position or vice versa (a phenomenon known as ‘back mutation’) is also very unlikely in this class of polymorphisms (Pereira, Carneiro et al. 2010).

The SPecies Identification by Insertions/Deletions (SPInDel) method uses the length of hypervariable genomic regions (regions containing multiples indels) that are found interspersed with highly conserved regions (regions presenting none or low sequence variability) that delimitate the variable segments like anchors. Therefore, each species can be identified by a unique numeric profile of fragment lengths resulting from the combination of the length of hypervariable regions (a 'SPInDel profile') (Figure 1) (Pereira, Carneiro et al. 2010, Gonçalves, Marks et al. 2015).

The SPInDel method has already been applied to discriminate a large sample of eukaryotes (1556 species) analysed through the rRNA genes of the mitochondrial genome and was able to assign a unique profile to 1451 species (95%) (Pereira, Carneiro et al. 2010). The red fox (*Vulpes vulpes*) was differentiated from the other species (human, common domestic livestock and Australian endemic wildlife species) through the combination of SPInDel method and multiplex PCR analysis of mitochondrial 12S and 16S gene regions. The strategy proved effective because at least two hypervariable regions had a significant divergence from all samples (Gonçalves, Marks et al. 2015).

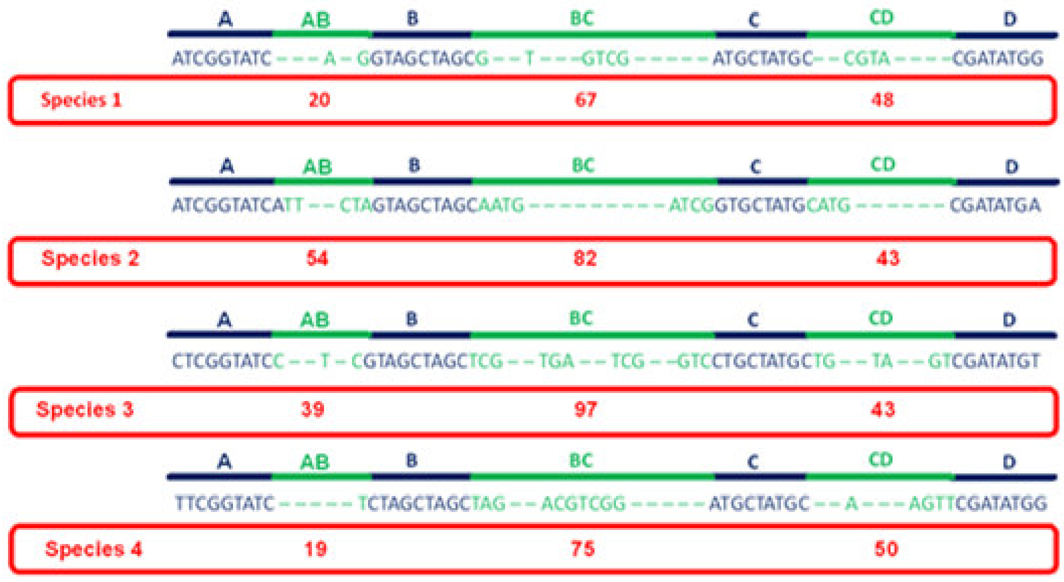
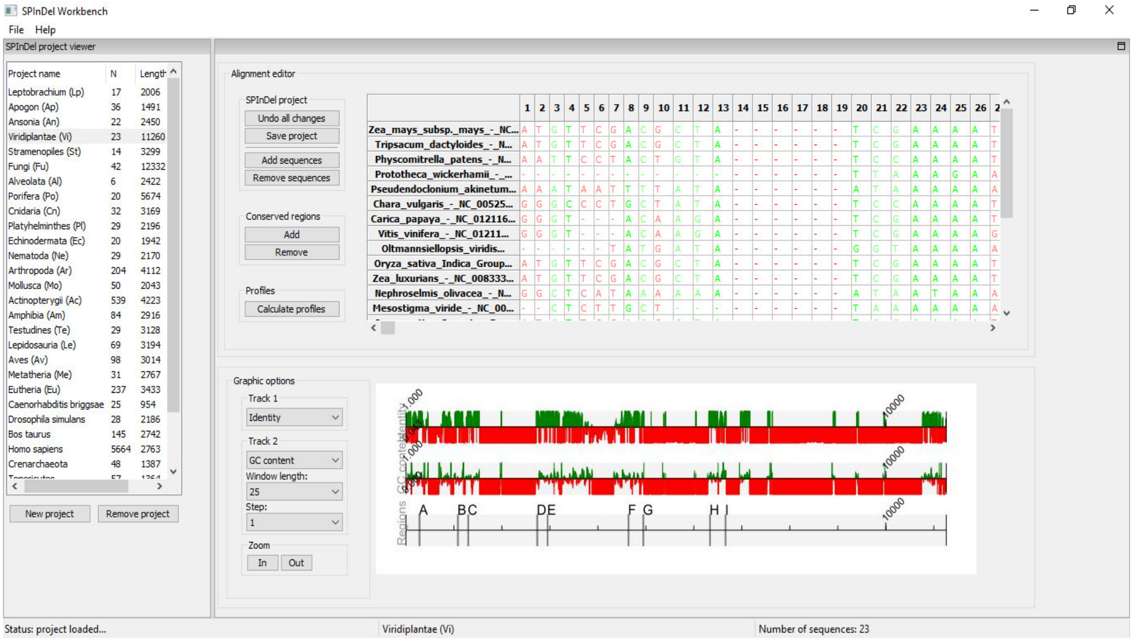

A AB B BC C CD D
ATCGGTATC - - - A - GGTAGCTAGCG - - T - - -GTCG - - - - - ATGCTATGC- - CGTA - - - - CGATATGG
Species 1 20 67 48
A AB B BC C CD D
ATCGGTATCATT - - CTAGTAGCTAGCAATG - - - - - - - - - - ATCGGTGCTATGCATG - - - - - - CGATATGA
Species 2 54 82 43
A AB B BC C CD D
CTCGGTATCC - - T - CGTAGCTAGCTCG - - TGA - - TCG - - GTCCTGCTATGCTG - - TA - - GTCGATATGT
Species 3 39 97 43
A AB B BC C CD D
TTCGGTATC - - - - - TCTAGCTAGCTAG - - ACGTCGG - - - - - ATGCTATGC - - A - - - AGTTCGATATGG
Species 4 19 75 50

Figure 1. Schematic illustration of the strategy used in the species identification by the insertions/deletions (SPInDel) method. Illustration of the sequence alignment for four hypothetical species. 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 box).

The SPInDel workbench is a computational platform that was built to facilitate the planning and project management and alignment of nucleotide sequences, visualization and selection of conserved regions, calculation of the properties of PCR primers properties, prediction of SPInDel profiles and diverse statistical and phylogenetic analyses. It includes a large database comprising nearly 1,800 numeric profiles for the identification of eukaryotic, prokaryotic and viral species. For 'Viridiplantae' SPInDel workbench provides 23 sequences (Figure 2). The SPInDel computational workbench available in http://www.portugene.com/SPInDel/SPInDel_web.html can be used with sequence data from any genomic region and is a useful tool to help researchers in all steps of the species-identification workflow.



Project name	N	Length
Leptobrachium (Lp)	17	2006
Apogon (Ap)	36	1491
Ansonia (An)	22	2450
Viridiplantae (Vi)	23	11260
Stramenopiles (St)	14	3299
Fungi (Fu)	42	12332
Alveolata (Al)	6	2422
Porifera (Po)	20	5674
Cnidaria (Cn)	32	3169
Platyhelminthes (Pl)	29	2196
Echinodermata (Ec)	20	1942
Nematoda (Ne)	29	2170
Arthropoda (Ar)	204	4112
Mollusca (Mo)	50	2043
Actinopterygii (Ac)	539	4223
Amphibia (Am)	84	2916
Testudines (Te)	29	3128
Lepidosauria (Le)	69	3194
Aves (Av)	98	3014
Metatheria (Me)	31	2767
Eutheria (Eu)	237	3433
Caenorhabditis briggsae	25	954
Drosophila simulans	28	2186
Bos taurus	145	2742
Homo sapiens	5664	2763
Crenarchaeota	48	1387

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	2
Zea_mays_subsp._mays_-_NC...	A	T	G	T	T	C	G	A	C	G	C	T	A	-	-	-	-	-	-	T	C	G	A	A	A	A	T
Tripsacum_dactyloides_-_N...	A	T	G	T	T	C	G	A	C	G	C	T	A	-	-	-	-	-	-	T	C	G	A	A	A	A	T
Physcomitrella_patens_-_N...	A	A	T	T	C	C	T	A	C	T	G	T	A	-	-	-	-	-	-	T	C	C	A	A	A	A	T
Prototheca_wickerhamii_-_...	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	T	T	A	A	A	G	A	A
Pseudendoclonium_akinetum...	A	A	A	T	A	A	T	T	T	T	A	T	A	-	-	-	-	-	-	A	T	A	A	A	A	A	A
Chara_vulgaris_-_NC_00525...	G	G	G	C	C	C	T	G	C	T	A	T	A	-	-	-	-	-	-	T	C	C	A	A	A	A	T
Carica_papaya_-_NC_012116...	G	G	G	T	-	-	-	A	C	A	A	G	A	-	-	-	-	-	-	T	C	G	A	A	A	A	T
Vitis_vinifera_-_NC_01211...	G	G	G	T	-	-	-	A	C	A	A	G	A	-	-	-	-	-	-	T	C	G	A	A	A	A	G
Oltmannsiellopsis_viridis...	-	-	-	-	-	-	T	A	T	G	A	T	A	-	-	-	-	-	-	G	G	T	A	A	A	A	A
Oryza_sativa_Indica_Group...	A	T	G	T	T	C	G	A	C	G	C	T	A	-	-	-	-	-	-	T	C	G	A	A	A	A	T
Zea_luxurians_-_NC_008333...	A	T	G	T	T	C	G	A	C	G	C	T	A	-	-	-	-	-	-	T	C	G	A	A	A	A	T
Nephroselmis_olivacea_-_N...	G	G	C	T	C	A	T	A	A	A	A	A	A	-	-	-	-	-	-	A	T	A	A	T	A	A	A
Mesostigma_viride_-_NC_00...	-	-	C	T	C	T	T	G	C	T	-	-	-	-	-	-	-	-	-	T	A	A	A	A	A	A	A

Graphic options
Track 1: Identity
Track 2: GC content
Window length: 25
Step: 1
Zoom: In Out

Status: project loaded... | Viridiplantae (Vi) | Number of sequences: 23

Figure 2. Main page of SPInDel workbench. The green segments represents conserved regions (potential primer binding sites), and the red ones represent hypervariable regions.

The plant genomes - nuclear DNA (nuDNA)

The nuclear genome of plants is diverse, ranging from 38Mb to 87,000Mb (Arumuganathan and Earle 1991, Bennetzen 2000, Su, Chao et al. 2013, Xu, Chen et al. 2013). The size and complexity of the nuDNA makes difficult its sequencing due to several types of rearrangements like inversions, deletions and translocations, besides polyploidy and gene duplication (Bennetzen 2000, Feuillet, Leach et al. 2011, Daniell, Lin et al. 2016, Gualberto and Newton 2017). The main factors responsible for the variation in the size of the nuclear genomes of the plant are the ploidy level (from diploid to octaploid and higher); number of repetitions (simple repeating tandem for example) and transposable elements and recurrent exclusions of DNA.

Closely related plant lineages may differ considerably in the size of the genome. Even in smaller genomes, such as *Arabidopsis*, repeated fragments represent more than 20% of the nuDNA. The low quantity of nuDNA is not always associated with the small size or short life cycle of the species. Within a species, nuDNA tends to be conserved, but between species, it can vary considerably, even among species of the same genus. The size of the genome varies greatly between species but is not related to the size and number of chromosomes. Genes of plants are relatively compact and often grouped with smalls introns. Nuclear genes from a single copy are less influenced by evolution and convergent recombination, but have rarely been used for plant phylogenetic reconstruction (Arumuganathan and Earle 1991, Koch, Haubold et al. 2001, Kellogg and Bennetzen 2004, Feuillet, Leach et al. 2011).

The plant genomes - Mitochondrial genome (mtDNA)

The mitochondrial genome (mtDNA) is derived from an ancestor of endosymbiotic prokaryotes. In most terrestrial plants, the mode of transmission of mtDNA is of maternal heritage. In plants as in other eukaryotes, mtDNA encodes a small number of essential components of the mitochondrial electron transfer chain. For the expression of these genes, the mitochondria have their own translation system, which is also partially encoded by mtDNA, including rRNAs, tRNAs and a varied number of ribosomal proteins. However, all the factors necessary for maintenance of mtDNA and the expression of its genes are encoded in the nucleus and imported from the cytosol, thus placing mtDNA replication, structural organization and expression of the genes under nuclear control (Parson, Pegoraro et al. 2000, Gualberto and Newton 2017).

The number of mitochondrial genes varies considerably between related species and even within a species. Many genomes include unknown genes and can be rapidly gained or lost, contributing to the intraspecific diversity of mtDNA. The mtDNA size is highly variable and the mitochondrial genomes of terrestrial plants are by far the largest, which vary between 200-700kb and can reach 11Mb (in *Silene conica*) (Gualberto and Newton 2017). Plant mtDNA contains some additional genes and several genes contain introns, characteristics that contribute to a large variation size. The mitochondrial genome of plants is abundant in non-coding repeated sequences of different sizes and numbers, usually not conserved within a species. The greatest variability in the structural organization of plant mtDNA is the presence of active recombination of long repeats. It is also possible that the mtDNA acquire new exogenous sequences by horizontal transfer derived from cpDNA, nuDNA or viral DNA (Parson, Pegoraro et al. 2000, Petit, Duminil et al. 2005).

The mtDNA of the plants evolves more slowly than of animals and genetic sequences have low nucleotide substitution rates, which does not promote sufficient variability for species discrimination (Bennetzen 2000, Lahaye 2008, Daniell, Lin et al. 2016, Staats, Arulandhu et al. 2016). The reason for this low variability may reside in existence of effective repair pathways, in particular an active homologous recombinant system, which potentially corrects the mutations (Notsu, Masood et al. 2002, Hebert, Cywinska et al. 2003).

The plant genomes - Chloroplastidial genome (cpDNA)

Plastids are essential organelles for plant physiology and development, including the synthesis of amino acids, nucleotides, fatty acids, phytorones, pigments, starches, vitamins and metabolites, the assimilation of sulphate and nitrogen, among others. Metabolites administered by plastids are important for the plant-environment relationship, for example, response to salinity, light, heat, drought, defence against pathogens, among others (Daniell, Lin et al. 2016).

Chloroplasts are a class of essential organelles, distinct and highly specialized plastids present in plant cells and algae. These intracellular organelles carry their own genome coding for many (but not all) genes essential for photosynthesis, so chloroplasts are responsible for capturing sunlight and converting the organic substance (carbohydrates) with the release of oxygen. Taking into account the size, content and gene organization of cpDNA, it is believed that chloroplasts evolved from endosymbiosis of a free-living cyanobacterium and were hosted by a nucleated cell, followed by several eukaryotic symbiosis and massive transfer of chloroplast genes to the nucleus. Although their evolution is strongly related to that of the host cell, the plastid genome does not necessarily follow the same evolutionary history of the host genome. Significantly different substitution rates, structurally independent replication and other biological processes, may lead to a divergent and incongruent evolution between chloroplast, mitochondrial and nuclear loci (Petit, Duminil et al. 2005, Pérez-Escobar, Balbuena et al. 2015, Wang, Cui et al. 2015, Daniell, Lin et al. 2016, Moon, Kim et al. 2016).

The cpDNA can range from 107kb to 2500kb. Despite this variation in length, generally associated with large scale rearrangements, gene duplication and small replicates; cpDNA is considered stable and conserved in terms of structure and genetic content. The cpDNA is present in several copies in one cell (Bennetzen 2000, Ronning, Rudi et al. 2005, Xu, Liu et al. 2015, Daniell, Lin et al. 2016, Gualberto and Newton 2017). The cpDNA is an independent and densely compact molecule of circular structure, usually divided into four sections, two of which are copies of an inverted region, IR-Inverted Region (+/- 25kb), separating two regions of single copy , LSC - large single copy (+/- 87kb) and SSC - small single copy (+/- 18kb) (Yang, Li et al. 2014, Zeng, Zhou et al. 2017) (Figure 3). The main cause of variation in cpDNA size is the difference in length of LSC and IR, particularly in the contraction and expansion of LSC and SSC junctions (Curci, De Paola et al. 2015).

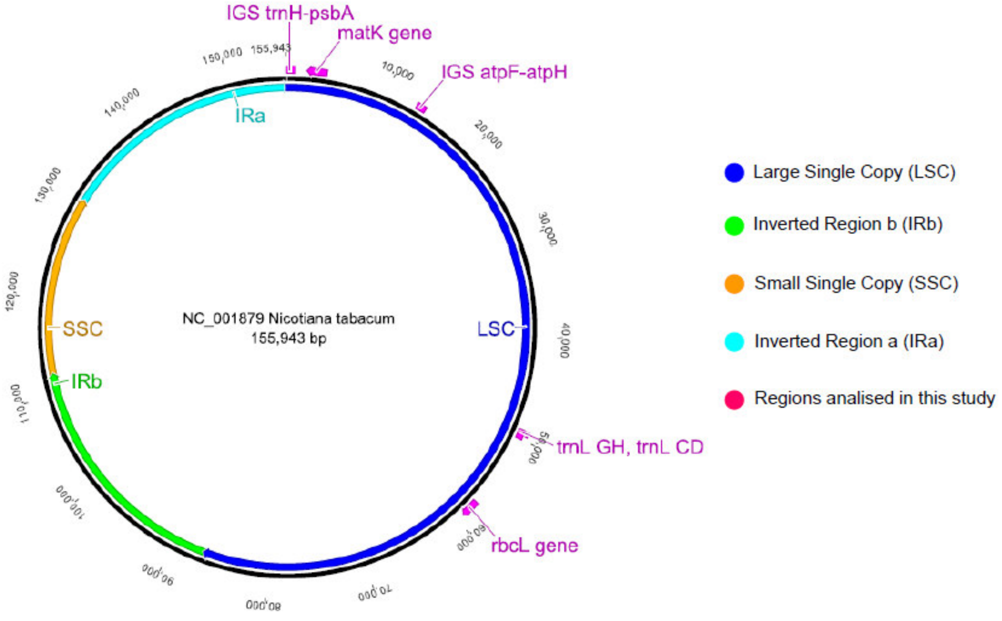


Figure 3. Representation of cpDNA of Nicotiana tabacum (NC_001879), highlighting the regions analysed in this thesis.

The typical cpDNA of terrestrial plants is formed by about 120-133 genes, which encode about 4 to 8 rRNAs, 30 to 37 tRNAs, 85 to 88 proteins, most of which have a known function and some of unknown function (Yang, Tang et al. 2013, Zeng, Zhou et al. 2017). The primary products of chloroplast genes have a role in photosynthesis and transcription-translation. Genes used in photosynthesis tend to be more conserved than ribosomal proteins and other genes. Many chloroplast genes are functionally grouped in polycistronic operons such as those containing the four ribosomal genes, *atpI*-H-F-A, *atpB*-E. The order and mode of expression of the genes in these operons are very similar to those observed in prokaryotes. The main structural difference between some chloroplastic and prokaryotic genes is the presence of introns. The cpDNA of some terrestrial plant lineages shows significant structural rearrangements, with an obvious loss of IR or whole genes. Although introns are generally conserved, most of the loss of these structures within the genes encoding was observed in specific groups or species. Comparative sequence analyses showed that the cpDNA has genes with similar sequences present in the mtDNA, but in the chloroplast the function of these genes is unknown (Palmer, Jansen et al. 1988, Xu, Liu et al. 2015, Daniell, Lin et al. 2016).

The cpDNA is haploid, with maternal inheritance, with little or no recombination, low nucleotide substitution rate and an average growth rate 4 times slower than nuDNA in plants. Variations in cpDNA provide higher resolution at the population level than nuclear markers, characteristics that make the cpDNA suitable for comparative genomic studies (Li, Yang et al. 2015, Moon, Kim et al. 2016). Mutations in cpDNA are essentially two types: point mutations (substitution of a single nucleotide pair) and rearrangements. The most frequent mutations are the point mutations and insertions/deletions (indels) in noncoding regions (Yang, Tang et al. 2013, Daniell, Lin et al. 2016). However, the rate of change of the chloroplast differs depending on its location in the genome and between genes. Typically, the rate of evolution and the nucleotide substitution rate of the LSC and SSC regions is higher than the IR. The IR and coding regions of the chloroplast genome are more conserved (low AT content high CG content) relative to the SC and non-coding regions, respectively (Zeng, Zhou et al. 2017). Direct sequencing studies reveal different levels of nucleotide substitution between chloroplast-specific genes. The rate of substitution in the cpDNA genes is on average two to three times lower than that of mitochondrial animal genes, but three to four times higher than mitochondrial plant genes. It is often the genome of choice for phylogenetic analysis in plants (Curci, De Paola et al. 2015, Li, Yang et al. 2015, Moon, Kim et al. 2016).

As the evolution of mitochondrial genome in most plants is too slow, it cannot be used to distinguish between species. Various genes and non-coding regions in the plastid genome have been put forward as alternatives (Sarwat and Yamdagni 2016). Molecular differentiation arisen in cpDNA among plant species and even individuals offer-promising tools for phylogenetic reconstruction and species identification. Recently, a few studies have discussed using complete chloroplast genomes to identify species or as organelle-scale barcodes (Yang, Tang et al. 2013, Li, Yang et al. 2015). Complete cpDNA sequencing is being used for obtaining evolutionary information that can be used to address questions of species identification and phylogenetic analyses of plants (Yang, Li et al. 2014).

The cpDNA has conserved coding regions that can be easily aligned and used for primer design, which can be intercalated by variable introns or intergenic regions. The analyses of both these regions produce a structure capable of resolving inter and intraspecific relationships at different phylogenetic levels (Panero and Crozier 2003, Neubig, Whitten et al. 2009, Yang, Kung et al. 2015). Molecular markers in cpDNA can be used to identify commercial varieties of cultivars, determine purity and preserve production resources (Wang, Cui et al. 2015, Daniell, Lin et al. 2016).

Genomic regions for plant species identification

atpF-atpH

The *atpF* and *atpH* genes encode the ATP synthase subunit CFO I and CFO III, respectively. It is a non-coding space with high inter-specific variability due to the presence of indels (Lin, Lin et al. 2015). It was reported that, compared to other markers, *atpF-atpH* was the one with the best inter- and intra-species ratio, with sufficient inter-specific but relatively low intra-specific divergence. The adequate variation and narrow range of overlap of the *atpF-atpH* marker can ensure correct identification of species. It is a recommended molecular marker due to high amplification in PCR, easy alignment and sufficient divergence in sequences (Table 1) (Wang, Wu et al. 2010).

psbA-trnH

The *psbA-trnH* region includes the chloroplast genome space between the *psbA* and *trnH* genes. The *psbA* regulatory region (3'UTR) is of utmost importance in the regulation and expression of the *psbA* gene, which encodes the chloroplast (D1 of photosystem II) protein (Daniell, Lin et al. 2016). It is a highly variable locus, with high interspecific divergence due to the high frequency of nucleotide repeats, micro inversions and indels. The presence of a duplicate loci and a pseudogene makes *psbA-trnH* sequences in some species (conifers >1000bp), shorter in others (monocotyledons <300bp) and extremely short in others (bryophytes <100bp). This variation in length is considered unfavourable because it imposes difficulties in the alignments (Chen, Yao et al. 2010, Wang, Wu et al. 2010, Li, Yang et al. 2015, Tang, Yukawa et al. 2015, De Castro, Comparone et al. 2017). However, it is a widely used plastid region, because short spaces show sufficient variation, being considered an excellent phylogenetic marker, even to resolve interspecific relationships. Long *psbA-trnH* regions can be difficult to recover without primers specially designed to obtain high-quality bidirectional sequences. However, the presence of highly conserved coding sequences at both ends allows the design of such oligonucleotides (Table 1) (Lahaye 2008, Kumar, Mishra et al. 2016).

trnL

The *trnL* intron (UAA) is a non-coding region of the chloroplast genome encoded in the large single-copy region (LSC) (Figure 3). It is part of the group I introns, which show a mosaic structure of conserved elements and common secondary structure elements, which are essential for correct splicing, and less constrained regions of variable size (Quandt and Stech 2005). The region presented low taxonomic resolution and was not variable enough to differentiate related species but can be used to identify commonly consumed plants (Bruneau, Forest et al. 2001). Its evolution in land plants is well understood and it has been often used to study relationships among genera, reconstructing phylogenies between distantly related groups or for identifying plant species. It shows an acceptable discrimination efficiency for the needs of food analysis, since it is sufficiently variable among species and conserved enough within species (Kajita, Kamiya et al. 1998, Quandt and Stech 2005, Spaniolas, Bazakos et al. 2010). The food industry and forensic science has used extensively the *trnL* (UAA) intron, in particular due to the small size of the P6 loop (10-143 bp), where it is difficult to obtain fragments greater than 150pb (Taberlet, Coissac et al. 2007, Thomsen and Willerslev 2015).

The *trnL* is not considered the most informative noncoding region of cpDNA, but a large number of nucleotide sequences are available in public databases. This abundance is due to the availability of highly conserved primers (important for PCR), from bryophytes to angiosperms. The presence of A/T >10bp stretches and the frequent presence of indels mutation makes the short P6 loop also exhibit some intraspecific variation (Quandt and Stech 2005, Taberlet, Coissac et al. 2007). The design of universal primers is viable due to highly conserved gene encoding sequences flanking interesting noncoding regions (Table 1). Hotspots rich in A/T nucleotides, with respect to the rest of the introns, have already been documented in this intron, resulting in variable length polymorphisms (Ronning, Rudi et al. 2005).

rbcL

The ribulose – 1,5 – biphosphate carboxylase/oxygenase is a cpDNA gene highly conserved, encode the big subunit of enzyme (RuBisCO) the 476 amino acids protein responsible for CO_2 binding. It has a relatively slow rate of evolution, being the locus with the slightest divergence between the plastid genes of the plants; therefore, it is not suitable at the species level because of the modest discriminatory power. The *rbcL* present low ability in resolving phylogenetic relationships below the family or gender levels (Taberlet, Coissac et al. 2007, Dong, Cheng et al. 2014), despite this it is one of the more characterized plastid coding regions, taking into account the number of sequences available in the databases. This sequence availability is due to its great universality, which allows the design of primers, easy amplification (despite the size), generating quality sequences and unequivocal alignments for most terrestrial plants (Table 1) (Mishra, Kumar et al. 2016, Staats, Arulandhu et al. 2016).

The *rbcL* alone does not fulfil the attributes for a barcode locus, although it can be useful for species identification when combined with other plastics or nuclear loci (Li, Yang et al. 2015). The Plant Working Group of Consortium for Barcode of Life (CBOL) suggested the use of approximately 650bp at the 5' end of the *rbcL* gene for the combination of two locus (*rbcL* and *matK*) as the nucleus barcode. Inadequate performance at species and genus levels is particularly due to the selection of a relatively conserved region in the gene; so that regions with greater variability may be present (Dong, Cheng et al. 2014).

matK

The *matK* plastidial gene codes for the maturase protein that is important in splicing (modification/binding) process. It is a region that is subject to different selective pressures that, when positive, help the species adapt to heat and dry climate (Daniell, Lin et al. 2016). It is a coding region that has a high rate of evolution and rapid substitution, rare occurrence of indels, adequate length and interspecific divergence (Table 1) (Mishra, Kumar et al. 2016). The *matK* sequences are used to study phylogenetic and evolutionary relationships at all taxonomic levels. The *psbA-trnK* space includes the complete *matK* gene and adjacent regions (Koch, Haubold et al. 2001). However, the *matK* barcode space used in the analyses consists of an 841bp segment at the centre of the gene and is considered to be a COI-like region used as a barcode in animals (Staats, Arulandhu et al. 2016).

This locus was proposed as a barcode for plants by Lahaye (2008), but the unavailability of universal primers for all taxa leads to a low rate of amplification by PCR and is often a limiting factor for the use of this region (Yu, Xue et al. 2011). The divergence of the *matK* sequences is greater than that of other coding regions, evolving about two to three times faster than *rbcL*, thus enhancing support at different taxonomic levels (Techen, Parveen et al. 2014, Sarwat and Yamdagni 2016). Although *matK* often does not show sufficient variability for discrimination at low taxonomic levels (Neubig, Whitten et al. 2009, Daniell, Lin et al. 2016), it showed highly variable sequences in the species *Oryza sativa*, *Zea mays* and *Triticum aestivum* (Poaceae) (Yang, Kung et al. 2015) and in Orquidaceae family species, but differentiated less than 49% of the Myristicaceae family species (Saddhe, Jamdade et al. 2017).

ITS

The Internal Transcribed Spacer (ITS) comprises the 5.8S locus and its adjacent regions ITS1 and ITS2, each with about 300bp. It is a nuclear ribosomal gene, considered to be a good phylogenetic marker, with high levels of inter and intraspecific divergence. It generally contains sufficient phylogenetic evidence for plant discrimination, even at low taxonomic levels (Table 1). Because of the discriminatory power of ITS on plastid regions, it has been proposed as a standard nuclear barcode (Chen, Yao et al. 2010).

The limitations associated with the use of this marker are the presence of putative pseudogenes leading to sequencing difficulties in many groups and paralogy. The fungal ITS sequences have a great similarity with those of plants. The primers used to amplify and sequence the two groups are similar, so that the fungal DNA can sometimes be amplified, preferably or confused, especially in plants containing fungal endophytes (Chen, Yao et al. 2010, Yao, Song et al. 2010). The available primer sets are problematic for several samples, making amplification difficult (Table 1). Despite the problems associated with its use, many studies suggest the use of ITS (Mishra, Kumar et al. 2016).

The ITS2 was considered a highly informative region to discriminate among related plant species and taxonomic studies (Gao, Yao et al. 2010, Liu, Zeng et al. 2012, Saddhe, Jamdade et al. 2017). The ITS2 was used to discriminate more than 6600 medicinal plants, showing a rate of identification of 92.7% at the species level. This markers has several available sequences, is a short region (160-320bp) easy to align and can be amplified using universal primers. It has a high and well-defined interspecific divergence (barcode gap). However, it often presents unsatisfactory quality levels in sequencing due to the existence of rich AT regions or homologous sequences (Chen, Yao et al. 2010).

Table 1. Comparative view of interest parameters used in species identification for the analysed genomic regions.

	atpF-atpH	***psbA-trnH***	***trnL***	***rbcL***	***matK***	**ITS**
Universality	intermediate	low	high	high	low	intermediate
Alignment	intermediate	low	high	high	intermediate	high
Amplification	high	high	high	high	low	low
Sequencing	high	low	high	high	low	low
Design/ availability of primers	high	high	high	high	low	high
Sequences available in Gene Bank	low	intermediate	high	high	high	high
Divergence/ variation	high	high	low	low	intermediate	high
Discriminatory power (species level)	high	high	interme diate	interme diate	intermediate	high
References	(97) (146) (168)	(40) (56) (105) (146)	(97) (162)	(56)	(56) (67) (92) (116) (146) (158)	(24) (40) (96) (181)

Other genomic regions

Other genomic regions have been used in different analyses and are proposed as complementary or ideal markers, depending on the objective of the study. These are less exploited regions and therefore fewer sequences are available in databases. The *rpoC2* (RNA polymerase beta-prime chain) chloroplast gene sequences were used to differentiate species from the Poaceae family (Moon, Kim et al. 2016, Zeng, Zhou et al. 2017). The *psbK-psbI* is the intergenic space between the *psbK* and *psbI* genes, which encode two low molecular weight polypeptides, K and I, respectively, of the photo system II. This region showed good PCR performance and sequencing, sequence alignments were not problematic and showed moderate inter-specific diversity (Lahaye 2008).

The *ycf1* (hypothetical chloroplast open reading frame 1) gene was analysed for Asteraceae species and observed a high number of SSRs (Simple Sequence Repeats) and a higher percentage of informative characters compared to the regions studied (*rbcL*, *matK* and *psbA-trnH*). For phylogenetic studies or low-level taxonomic DNA barcoding, this highly variable region was effective showing simple amplification and align due to its conserved reading structure. It is an unusual gene among plastid genes for DNA barcode or systematic molecular targets because of its length (5709 bp in *Nicotiana tabacum*), few sequences available and is incomplete or absent in some taxa but not a common loss (Neubig, Whitten et al. 2009, Curci, De Paola et al. 2015, Dong, Xu et al. 2015, Xu, Liu et al. 2015).

The marker *rps16-trnQ* showed the best discriminatory power on the variation of length, as well as the variation of sequence. Therefore, is suggested that *rps16-trnQ* could serve as a better barcode in orchids at the species level (Lin, Lin et al. 2015). The *rps7-trnV* segment was sequenced and genotyped among other markers for commercial teas authentication. The region was indicated as a suitable marker to identify possible contaminants, although not yet well represented in GenBank (De Castro, Comparone et al. 2017).

The availability of a large number of sequences was one of the requirements for the choice of regions analysed in this work. Now, we address about the availability of plant nucleotide sequences.

Available DNA sequences and databases

The amount of available genomic sequences has increased dramatically due to the fast advances in high-throughput DNA sequencing technologies (Peyachoknagul, Mongkolsiriwatana et al. 2014, Zeng, Zhou et al. 2017). This wealth of genomic data arising from plant genome sequencing projects reflects the growing awareness of the importance of plants as a resource for secure food production, and in bioenergy production pharmacology and other plant biotechnology applications (Lohse, Nagel et al. 2014). However, is a challenge organize such huge amount of data in an integrated, functional, and engaging way (Lai, Berkman et al. 2012, Sakai, Lee et al. 2013, Lohse, Nagel et al. 2014).

The chloroplast genome of the tobacco (*Nicotiana tabacum*) was the first to be sequenced. Thereafter, more than 800 complete chloroplast genomes and a multitude of partial sequences are available from the National Center for Biotechnology Information (NCBI), obtained from a wide variety of environmental samples. It may seem a significant number, but it is still unrepresentative in view of the number of existing plants species (Apweiler, Attwood et al. 2001, Abe, Inokuchi et al. 2014, Curci, De Paola et al. 2015, Yu, Dossa et al. 2017, Zeng, Zhou et al. 2017).

The accumulation of raw data led to the construction of public genomic databases, usually from independent initiatives. The information contained in the sequences is often fragmented, with some annotations, or only for a particular group or species (Apweiler, Attwood et al. 2001, Meyer, Nagel et al. 2005, Jung, Staton et al.

2007). The genome annotation is one of the most fundamental and indispensable steps, directly affecting further experiments (Numa and Itoh 2014). A lack of annotations can seriously harm and hinder the interpretation of sequence data. Identification of uncharacterized DNA sequences depends primarily on good reference database containing accurate, reliable and trustworthy genomic sequences with well-designed interfaces that allow selection, analysis, integration of information and the correct assignment of species (Sakai, Lee et al. 2013, Zhang, Chen et al. 2013). Databases are used as anchors in genetic mapping studies of other species, linking structural analysis with the functional genome (Meyer, Nagel et al. 2005). They also serve as tools for the development of molecular markers and studies of inter and intraspecific variability (Jung, Staton et al. 2007).

Although there is an overlap between available databases, the content of the repositories differs. It is therefore advisable to search all available repositories to ensure that the analysis performed to generate the data are as persistent as possible and to take advantage of the variety of search methods. The unbalanced representation of some species may strongly affect analysis (Attwood 2002, Hebert, Cywinska et al. 2003, Yang, Tang et al. 2013). A database can integrate multiple data from different sources, facilitating analysis through search and filtering processes (Carneiro, Resende et al. 2017). A way to group and organize the data visually and intuitively through multiple sequence alignment. A large number of aligned sequences allow for an in-depth evaluation of the universality of the genomic region (Attwood 2002, Taberlet, Coissac et al. 2007). Multiple sequence alignments provide an integral view of the conservation of sequences for each target region (Figure 4). The sequence alignments define homologous characters on which phylogenetic inferences are based (Veidenberg, Medlar et al. 2016).

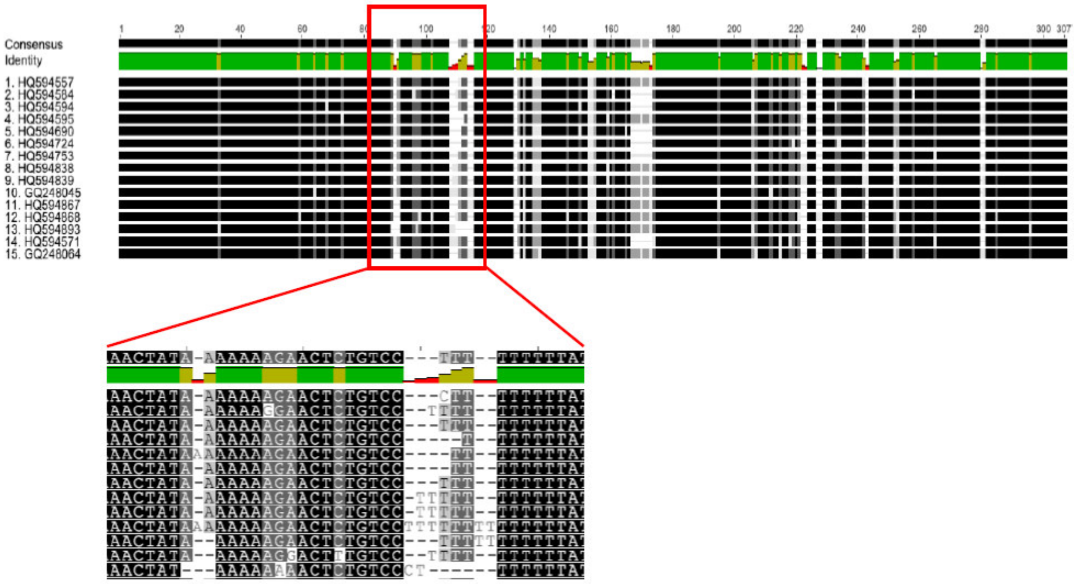

1
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300 307
Consensus
Identity
1. HQ594557
2. HQ594584
3. HQ594594
4. HQ594595
5. HQ594690
6. HQ594724
7. HQ594753
8. HQ594838
9. HQ594839
10. GQ248045
11. HQ594867
12. HQ594868
13. HQ594893
14. HQ594571
15. GQ248064
AACTATA-AAAAAAGAACTCTGTCC---TTT--TTTTTTAT
AACTATA-AAAAAAGAACTCTGTCC---CTT--TTTTTTAT
AACTATA-AAAAAGGAACTCTGTCC--TTTT--TTTTTTAT
AACTATA-AAAAAAGAACTCTGTCC---TTT--TTTTTTAT
AACTATA-AAAAAAGAACTCTGTCC-----T--TTTTTTAT
AACTATAAAAAAAAGAACTCTGTCC----TT--TTTTTTAT
AACTATA-AAAAAAGAACTCTGTCC----TT--TTTTTTAT
AACTATA-AAAAAAGAACTCTGTCC---TTT--TTTTTTAT
AACTATA-AAAAAAGAACTCTGTCC-TTTTT--TTTTTTAT
AACTATA-AAAAAAGAACTCTGTCC-TTTTT--TTTTTTAT
AACTATAAAAAAAAGAACTCTGTCCTTTTTTTTTTTTTTAT
AACTATA--AAAAAGAACTCTGTCC---TTTTTTTTTTTAT
AACTATA--AAAAAGGACTTTGTCC--TTTT--TTTTTTAT
AACTAT---AAAAAAAACTCTGTCCCT------TTTTTTAT

Figure 4. Example of a multiple sequences alignment. The green blocks represents identity, the conservation degree in that regions for all sequences present in the alignment. The black blocks are conserved regions. The amplified section of the alignment show the nucleotide bases variation in hypervariable regions (grey blocks).

Several databases are dedicated for a particular groups of species or single species. For example, a browser to display nucleotide sequence alignments, generic annotations, and single nucleotide polymorphisms (SNP) was used to comparatively analyse the rice genomes, to identifying the loss of genes from wild species to domestic, genes that may be related to the loss of recent cultivar characteristics as stress tolerance. The researchers also used plant families as preferred taxonomic rank to show how genes are conserved between plant species and how family genes evolve in each species (Sakai, Lee et al. 2013). In another example, the Oryzabase, is dedicated to rice (*Oryza sativa*) where anatomical and development descriptions are correlated with molecular genomic information like mutations and gene expression (Kurata and Yamazaki 2006). The AppleGFDB collects function, expression and annotated genes in the genome of apple (*Malus domestica*). These repositories can be used to access gene information of this important species (Zhang, Chen et al. 2013). The RadishBase, facility identification of possible genes associated with agriculturally important traits and understanding of important evolutionary process through the large-scale genome, expressed sequence tag (EST) sequences and high-density genetic maps of *Raphanus sativus* (Shen, Sun et al. 2012). The WheatGenome.info provides several web-based tools to analyze the wheat (*Triticum aestivum*) genome complex, allowing for genomic research that improves the production of this important cereal (Lai, Berkman et al. 2012).

Among other examples of plant sequence repositories, PoMaMo contains molecular maps of the chromosomes, putative gene functions and mutations information for analysis of potato (*Solanum tuberosum*), tomato (*Solanum lycopersicum*) and other related species of the family Solanaceae (Meyer, Nagel et al. 2005). The Genome Database for Rosaceae (GDR) combine physical, genetic and transcriptome maps, besides mutations and markers of the main species belonging to this group (Jung, Staton et al. 2007).

The Plant Microsatellite DNAs Database (PMD-Base), integrates a large number of genome microsatellites from most of the plant species grown or used as models for research and development (Yu, Dossa et al. 2017). The InterPro makes it possible to diagnose and document proteins from nucleotide sequences of unknown function (Apweiler, Attwood et al. 2001). The tRNA gene database (tRNADB-CE) which, in addition to several other genomes and sequences, provides analysis of 121 cpDNAs regarding tRNAs (Abe, Inokuchi et al. 2014). The Plant Long non-coded RNA Database (PLncDB) is an on-line repository that provides a complete genomic overview of RNAs long non-coding of *Arabidopsis* and can be used as a source of information for this genetic content for research in other plant species (Jin, Liu et al. 2013).

Many of these online repositories display the data in the phylogenetic level of the family because this category provides an adequate quantity of information that can be easily standardized and compared. Families with species of commercial interests are often analysed.

Laboratory procedures for DNA extraction

Variations in the growth and harvesting process, extraction and growth conditions, may also lead to failures in species identification and standardization of characterization techniques (Daniell, Lin et al. 2016, Mishra, Kumar et al. 2016). Many plant species produce secondary metabolites or bioactive substances such as alkaloids, flavonoids, tannins, cumarins, glycosides, phenylpropannes, organic acids, phenols, viscous polysaccharides, phytoalexins, terpenes and quinones which are used for plant protection and in food, pharmaceuticals, cosmetics and pesticides (Ma, Xie et al. 2010, da Cruz Cabral, Pinto et al. 2013, Barolo, Mostacero et al. 2014, Staats, Arulandhu et al. 2016). However, these same metabolic compounds are responsible for the reduce yield and in certain laboratory procedures, such as DNA extraction, amplification and cloning, among other analyses that can be done subsequently (Vassou, Kusuma et al. 2015, Kikkawa, Tsuge et al. 2016). For example, the plant *Taxus wallichiana* produces the secondary metabolite taxol and its precursors, which is known to inhibit the growth of some types of cancers. These compounds, when isolated together with DNA, inhibit PCR amplification (Khanuja, Shasany et al. 1999, Thomsen and Willerslev 2015).

An important step for the laboratory procedures in molecular genetics is the sample preparation and DNA extraction. The standardization and optimization of such procedures can be laborious because of the complexity and diversity of the matrices found (Cheng, Guo et al. 2003). In particular, low quality samples from processed or fragmented specimens, with little DNA or mixtures, pose a challenge for obtaining sufficient and quality material for subsequent analyses. Moreover, the reduction in the size of the fragments obtained, the lack of elimination of the potential inhibitory components and of the interfering substances of the material studied may compromise the results (Khanuja, Shasany et al. 1999, Parson, Pegoraro et al. 2000, Ronning, Rudi et al. 2005).

In plants, the biochemical composition of tissues may differ considerably, which complicate the obtaining of DNA and possibly related species require different extraction protocols (Dellaporta, Wood et al. 1983). With respect to available DNA extraction methods, commercial kits offer the advantage of standardization, being easily implemented in any laboratory, but often they yield low quality and quantity DNA. In this regard, specific protocols exist to improve DNA extraction efficiency. For example, the DNA extraction with cetyltrimethylammonium bromide detergent buffer (CTAB) combined with some purification step based on resin is widely used for diverse plants and derived products. This method does not require the use of expensive equipment or reagents (Rogers and Bendich 1985). The CTAB protocol was initially proposed by Murray and Thompson (1980), but widely used after the adaptations of Doyle (1987). Subsequently, on the basis of these publications and contributions such as Dellaporta, Wood et al. (1983), Rogers and Bendich (1985) and Gawel and Jarret (1991), the improvements proposed by the particular groups according to the material to be analysed. For example, certain reagents have been added to the process (b-mercaptoethanol helps remove polyphenols, NaCl solves the problem of high levels of polysaccharides) and other protocols can be used in a complementary manner (e.g. phenol-chloroform) (Cheng, Guo et al. 2003).

The CTAB method is extremely effective in recovering large amounts of total DNA from cells. A few fresh (or frozen) leaves, like 0.5 to 2 grams, can produces 20 to 100 micrograms of high molecular weight DNA, which represents quantity enough to perform subsequent analysis. However, isolation of genomic DNA from dry parts has been more difficult due to DNA degradation and the presence of unknown inhibitors (Yang, Tang et al. 2013, Yang, Li et al. 2014, Tang, Yukawa et al. 2015, Vassou, Kusuma et al. 2015).

Most protocols recommend extraction from fresh leaf tissue, but these material is not always available (Khanuja, Shasany et al. 1999). Despite the standardization efforts, the most appropriate DNA extraction method depends strongly on the matrix, and there is no universally accepted approach that allows simultaneously: (a) the recovery of large amounts of DNA; (b) from several parts of the plant; c) which can be used in diverse samples; and (d) ensuring the purity and cleanliness of isolated DNA for future processes. Therefore, the improvement of DNA isolation protocols is necessary (Staats, Arulandhu et al. 2016, De Castro, Comparone et al. 2017).

The main advantages of using total DNA extraction are to obtain sufficient DNA for analysis with little material; the flexibility, once that total DNA preparations can be used to study variations in all three genomes; and the adaptation, total DNA can be extracted from several groups of plants in which the current cpDNA extraction methods do not work). Researchers proposed targeted and standardized enrichment protocols for extraction using total DNA as template for cpDNA sequencing, this strategy could solve problems of cpDNA extraction of dry and degraded materials, but also simplify the extraction process (Cheng, Guo et al. 2003).

We have barely begun to explore cpDNA sequencing; two major reasons contribute to the current low numbers of completely sequenced chloroplast genomes. First, a large quantity of fresh leaves is needed for chloroplast DNA extraction. Second, it is difficult in many plants to isolate high-quality cpDNA, and considerable gaps were produced using low-quality cpDNA, which made it troublesome to assemble complete cpDNA. Owing to these difficulties, obtaining complete cpDNA sequences has been limited. These limitations severely restrict the extent to which investigators can analyse complete cpDNA data. A strategy for obtaining sufficient amounts of high quality, pure and complete chloroplast genome from a small number of fresh leaves and acquiring higher coverage of sequencing is urgently needed. The technologies involved in long-range PCR amplification and NGS methods make it possible to amplify whole cpDNA using several pairs of primers and then sequencing. Universal primers are the key for amplifying whole cpDNA of plants (Yang, Li et al. 2014).

Our research is justified because it offers effective tools for identifying species of the most important plant families. Plants is a taxonomic group in which there are several limits for attributing the correct assignment of the organisms, we have verified that the identification of plant species can be made by the variation in size of the nucleotide sequences of the chloroplast genome. We have designed a set of PCR-conserved primers that efficiently amplify the highly informative regions of the chloroplast DNA for major plant families. In addition to having a database that brings together in an accessible and intuitive way family-organized sequence alignments for the main genomic regions used in plant studies.

CHAPTER II
OBJECTIVES

OBJECTIVES

The high genetic diversity of plants is a challenge to the development of molecular methods and tools for population and species characterization. The main objective of this thesis is to provide new molecular and bioinformatics tools to study the most relevant families of plants.

Our specific objectives are:

1. Demonstrate that the identification of plant species can be achieved using variable length chloroplast DNA sequences. Our aim is to demonstrate the utility of the SPInDel concept for the identification of plants.

2. Design and evaluate the utility of universal PCR primers for amplification of informative chloroplast DNA regions in plants. The conserved genomic regions and PCR primers will be useful in diverse areas of plant research, including DNA barcoding, molecular ecology, metagenomics or phylogeny.

3. Construct a comprehensive on-line resource of curated nucleotide sequence alignments for plant research. The website will provides a complete, quality checked and regularly updated collection of alignments that can be used in taxonomic, molecular genetics, phylogenetic and evolutionary studies.

CHAPTER III

ORIGINAL RESEARCH

Study 1
Identification of plant species using variable length chloroplast DNA sequences

(*Accepted in Forensic Science International: Genetics, 2018*)

Identification of plant species using variable length chloroplast DNA sequences

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.

Keywords: cpDNA; plants; SPInDel; species identification; forensic botany.

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 (Coyle 2004, Zaya and Ashley 2012, Ogden and Linacre 2015, Bell, Burgess et al. 2016, Arenas, Pereira et al. 2017, Moreira, Carneiro et al. 2017). 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 (Woolfe and Primrose 2004, Pereira, Carneiro et al. 2008, Linacre and Tobe 2011). Such problems limit the standardization of results for inter and intra-laboratory comparisons. Among the available methods, DNA sequencing is currently the most used procedure. The 'DNA barcoding' has proved more difficult to use in plants than in animals (Pennisi 2007, Hollingsworth, Andra Clark et al. 2009, Ferri, Corradini et al. 2015). A few years ago, the Consortium for the Barcode of Life (CBoL) Plant Working Group (PWG) presented a final evaluation of seven candidate regions, recommending the use of a standard plant barcode comprising the combination of *rbcL* and *matK* (Pennisi 2007). According to the PWG reports, these combined loci identified 72% of all species (Chase and Fay 2009), which is still far from being a reliable identification system.

As an alternative to DNA sequencing, we have previously developed the SPInDel (Species Identification by Insertions/Deletions) method for molecular species identification (Pereira, Carneiro et al. 2010, Carneiro, Pereira et al. 2012). 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 (Figure 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 (Pereira, Carneiro et al. 2010, Carneiro, Pereira et al. 2012, Gonçalves, Marks et al. 2015, Alves, Pereira et al. 2017).

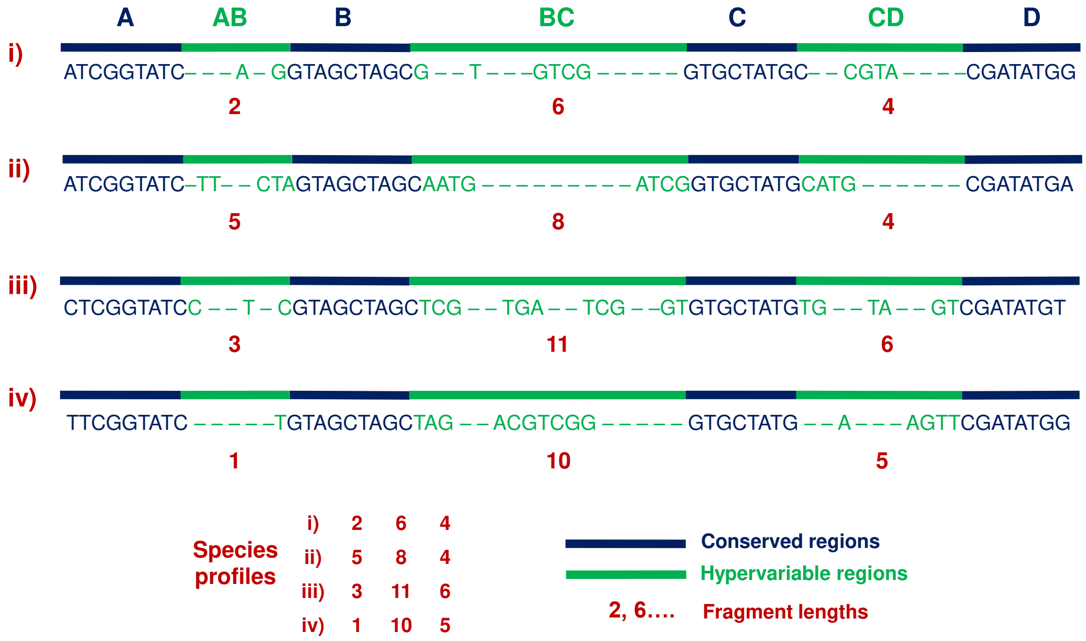

A
AB
B
BC
C
CD
D
i)
ATCGGTATC– – – A – GGTAGCTAGCG – – T – – –GTCG – – – – – GTGCTATGC– – CGTA – – – –CGATATGG
2
6
4
ii)
ATCGGTATC–TT– – CTAGTAGCTAGCAATG – – – – – – – – – ATCGGTGCTATGCATG – – – – – – CGATATGA
5
8
4
iii)
CTCGGTATCC – – T – CGTAGCTAGCTCG – – TGA – – TCG – –GTGTGCTATGTG – – TA – – GTCGATATGT
3
11
6
iv)
TTCGGTATC – – – – –TGTAGCTAGCTAG – – ACGTCGG – – – – – GTGCTATG – – A – – – AGTTCGATATGG
1
10
5
Species profiles
i) 2 6 4
ii) 5 8 4
iii) 3 11 6
iv) 1 10 5
Conserved regions
Hypervariable regions
2, 6…. Fragment lengths

Figure 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).

Our previous works have targeted the mitochondrial DNA (mtDNA) of animals, taking advantage of its relatively high mutation rate (Pereira, Carneiro et al. 2010, Gonçalves, Marks et al. 2015, Alves, Pereira et al. 2017). 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 (Wolfe, Li et al. 1987, Lynch, Koskella et al. 2006, Pereira, Carneiro et al. 2010). Therefore, researches have used the chloroplast DNA (cpDNA) for identification of plant species (Chase and Fay 2009, Ford, Ayres et al. 2009, Hollingsworth, Graham et al. 2011). 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 highly conserved regions suitable for the design of 'universal' primers and e) it is usually uniparentally inherited, and non-recombinant, making it effectively haploid (Olmstead and Palmer 1994, Shaw, Lickey et al. 2005, Santos and Pereira 2017).

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.

Materials and Methods

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 32kDa 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 (Edgar 2004) running in the Geneious version 5.5.8 (Kearse, Moir et al. 2012). 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).

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*	Number of families	Number of families with N≥10	Number of conserv ed regions	Number of hypervaria ble 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

* filtered - one per species, complete taxonomy and covering the region of interest

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 none or small variability at the sequence level that can be used as primer-binding sites (Figure 1). The SPInDel conserved regions were selected according to the criteria previously described (Pereira, Carneiro et al. 2010, Carneiro, Pereira et al. 2012). In the case of *trnL* (UAA), we used as conserved regions those named "C", "D", "G" and "H" by Taberlet *et al.* 2007 (Taberlet, Coissac et al. 2007). 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 Figure S1).

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 (Carneiro, Pereira et al. 2012) 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 the 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 alignments of different cpDNA regions in order to perform the diverse statistical analyses available on the SPInDel workbench. We started by using those species that were represented in the datasets of *atpF-atpH*, *psbA-trnH* and *trnL* CD. A total of 38 species were identified for the three regions. The sequence alignments of the three target regions were concatenated using the Geneious software (Figure 2a). We also concatenated the *atpF-atpH*, *psbA-trnH* and *trnL* GH regions using sequences from 170 species. The concatenated alignments were exported to the SPInDel workbench and analysed as previously described using the conserved regions defined for the individual regions. In these analyses, we have excluded the hypervariable regions defined by the peripheral conserved region of adjacent targets since they are not close to each other in the cpDNA. Therefore, the obtained profiles are only composed of the hypervariable regions inside each target region.

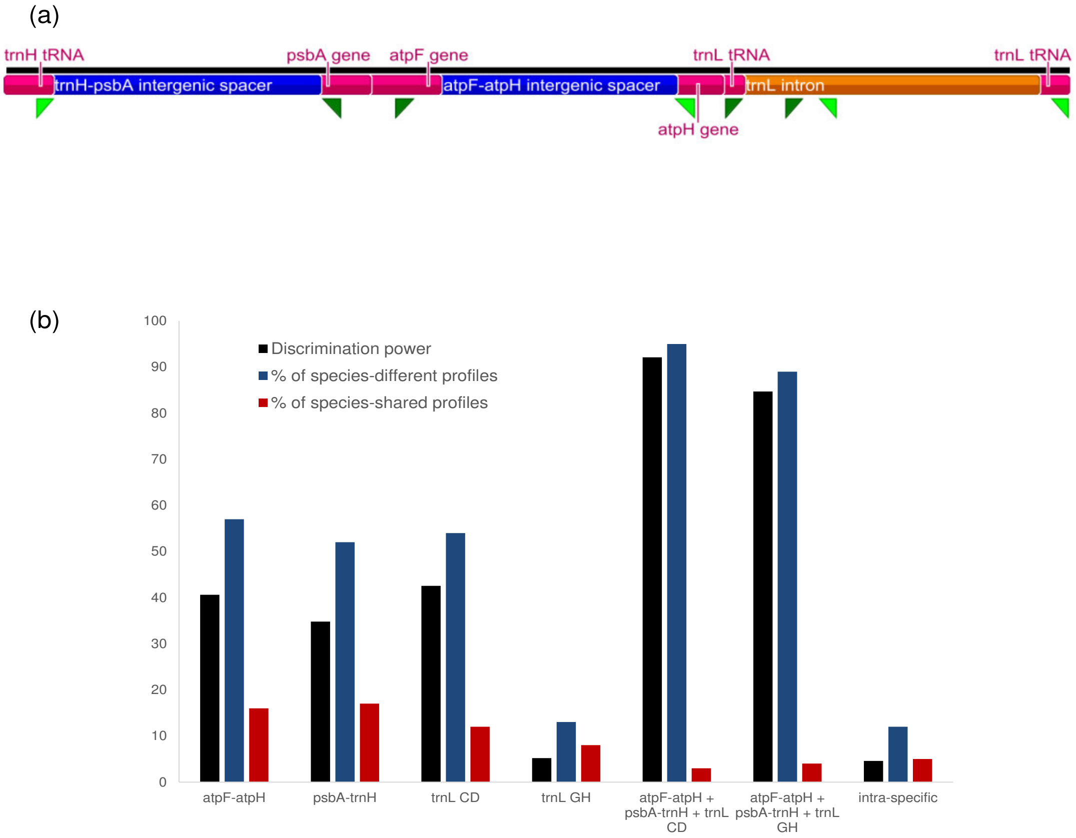


Figure 2. The cpDNA regions tested in the SPInDel approach. a) Schematic representation of the concatenated *atpF-atpH*, *psbA-trnH* and *trnL CD* cpDNA regions. Green arrows indicate the SPInDel conserved regions. b) The informative power of the cpDNA target regions considering the discrimination power (DP) and the percentage of species different and shared SPInDel profiles.

Results

atpF-atpH

We identified two SPInDel conserved regions in the *atpF* and *atpH* genes that delimitate the *atpF-atpH* spacer region (Supplementary Figure S1). We retrieved 2,360 sequences from the *atpF-atpH* target region, from which 1,317 (55.8%) were selected after removing redundant and incomplete sequences. These 1,317 sequences were organized in 156 families, from which 29 had 10 or more species (Table 1). These 29 families had a mean value of 33 sequences, with a minimum of 11 and maximum of 181 species (Table 2).

The potential use of SPInDel profiles for species identification purposes requires the existence of “species-specific SPInDel profiles”: those that are only found in one species within a taxonomic group and allow their unequivocal identification. The mean number of species-specific profiles (N_{sp}) was 9 (from 2 to 21), while the mean number of species with shared profiles ($N_{(species)\ sh}$) was 23 (varied from 0 to 160) (Table 2; Supplementary Table S3). Within each group, a species can present a profile that is unique or shared (i.e., common to more than one species). If all profiles were specific, N_{sp} will be equal to number of different profiles (N_{dp}). If some profiles are shared, then $N_{dp} > N_{sp}$. A profile can be shared between two or more species, therefore the number of species with shared profile (e.g., $N_{(species)\ sh}^{Araceae} = 22$) can be higher than the number of species-shared profiles (e.g., $N_{(profile)\ sh}^{Araceae} = 9$) (Supplementary Table S3).

The mean frequency of species-specific profiles (f_{sp}) was 0.41, ranging from 0.03 to 1 (Table 2). We also observed that the number of specific profiles was in general higher in families with fewer individuals (Figure 3a). For instance, the family Apiaceae had the lowest number of sequences (N=11) and the maximum frequency of species-specific profiles ($f_1^{Apiaceae}$=1). This result suggests that all species in this family had a unique combination of fragments lengths. On the other hand, families with a high N had usually a lower value of f_n^G, as observed in Poaceae with N=181 and $f_1^{Poaceae} = 0.12$. The lowest f_n^G value was observed in Zamiaceae with $f_1^{Zamiaceae}$ =0.03 and N=64 (Supplementary Table S3).

The family Apiaceae had an $N_{sh}^{Apiaceae}$=0, i.e. no species in this group had shared profile. Therefore, $N_{sp}^{Apiaceae}$ is equal to $N^{Apiaceae}$. However, it is important to consider the number of individuals in each group. The family Poaceae had a high $N_{(species)\ sh}$ due to the high N. Therefore, $N_{sp}^{Poaceae}$ 21 + $N_{(species)\ sh}^{Poaceae}$ 160 = $N^{Poaceae}$ 181 (Supplementary Table S3). The mean f_{sh} in the *atpF-atpH* spacer was 0.16, varying of 0 to 0.26 (Table 2). The families with the highest frequency of species-shared profile were Araceae with (f_{sh}=0.26; N=34) and Melanthiaceae (f_{sh}=0.26; N=19). Apiaceae was the only family with f_{sh}=0 (Supplementary Table S3). The mean frequency of different profiles (f_{dp}) was 0.57, ranging from 0.12 to 1 (Table 2). The family Zamiaceae had the lowest f_{dp} value, with 0.12. However, the family Apiaceae presented the maximal value (f_{dp} =1) indicating that all species had a different profile (Supplementary Table S3). The profiles from this target region presented an f_n^G= 0.9, ranging from 0.54 to 1 (Table 2). The family with the highest f_{sp} was Apiaceae (f_{sp}=1) (Supplementary Table S3). Most families presented values of $\bar{p}_n^G$ above 0.8, including those with a large number of individuals. For example, Poaceae with N=181 had $\bar{p}_n^G$=0.95, while Araucariaceae with N=15 had $\bar{p}_n^G$= 0.54 (Supplementary Table S3).

The mean discrimination power (DP), i.e., percentage of species that present a unique profile on a particular group, of *atpF-atpH* target region was 40.62%, ranging from 3.13% to 100% (Figure 2b and Table 2). The highest DP values were found in the family Apiaceae, where we are able to discriminate all species (DP=100%). On the other hand, Zamiaceae was the family with the lowest DP value (3.13%) (Supplementary Table S3). The DP increases with the increase in the frequency of different profiles (f_{dp}), as shown in Figure 3b.

psbA-trnH

Two SPInDel conserved regions were identified in the *psbA* and *trnH* genes that delimitate the *psbA-trnH* intergenic region (Supplementary Figure S1). We retrieved 14,550 sequences from the *psbA-trnH* spacer, from which 5,632 (38.7%) were selected for SPInDel analyses after removing redundant and incomplete sequences. These 5,632 sequences were organized in 327 families, from which 79 had 10 or more species (Table 1). These 79 families have a mean value of 64 species (Table 2).

The mean f_{sp} was 0.35, ranging from 0 to 1.00 (Table 2). The Hymenophyllaceae family with N=14 had $f_1^{\text{Hymenophyllaceae}}$ =1, meaning that all species of this group have a unique profile. Ephedraceae had no species-specific profile ($f_1^{\text{Ephedraceae}}$ =0) (Supplementary Table S4). The mean N $_{(species)\ sh}$ was 50, ranging from 0 to 412 (Table 2). The Poaceae family had the highest value ($N_{(species)\ sh}$ =412), while Hymenophyllaceae had the lowest value for this parameter ($N_{(species)\ sh}$ =0) (Supplementary Table S4). The mean f_{sh} in the *psbA-trnH* spacer was 0.17 with a minimum of 0 and maximum of 0.32 (Table 2). The families with the highest number of sequences (N^{Poaceae} =425 and $N^{\text{Orchidaceae}}$ =406) showed low f_{sh} (f_{sh}^{Poaceae} =0.08 and $f_{sh}^{\text{Orchidaceae}}$ =0.10) (Supplementary Table S4). The mean f_{dp} across the all families was 0.52, varying from 0.1 to 1.0 (Table 2). The f_n^G was 0.89, ranging from 0.27 to 1.0 (Table 2). The family with the lowest average (f_n^G =0.27), was Cyperaceae, although most families had values near to the possible maximum (Supplementary Table S4). The mean DP for of the *psbA-trnH* target region was 34.80% (Figures 2b and 3 and Table 2). The best results can be found in Hymenophyllaceae (DP=100%), since we are able to discriminate all species. On the other hand, no species were discriminated in Ephedraceae (DP=0%) (Supplementary Table S4).

trnL CD

We identified four SPInDel conserved regions in the *trnL* (UAA) intron target region (Supplementary Figure S1). We recovered 4,083 sequences from the *trnL* CD target region, from which 2,714 (66.5%) were selected for SPInDel analyses after removing redundant and incomplete sequences. These 2,714 species were organized in 117 families, from which 44 had 10 or more sequences (Table 1). The mean number of species per family was 57, varying of 10 to 397 sequences (Table 2). The families with the highest number of species were Poaceae (N=397) and Rubiaceae (N=335) (Supplementary Table S5). The mean f_{sp} was 0.43, ranging of 0.06 to 1.0 (Table 2).

The mean f_{sh} was 0.12 with a maximum value of 0.26 and minimum of 0 (Table 2). In families with a high number of species, the f_{sh} was low. For example, $f_{sh}^{Poaceae}$ =0.12 and $f_{sh}^{Rubiaceae}$ =0.09. The minimum value of f_{sh} was found in families Ericaceae, Goodeniaceae and Saxifragaceae, with f_{sh}=0. The maximum f_{sh} value was found in the family Theaceae with $f_{sh}^{Theaceae}$ = 0.26 (Supplementary Table S5). The mean f_{dp} was 0.54, varying of 0.1 to 1.0 (Table 2). Brassicaceae had the lowest frequency of species-different profile ($f_{dp}^{Brassicaceae}$ = 0.10). On the other extreme, Ericaceae had $f_{dp}^{Ericaceae}$ =1 (Supplementary Table 5). The mean $N_{(species)\ sh}$ was 43, ranging of 0 to 336 (Table 2). The Poaceae had the highest $N_{(species)\ sh}$ (N_{sh} = 336). One of the lowest values was found in Gnetaceae, with $N_{(species)\ sh}$ = 2 (Supplementary Table S5). The f_n^G was 1.55 (range 0.11 to 2.6; Table 2). The maximum value in this target region reached three, because *trnL* CD has three hypervariable regions (C-G, G-H and H-D) (Supplementary Figure S1). The highest values of f_n^G were observed in families with less species (Supplementary Table S5). The mean DP for *trnL* CD was 42.54% ranging from 5.56% to 100% (Figures 2b and 3 and Table 2). The best results were found in the Ericaceae, Goodeniaceae and Saxifragaceae families, where all species were discriminated. The Amaryllidaceae family had the lowest DP value (5.56%) (Supplementary Table S5).

trnL GH

We identify two SPInDel conserved regions in the *trnL* (UAA) intron. These conserved regions delimited a shorter segment located inside the *trnL* CD spacer, defined by regions G and H (*trnL* GH) (Supplementary Figure S1). We retrieved 54,494 sequences from the *trnL* GH target region from which 35,198 (64.6%) were selected for SPInDel analyses after removing redundant and incomplete sequences. These selected sequences were organized in 351 families, from which 173 had 10 or more sequences (Table 1). The target regions *trnL* GH and *psbA-trnH* had a similar number of families, although *psbA-trnH* had fewer species per family. The mean number of species in *trnL* GH was N=200, while the mean was N=64 in *psbA-trnH* (Table 2). The 173 families had a mean value of 200 species, with the maximum value reached in Fabaceae (N=2,599) and in Poaceae (N=2,078) (Supplementary Table S6).

The mean f_{sp} across all families was 0.05, with a minimum of 0 and maximum of 0.47 (Table 2). This cpDNA region is shorter and less informative that *atpF-atpH, psbA-trnH* and *trnL* CD. No family had all species with a unique profile (i.e., f_{sp} was always <1).

Therefore, there was no family with $N_{(species)\ sh}$ = 0 (Table 2). Most families showed a f_{sp} smaller than 0.2. The family with the highest f_{sp} was Clusiaceae, with $f_1^{Clusiaceae}$ = 0.47 (Supplementary Table S6). The mean f_{sh} was 0.08, varying of 0 to 0.27 (Table 2). The family Plumbaginaceae had the maximum frequency of species-shared profile, f_{sh}=0.27. The families Sapindaceae and Colchicaceae had no species with specific profiles, having only shared profiles. As previously shown, the highest frequencies of species-shared profiles was reached in families with the lowest number of sequences (Supplementary Table S6). The mean f_{dp} was 0.13, ranging of 0.01 to 0.71 (Table 2). The family Clusiaceae had the highest frequency of species-different profiles (f_{dp}=0.71). On the other hand, the families Araliaceae, Asteraceae, Brassicaceae, Bromeliaceae, Fabaceae, Poaceae and Sapindaceae had the lower frequency of species-different profiles, with f_{dp}=0.01 (Supplementary Table S6). The mean $N_{(species)\ sh}$ was 196, with a minimum of 8 and a maximum 2593 (Table 2). The families with the lowest number of species with shared profiles were Coriaceae and Chrysobalanaceae ($N_{(species)\ sh}$ = 8) (Supplementary Table S6).

The f_n^G was 0.53, with a maximum value of 0.97 (Table 2). No family yielded a value of 1 (i.e., all species were different), because there was always some species with equal profiles (no family had N_{sp}=1) (Supplementary Table S6). The families with low values of f_{sp} showed diverse values of f_n^G. However, all families with a f_{sp} above 0.1 had an higher than 0.35. For instance, the family Clusiaceae had f_{sp} =0.47 and f_n^G =0.96 (Supplementary Table S6). The mean DP for the families of *trnL* GH target region was 5.18% ranging from 0% to 47.06% (Figures 2b and 3 and Table 2). Clusiaceae had the highest DP (47.06%). In 21 families, no species had a unique profile (DP=0%) (Supplementary Table S6).

Table 2. Main SPInDel analyses performed for each cpDNA target region.

	Mean (min - max)									
Genomic Region	Number of species per family (N)	Number of species-specific profiles (N_{sp})	Frequency of species-specific profiles (f_{sp})	Number of species with shared profiles ($N_{(species)\ sh}$)	Number of species-shared profiles ($N_{(profile)\ sh}$)	Frequency of species-shared profiles (f_{sh})	Number of species-different profiles (N_{dp})	Frequency of species-different profiles (f_{dp})	Average number of pairwise differences (f_n^G)	Discrimination power (%)
atpF-atpH	33	9	0.41	23	5	0.16	14	0.57	0.90	40.62
	(11 - 181)	(2 - 21)	(0.03 – 1.00)	(0 - 160)	(0 - 25)	(0.00 - 0.26)	(4 - 46)	(0.12 - 1)	(0.54 - 1)	(3.13 - 100.00)
psbA-trnH	64	14	0.35	50	10	0.17	24	0.52	0.89	34.80
	(10 - 425)	(0 - 75)	(0.00 - 1.00)	(0 - 412)	(0 - 66)	(0.00 - 0.32)	(2 - 141)	(0.10 - 1)	(0.27 - 1)	(0.00 - 100.00)
***trnL* CD**	57	14	0.43	43	6	0.12	21	0.54	1.55	42.54
	(10 - 397)	(1 - 61)	(0.06 – 1.00)	(0 - 336)	(0 - 48)	(0.00 - 0.26)	(2 - 109)	(0.10 - 1)	(0.11 - 2.6)	(5.56 - 100.00)
***trnL* GH**	200	3	0.05	196	7	0.08	11	0.13	0.53	5.18
	(10 - 2599)	(0 - 26)	(0.00 - 0.47)	(8 - 2593)	(1 - 43)	(0.00 - 0.27)	(1 - 69)	(0.01 - 0.71)	(0 - 0.97)	(0.00 - 47.06)

Intra-specific SPInDel diversity

The effectiveness of the SPInDel concept depends upon the existence of low intraspecific variation (Pereira, Carneiro et al. 2010). We analysed 14 intra-species datasets representing the species with the largest number of sequences available in GenBank (Supplementary Table S2). The mean number of individuals per species was 87, with the highest number of sequences obtained for the *Acer rubrum* species (N=261) of the Aceraceae family (Figure 3a and Table 3). The mean f_{sp} in all target regions was 0.07, with most species presenting low values: $f_1^{Onobrychis\ viciifolia}$ = 0.01, $f_1^{Ranunculus\ kuepferi}$ = 0.01 and $f_1^{Potentilla\ argentea}$ =0.01, meaning that the most profiles were equal inside each species (Table 3).

Individuals from *Justicia adhatoda* (a), *Lepidium montanum* (g), *Boechera holboelli* (l) and *Carapichea ipecacuanha* (n) had no differences among them, with f_n^G =0 (Supplementary Figure S2). Values of f_n^G lower than 0.11 were observed in *Musa acuminate (atpF-atpH)*, *Phalaris arundinaceae* (*psbA-trnH*) and *Poa annua* (*trnL* CD)*,* which indicates that all profiles from the same species diverge by a small number of differences. However, $\bar{p}_3^{Silene\ latifolia}$= 0.73 and $\bar{p}\,_3^{Ficus\ carica}$= 1.32, suggesting that there are divergent hypervariable region in these species.

From the 261 *A. rubrum* cpDNA sequences (N=261), only three individuals (f_{sp}=0.01%) had unique profiles (N_{sp}=3), therefore the number of species with shared profile was $N_{(species)\ sh}$=258. There was no individual with a unique profile (N_{sp}=0) in the 167 *Populus balsamifera* sequences (N=167), i.e., all individuals shared profiles ($N_{(species)\ sh}$ = 167). When considering the target region *psbA-trnH,* the family Poaceae had an $\bar{p}_1^{Poaceae}$ =0.95 (Supplementary Table S4), while the representative species from Poaceae had in *psbA-trnH* $\bar{p}_1^{Phalaris\ arundinaceae}$ = 0.11 and in *trnL* CD $\bar{p}_1^{Poa\ annua}$=0.08. For the *trnL* GH region, the Rubiaceae family had $\bar{p}_1^{Rubiaceae}$=0.75, while the representative species of this family had $\bar{p}_1^{Carapichea\ ipecacuanha)}$ = 0. The lowest values for the f_{sp} were observed for the *trnL* GH (mean f_{sp}=0.00) (Table 3).

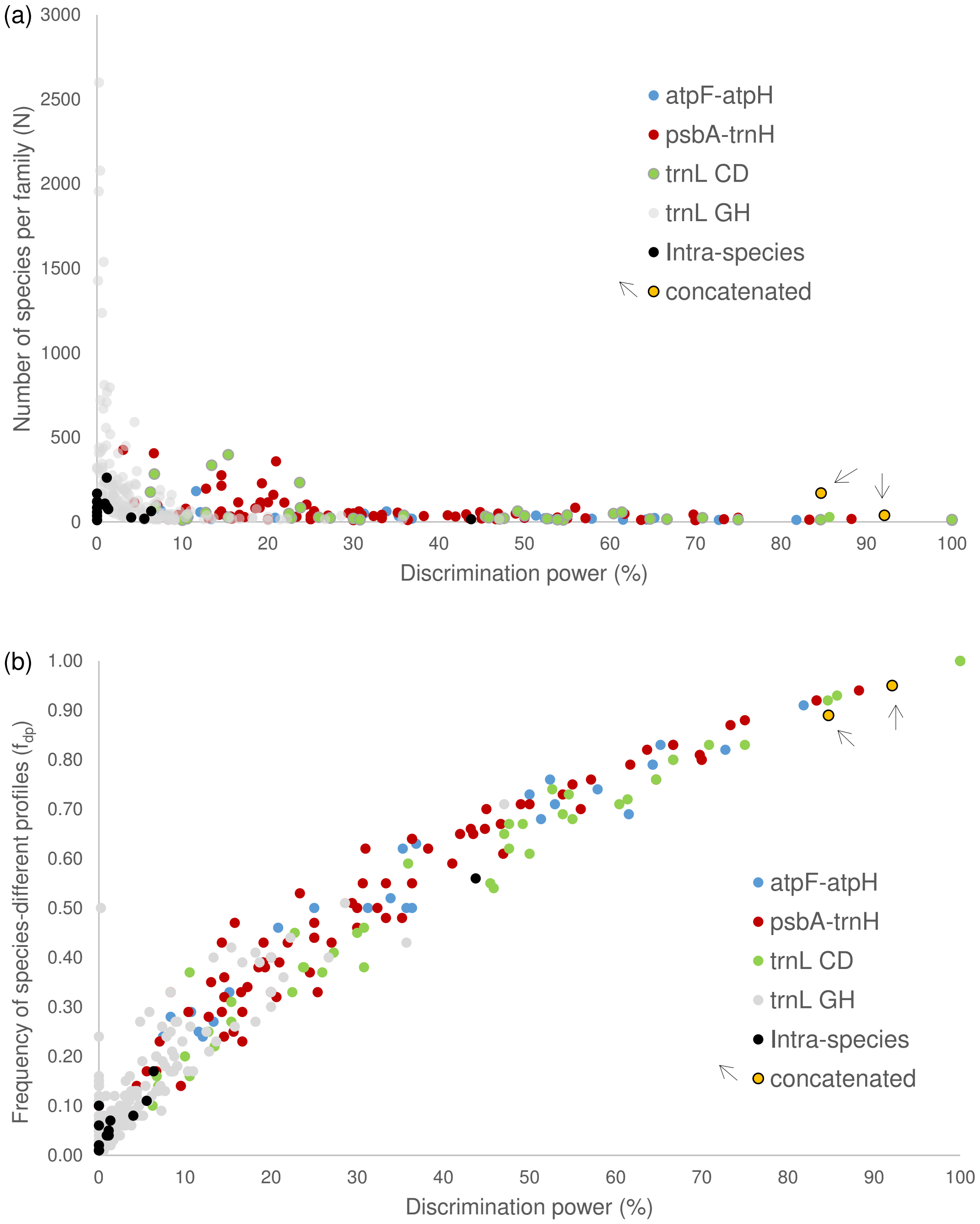


Figure 3. The Discrimination Power (DP) of the SPInDel approach in plant families. Variation of DP values considering a) the number of species in each family and b) frequency of species-different profiles (f_{dp}).

Table 3. Intraspecific SPInDel statistics for 14 species.

Genomic Region	Species (Family)	Number of sequences per family (N)	Number of species-specific profiles (N_{sp})	Frequency of species-specific profiles (f_{sp})	Number of species with shared profiles ($N_{(species)\ sh}$)	Number of species-shared profiles ($N_{(profile)\ sh}$)	Frequency of species-shared profiles (f_{sh})	Number of species-different profiles (N_{dp})	Frequency of species-different profiles (f_{dp})	Average number of pairwise differences ($\bar{p}_n^G$)	Discrimination power (%)
atpF-atpH	*Justicia adhatoda* (Acanthaceae)	10	0	0.00	10	1	0.10	1	0.10	0.00	0.00
atpF-atpH	*Musa acuminata* (Musaceae)	18	1	0.06	17	1	0.06	2	0.11	0.11	5.56
	Mean	14	1	0.03	14	1	0.08	2	0.11	0.06	2.78
psbA-trnH	*Acer rubrum* (Aceraceae)	261	3	0.01	258	8	0.03	11	0.04	0.74	1.15
psbA-trnH	*Onobrychis viciifolia* (Fabaceae)	87	1	0.01	86	3	0.03	4	0.05	0.51	1.15
psbA-trnH	*Phalaris arundinaceae* (Poaceae)	35	0	0.00	35	2	0.06	2	0.06	0.11	0.00
psbA-trnH	*Potentilla argentea* (Rosaceae)	75	1	0.01	74	4	0.05	5	0.07	0.70	1.33
	Mean	115	1	0.01	113	4	0.04	6	0.06	0.50	0.86
trnL CD	*Lepidium montanum* (Brassicaceae)	57	0	0.00	57	1	0.02	1	0.02	0.00	0.00
trnL CD	*Silene latifolia* (Caryophyllaceae)	63	4	0.06	59	7	0.11	11	0.17	0.73	6.35
trnL CD	*Ficus carica* (Moraceae)	16	7	0.44	9	2	0.12	9	0.56	1.32	43.75
trnL CD	*Poa annua* (Poaceae)	25	1	0.04	24	1	0.04	2	0.08	0.08	4.00
	Mean	40	3	0.14	37	3	0.07	6	0.21	0.53	13.53
trnL GH	*Boechera holboelli* (Brassicaceae)	84	0	0.00	84	1	0.01	1	0.01	0.00	0.00
trnL GH	*Ranunculus kuepferi* (Ranunculaceae)	108	1	0.01	107	3	0.03	4	0.04	0.24	0.93
trnL GH	*Carapichea ipecacuanha* (Rubiaceae)	119	0	0.00	119	1	0.01	1	0.01	0.00	0.00
trnL GH	*Populus balsamifera* (Salicaceae)	167	0	0.00	167	2	0.01	2	0.01	0.35	0.00
	Mean	120	0	0.00	119	2	0.02	2	0.02	0.15	0.23
All target regions	Minimum	10	0	0	9	1	0.01	1	0.01	0	0.00
	Maximum	261	7	0.44	258	8	0.12	11	0.56	1.32	43.75
	Mean	87	1.63	0.07	85.81	2.88	0.05	4.25	0.12	0.39	4.59

Concatenated *atpF-atpH* + *psbA-trnH* + *trnL* CD regions

We concatenated the *atpF-atpH*, *psbA-trnH* and *trnL* CD targets from 38 species that had available sequences in GenBank for the three genomic regions (Figure 2a). The species and length of each target region is described in Supplementary Table S7. The merging of these three regions allows the identification of eight SPInDel conserved regions. We then selected five hypervariable regions for the SPInDel analyses: *atpF* F - *atpH* R*; psbA* F - *trnH* R*; trnL* CG*, trnL* GH and *trnL* HD (Figure 2a and Table 4).

When considering each target region alone, we observed that the *atpF* F - *atpH* R and *psbA* F - *trnH* R regions has the highest diversity (different lengths), with 27 different length out of 38 in *atpF-atpH* and 29 different length out of 38 in *psbA-trnH.* The hypervariable region *trnL* CG was the less informative with only six different lengths. The *trnL* GH had 16 and *trnL* HD 24 different length (Supplementary Table S7). The profiles that result from the combination of the length of the five cpDNA regions were unique in all species, with exception of three species from the same genus *(Picea abies, P. jezoensis* and *P. koraiensis*) (Supplementary Table S7). For this reason, the number of species-specific SPInDel profile in the concatenated alignment was 35, while the total number of different profiles (N_{dp}) was 36 (Table 4). In any case, different lengths were obtained for all species representing 25 genera, such as *Hordeum bulbosum, H. pusillum* and *H. vulgare; Poa annua* and *P. compressa; Silene latifolia* and *S. vulgaris; Viola dissecta, V. albida* and *V. chaerophylloide* (Supplementary Table S7). Overall, it is possible to discriminate 35 species in a total of 38 through of the combination of these five hypervariable regions. Moreover, the maximum frequency of species-specific profile of the concatenation (f_{sp}=0.92) is reached with the use of only three hypervariable regions (*atpF* F–*atpH* R, *psbA* F–*trnH* R and *trnL* HD) (Figure 4a).

The average number of pairwise differences for the concatenated regions was $\bar{p}_5^G$ = 4.55 (Table 4), a value close to the maximum that can be obtained with five hypervariable regions. A total of 462 pairwise comparisons (66% of the total combinations) yielded differences in the five hypervariable regions, while 200 cases (28%) had four differences. Only 41 cases were different by less than four hypervariable regions (Figure 4b). Figure 4c shows the 'region by region' analysis for the concatenate *atpF-atpH* + *psbA-trnH* + *trnL* CD*.* The regions *psbA-trnH* and *atpF-atpH* had the highest average pairwise differences, both with p = 0.98. The DP of this concatenated set was 92.1% (Figures 2b and 3 and Table 4).

Table 4. Diverse SPInDel statistics for the concatenated cpDNA regions.

Concatenated	**Number of sequences in the project (N)**	**Conserved regions**	**Hypervariable regions (n)**	**Number of species-specific profiles (N_{sp})**	**Frequency of species-specific profiles (f_{sp})**	**Number of species with shared profiles (N $_{(species)}$ *sh*)**	**Number of species-shared profiles (N $_{(profile)}$ *sh*)**	**Frequency of species-shared profiles (f_{sh})**	**Number of species-different profiles (N_{dp})**	**Frequency of species-different profiles (f_{dp})**	**Average number of pairwise differences ($\bar{p}_5^G$)**	**Discrimination Power (%)**
atpF-atpH* + *psbA-trnH* + *trnL CD	38	8	5	35	0.92	3	1	0.03	36	0.95	4.55	92.1
atpF-atpH* + *psbA-trnH* + *trnL GH	170	6	3	144	0.85	26	7	0.04	151	0.89	2.88	84.7

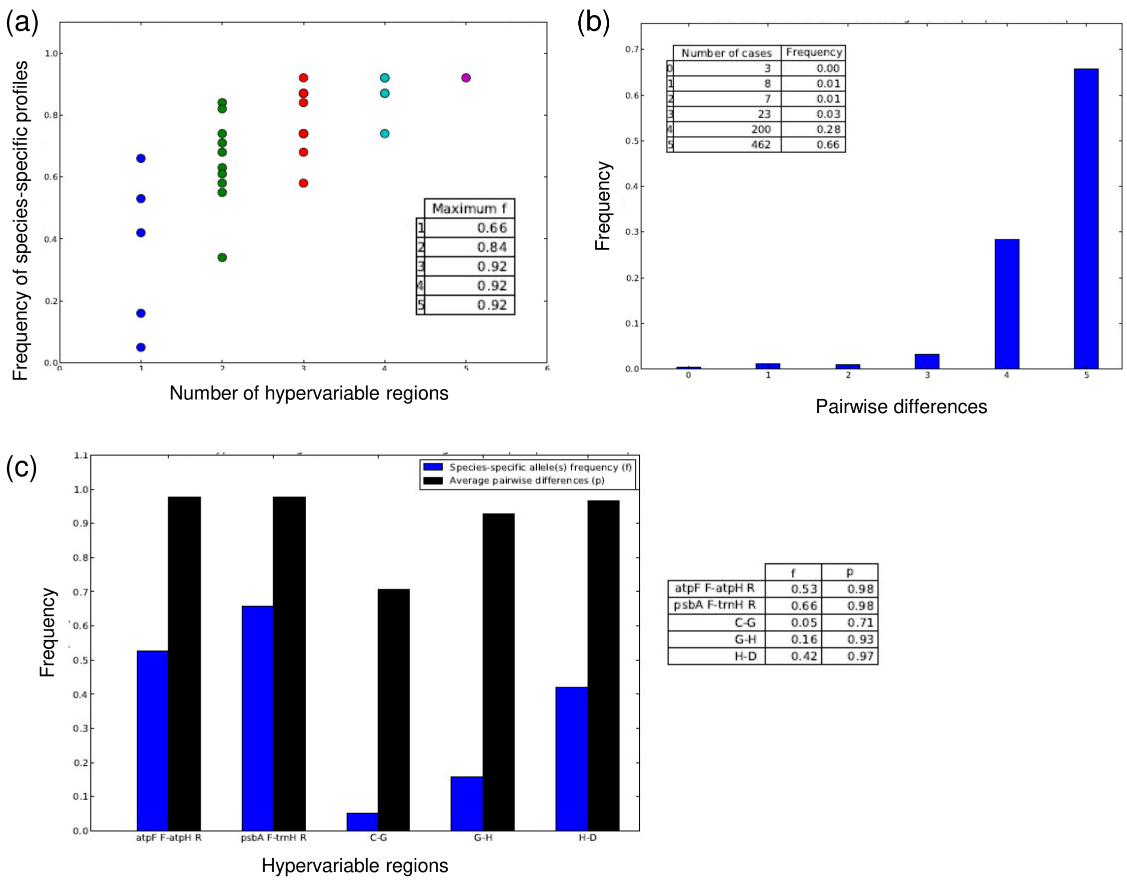


Figure 4. SPInDel analysis of concatenated *atpF-atpH* + *psbA-trnH* + *trnL* CD regions. a) The frequency of species-specific profile in all combinations of hypervariable regions. b) Mismatch distribution. i.e. the frequency distribution of the number of SPInDel hypervariable regions that differ between all pairs of SPInDel profiles in a taxonomic group. c) The discriminatory potential of each hypervariable region individually (region by region analyses).

Concatenated *atpF-atpH* + *psbA-trnH* + *trnL* GH regions

We also concatenated the *atpF-atpH, psbA-trnH* and *trnL* GH regions from 170 species that had available sequences in GenBank. The merging of these three genomic regions allowed the definition of six SPInDel conserved regions (*atpF*, *atpH*, *psbA*, *trnH*, *trnL* G and *trnL* H) (Figure 2a and Table 4). Three hypervariable regions (*atpF* F - *aptH* R, *psbA* F-*trnH* R and *trnL* GH) were selected for the SPInDel analyses. The *atpF-atpH* and *psbA-trnH* regions had the highest number of different sequence lengths, with 94 different lengths out of 170 in *atpF-atpH* and 97 different lengths out of 170 in *psbA-trnH.* The hypervariable region *trnL* GH was the less informative with 29 different sequence lengths (Supplementary Table S8).

The frequency of species-specific allele(s) for *atpF-atpH* (f_{sp}=0.36) and *psbA-trnH* (f_{sp}=0.40) was higher than for *trnL* GH (f_{sp}=0.04), in line with their high variability (*atpF-atpH* and *psbA-trnH* with $\bar{p}_n^G$ = 0.98, while *trnL* GH had $\bar{p}_n^G$ = 0.92; Figure 5). We then considered the profiles that result from the combination of the sequence length of the three cpDNA regions. There were 26 species with shared profiles (N $_{(species)\ sh}$ = 26) (Table 4). Twenty-two out of 26 species that had equal SPInDel profiles are from the same genus, e.g. *Lolium multiflorum* and *L. perenne*; *Thypa angustifolia* and *T. latifolia.* Only 4 species with equal profiles are from different genera: *Zoysia japonica* and *Arundinella hirta; Phyllostachys nigra var. henonis and Sasa palmata* (Supplementary Table S8).

In most cases, different numeric profiles were obtained for species from the same genus, such as *Passiflora incarnata* and *P. quadrangularis; Poa annua* and *P. compressa; Solanum dulcamara, S. lycopersicum* and *S. nigrum.* Different profiles were also were obtained for the 23 species of the genus *Viola*, eight of the genus *Ficus* and five of the genus *Acer* (Supplementary Table S8).

It is possible to discriminate 144 species in a total of 170 using these three target regions. Therefore, N$_{sp}$ in the concatenated alignment was 144, N $_{(profile)\ sh}$ was 7 (then N$_{dp}$ = 151). The average number of pairwise differences was $\bar{p}_3^G$ = 2.88 (Table 4), a value close to the maximum that can be obtained with three hypervariable regions (n = 3). The high discriminatory capacity of the SPInDel approach is clearly seen in the histograms representing the mismatch distribution (Figure 5a). A total of 12,988 pairwise comparisons (90% of the total combinations) yielded differences in the three hypervariable regions, while 1120 cases (8%) had two differences. Only 257 cases (2%) were different by one or none hypervariable regions (Figure 5b). The discriminatory potential of hypervariable regions *psbA-trnH* and *atpF-atpH* was higher than *trnL* GH (Figure 5c). The DP of this concatenated set was 84.7% (Figures 2b and 3 and Table 4).

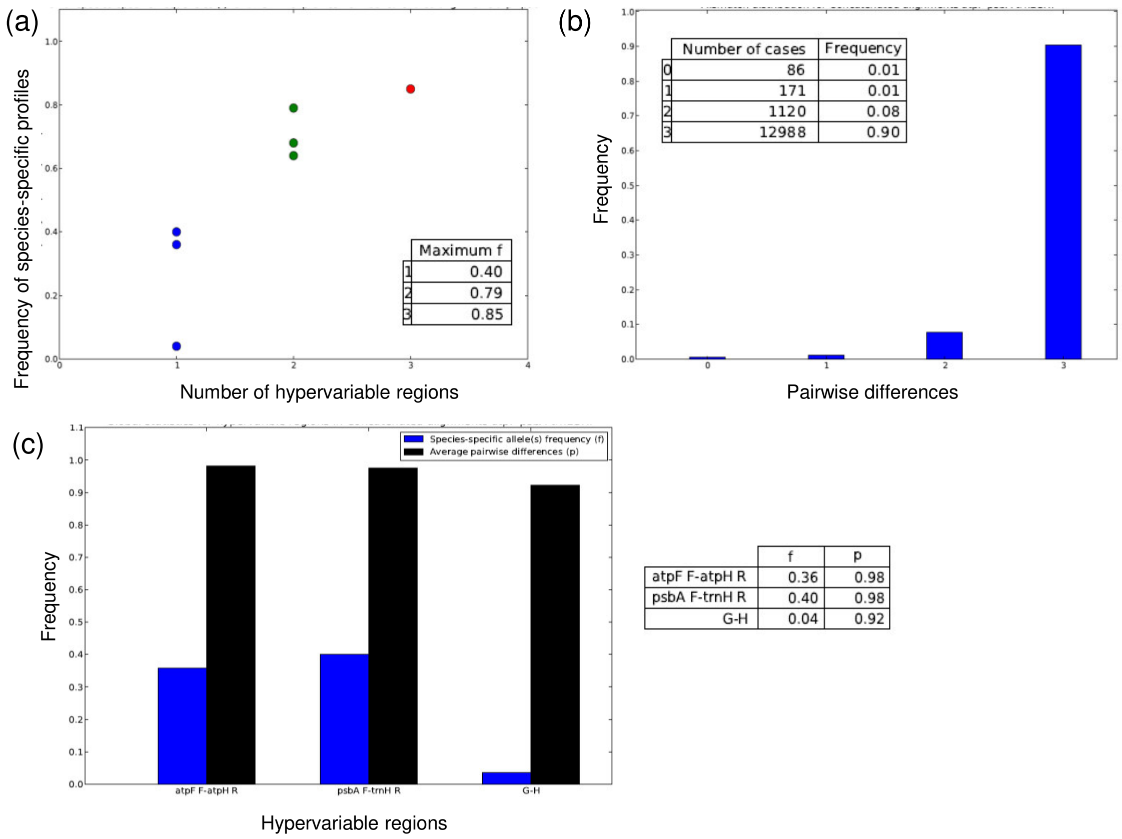


Figure 5. SPInDel analysis of concatenated *atpF-atpH* + *psbA-trnH* + *trnL* GH regions. a) The frequency of species-specific profile in all combinations of hypervariable regions. b) Mismatch distribution. i.e. the frequency distribution of the number of SPInDel hypervariable regions that differ between all pairs of SPInDel profiles in a taxonomic group. c) The discriminatory potential of each hypervariable region individually (region by region analyses).

Discussion

It has been suggested that the use of cpDNA for broad taxonomic identifications is constrained by the prevalence of indels that greatly complicate sequence alignments (Graham, Reeves et al. 2000, Kelchner 2000, Yamane, Yano et al. 2006, Ford, Ayres et al. 2009). The presence of indels is often regarded as a problem for DNA sequencing and indel-rich regions have been avoided for species identification purposes. However, the SPInDel concept for biological identification circumvents this apparent limitation by using cpDNA in a different manner: conserved regions are used to define variable segments in which a combination of sequence lengths (caused by indels) is characteristic of each species (Figure 1). The pattern of interspersed conserved and hypervariable regions is common in the cpDNA of plant species with the coding region being often very conserved, while the non-transcribed regions shows usually extensive sequence divergence and length heterogeneity (Xiong, Peng et al. 2009, Green 2011).

One of these non-transcribed spacer regions, the chloroplast *trnL* (UAA) intron, is known for its potential as species-specific marker due to low intra- and higher inter-specific genetic variation (Wallinger, Juen et al. 2012). This region has a conserved secondary structure with alternation of conserved and variable regions. Consequently, the alignment of diverse *trnL* intron sequences might allow the design of primers in conserved regions to amplify the short variable region in between (Taberlet, Coissac et al. 2007), which is suitable for the SPInDel concept. However, our results show that *trnL* does not represent the most variable non-coding region of chloroplast DNA (Figure 3, Table 2 and Supplementary Table S7). The main drawback of *trnL* (UAA) intron is the relative low resolution compared with other non-coding cpDNA regions, which is more evident for the short G-H segment. For instance, the discriminatory capacity of *trnL* GH was 5.18% and the mean f_{sp} across families was 0.05, while the discriminatory capacity of *atpF-atpH* was 40.62% and the mean f_{sp} was 0.41 (Figure 2b and Table 2). The levels of diversity in *trnL* CD are higher than in *trnL* GH mainly because *trnL* CD has three hypervariable regions (all other targets have only one hypervariable region). The low resolution of *trnL* GH is associated with a low intraspecific variation (Table 3).

The *psbA-trnH* target region was one of the first chloroplast locus to be suggested as a universal DNA barcode in plants (Kress and Erickson 2007, Yao, Song et al. 2009). This intergenic spacer is one of the most variable regions of the plastid genome and much of it is variability occurs as indels, exhibiting considerable variation in size. The *psbA-trnH* intergenic spacer is relatively short (~200-500bp) and has been recommended for species identification and phylogenetic studies as it evolves comparatively rapidly, offers useful levels of interspecific variation in nucleotide sequence and enables design of universal primers (Kress, Wurdack et al. 2005, Ford, Ayres et al. 2009, Pang, Luo et al. 2012). We found that the *psbA-trnH* length variation was sufficient to discriminate several species (Supplementary Tables S7 and S8). Moreover, interspecific analysis at the *psbA-trnH* in the family Poaceae was $\bar{p}_1^{Poaceae}$ =0.95 (Supplementary Table S4), suggesting that it is suitable for accurate species identifications. Similarly, the intraspecific diversity in a Poaceae species (*Phalaris arundinaceae*) was low ($\bar{p}_1^{G}$ =0.11), suggesting the existence of a gap between intra- and inter-species divergence (Table 3).

We found that the *atpF-atpH* target region has a moderate discriminatory power by length variability, with a mean f_{sp} across families of 0.35 (Table 2). The *atpF-atpH* was one of the intergenic spacers proposed as plant barcoding regions at the second international Barcode of Life Conference (Hollingsworth, Graham et al. 2011), often having a high interspecific diversity (Lahaye 2008). When considering length variation, we also found that *atpF-atpH* is moderately variable, with a mean f_{sp} across families of 0.41 and a discriminatory capacity of 40.62% (Table 2).

The *trnL* GH target region had the highest number of species with shared profile (mean N $_{(species)\ sh}$ =196) (Table 2). Among the families of the *trnL* GH target region, the lowest N $_{(species)\ sh}$ was found in the families Chrysobalanaceae and Coriariaceae. In the families of *trnL* GH region the N $_{(profile)\ sh}$ ranging 1 to 43 because in all families had at least one shared profile (Supplementary Table S6).

We detected low diversity values in the intraspecific data sets for all target regions (Figures 2b and 3 and Table 3), corroborating previous observations in SPInDel analyses (Pereira, Carneiro et al. 2010). With the exception of *Ficus carica* (Moraceae), all species had a frequency of species-different profiles lower than 0.17 (Figure 3 and Table 3). *F. carica* presented the highest values for the frequency of species-different profiles (f_{dp} = 0.56) and frequency of species-specific profiles (f_{sp} = 0.44) in the *trnL* CD region (Table 3). This high level of intra-species diversity may result from the fact that *F. carica* (Moraceae) is one of the early domesticated fruit species, where extensive sequence variation has been observed between and within cultivar groups (Baraket, Olfa et al. 2008). The evolutionary history of *F. carica* is linked to a high level of cpDNA polymorphism, which has allowed mutations to accumulate within closely related lineage (Ghada, Ahmed et al. 2010).

Despite the low intra-specific diversity in cpDNA genes, indels polymorphisms have a sufficiently rapid evolutionary rate of accumulation that allows for discrimination between closely related taxa (Pereira, Carneiro et al. 2010). The frequency of species-specific SPInDel profiles in some families reached the maximal possible value (f_{sp}=1), e.g. Apiaceae in *atpF-atpH* intergenic spacer, which indicates that all species of this group had a unique profile for this target region (Supplementary Table S3). The families of the *psbA-trnH* target region had a mean value of N=64 and a mean of N_{sp}=14, while the families of the *trnL* CD target region has a mean of N=57 and the same mean number of species-specific profiles (N_{sp}=14) (Table 2). Taken together these results suggest that *trnL* CD is slightly more informative than *psbA-trnH.* The mean f_{sp} for the families of *atpF-atpH* target region (f_{sp}=0.41) was very close to the mean f_{sp} for the *trnL* CD families (f_{sp}=0.43). However, the former has a mean N far below the latter. These values suggest that even with few species analysed, the region *atpF-atpH* had nearly the same f_{sp} of a region (*trnL* CD) with almost twice of sequences (Table 2).

The concatenation of the cpDNA target regions revealed the real potential of the SPInDel concept (Figure 2a and 2b and Tables 4 and 5). Combining two or three hypervariable regions results in high frequency values of species-specific profiles, reaching a discriminatory power of 92.1% for *atpF-atpH + psbA-trnH + trnL* CD and 84.7% for *atpF-atpH + psbA-trnH + trnL* GH (Figures 2 and 3). The occurrence of hypervariable regions with the same length in different species might not be a problem for the SPInDel approach because it relies on the analysis of multiple loci, which presents a clear advantage over methods targeting a single locus. In cases where one (or more) SPInDel hypervariable region(s) had the same length for two species, a correct identification was still possible based on the information from the remaining regions. For example, *Solanum lycopersicum* and *S. nigrum* presented the same length for *atpF-atpH* (502bp) and for *trnL GH* hypervariable regions (78bp), but they were different for *psbA-trnH* (512bp for *S. lycopersicum* and 497bp for *S. nigrum*; Supplementary Table S8).

When considering the concatenation of *atpF-atpH* + *psbA-trnH* + *trnL* CD*,* the *Picea abies, P. jezoensis* and *P. koraiensis* species had equal SPInDel profiles (Supplementary Table S7). A previous work showed that the genus *Picea* is morphologically uniform and discrete from other genera of the Pinaceae family (Sigurgeirsson and Szmidt 1993). The *Picea* genus is also considered uniform in wood anatomy, growth and ecological preference. The study of 31 species of *Picea* revealed a low level of cpDNA divergence that might result from a slow rates of cpDNA evolution or a recent radiation of *Picea* species from their common ancestor (Sigurgeirsson and Szmidt 1993), which may explain the equal SPInDel profiles. The concatenated analysis of *atpF-atpH* + *psbA-trnH* + *trnL* GH revealed a few shared profiles among species belonging to the same genus (e.g. *Lolium multiflorum* and *L. perenne*). Overall, our approach can discriminate several species from the same genus, such as *Populus balsamifera, P. tremuloides* and *P. alba*, each with a unique SPInDel profile. It has been shown that the DNA barcode combining *matK* and *rbcL* provides a discrimination close to 70-75%, which is far from the mtDNA COI used in metazoan (95%). The SPInDel approach using concatenated regions can discriminate at least 84% of species analyzed (Tables 4 and 5).

Table 5. The discriminatory power (%) of different approaches for the identification of plant species.

Number of markers	Markers	Number of samples	Taxonomic group	Discriminatory power (%)	Reference
3	*atpF-atpH+psbA-trnH+trnL CD*	38	Diverse plant genera	92.1	This work
3	*matK+psbA-trnH+psbK-psbI*	101	Monocotyledons	90.3	(Lahaye 2008)
2	ITS2+*matK*	44	mangrove	89.74	(Saddhe, Jamdade et al. 2017)
4	*atpF-atpH+matK+psbA-trnH+psbK-psbI*	101	Monocotyledons	89.3	(Lahaye 2008)
2	*matK+psbK-psbI*	101	Monocotyledons	87.5	(Lahaye 2008)
2	*psbA-trnH+rbcL*	48	Diverse plant genera	87.5	(Kress and Erickson 2007)
2	*psbA-trnH+rpoB2*	48	Diverse plant genera	87.5	(Kress and Erickson 2007)
2	*psbA-trnH+rpoC1*	48	Diverse plant genera	87.5	(Kress and Erickson 2007)
2	*matK+psbA-trnH*	101	Monocotyledons	87.1	(Lahaye 2008)
3	*atpF-atpH+matK+psbK-psbI*	101	Monocotyledons	86.2	(Lahaye 2008)
3	*atpF-atpH+matK+psbA-trnH*	101	Monocotyledons	85.7	(Lahaye 2008)
3	*atpF-atpH+psbA-trnH+trnL GH*	170	Diverse plant genera	84.7	This work
2	ITS1+*psbA-trnH*	48	Diverse plant genera	83.3	(Kress and Erickson 2007)
2	*atpF-atpH+matK*	101	Monocotyledons	82.8	(Lahaye 2008)
2	*matK+rbcL*	48	Diverse plant genera	79.2	(Kress and Erickson 2007)
2	*rbcL+rpoB2*	48	Diverse plant genera	77.1	(Kress and Erickson 2007)
2	*rbcL+rpoC1*	48	Diverse plant genera	77.1	(Kress and Erickson 2007)
2	*matK+psbA-trnH*	48	Diverse plant genera	75.0	(Kress and Erickson 2007)
7	*atpF-atpH+matK+psbA-trnH+psbK-psbI+rbcL+rpoB+rpoC1*	397	Seed plants	73.0	(Group, Hollingsworth et al. 2009)
2	ITS1+*rbcL*	48	Diverse plant genera	72.3	(Kress and Erickson 2007)
2	*matK+rbcL*	397	Seed plants	72.0	(Group, Hollingsworth et al. 2009)
7	*atpF-atpH+matK+psbA-trnH+psbK-psbI+rbcL+rpoB+rpoC1*	251	Land plants	71.0	(Fazekas, Burgess et al. 2008)
6	*atpF-atpH+matK+psbA-trnH+psbK-psbI+rbcL+rpoC1*	251	Land plants	71.0	(Fazekas, Burgess et al. 2008)
6	*atpF-atpH+matK+psbA-trnH+psbK-psbI+rpoB+rbcL*	251	Land plants	71.0	(Fazekas, Burgess et al. 2008)
6	*atpF-atpH+matK+psbA-trnH+rbcL+rpoB+rpoC1*	251	Land plants	71.0	(Fazekas, Burgess et al. 2008)
6	*atpF-atpH+matK+psbK-psbI+rbcL+rpoB+rpoC1*	251	Land plants	71.0	(Fazekas, Burgess et al. 2008)
6	*atpF-atpH+psbA-trnH+psbK-psbI+rbcL+rpoB+rpoC1*	251	Land plants	71.0	(Fazekas, Burgess et al. 2008)
6	*matK+psbA-trnH+psbK-psbI+rbcL+rpoB+rpoC1*	251	Land plants	71.0	(Fazekas, Burgess et al. 2008)
5	*matK+psbA-trnH+psbK-psbI+rbcL+rpoC1*	251	Land plants	71.0	(Fazekas, Burgess et al. 2008)
4	*psbA-trnH+psbK-psbI+rbcL+rpoC1*	251	Land plants	71.0	(Fazekas, Burgess et al. 2008)
5	*atpF-atpH+matK+psbA-trnH+rpoB+rbcL*	251	Land plants	70.0	(Fazekas, Burgess et al. 2008)
5	*atpF-atpH+matK+psbK-psbI+rpoB+rpoC1*	251	Land plants	70.0	(Fazekas, Burgess et al. 2008)
5	*atpF-atpH+matK+rbcL+rpoB+ rpoC1*	251	Land plants	70.0	(Fazekas, Burgess et al. 2008)
5	*atpF-atpH+psbA-trnH+psbK-psbI+rbcL+rpoB*	251	Land plants	70.0	(Fazekas, Burgess et al. 2008)
3	*atpF-atpH+matK+psbK-psbI*	251	Land plants	69.0	(Fazekas, Burgess et al. 2008)

2	ITS1+*rpoC1*	48	Diverse plant genera	68.8	(Kress and Erickson 2007)
4	*atpF-atpH+trnH-psbA+psbK-psbI+rbcL*	251	Land plants	68.0	(Fazekas, Burgess et al. 2008)
4	*matK+rpoB+rpoC1+rbcL*	251	Land plants	68.0	(Fazekas, Burgess et al. 2008)
4	*psbK-psbI+rbcL+rpoB+rpoC1*	251	Land plants	68.0	(Fazekas, Burgess et al. 2008)
5	*atpF-atpH+matK+psbA-trnH+psbK-psbI+rbcL+rpoB*	28	Pinaceae	67.86	(Ran, Wang et al. 2010)
4	*atpF-atpH+matK+psbA-trnH+psbK-psbI*	251	Land plants	67.0	(Fazekas, Burgess et al. 2008)
6	*atpF-atpH+matK+psbA-trnH+psbK-psbI+rpoB+rpoC1*	251	Land plants	67.0	(Fazekas, Burgess et al. 2008)
5	*atpF-atpH+matK+psbA-trnH+psbK-psbI+rpoC1*	251	Land plants	67.0	(Fazekas, Burgess et al. 2008)
4	*atpF-atpH+matK+rpoB+rpoC1*	251	Land plants	67.0	(Fazekas, Burgess et al. 2008)
5	*matK+psbA-trnH+rbcL+rpoB+rpoC1*	251	Land plants	67.0	(Fazekas, Burgess et al. 2008)
3	*psbK-psbI+rbcL+rpoB*	251	Land plants	67.0	(Fazekas, Burgess et al. 2008)
2	ITS+*rpoB2*	48	Diverse plant genera	66.7	(Kress and Erickson 2007)
3	*atpF-atpH+matK+psbA-trnH*	251	Land plants	66.0	(Fazekas, Burgess et al. 2008)
4	*atpF-atpH+matK+psbA-trnH+rpoB*	251	Land plants	66.0	(Fazekas, Burgess et al. 2008)
3	*atpF-atpH+psbA-trnH+psbK-psbI*	251	Land plants	66.0	(Fazekas, Burgess et al. 2008)
3	*matK+psbA-trnH+rpoC1*	251	Land plants	65.0	(Fazekas, Burgess et al. 2008)
3	*matK+psbA-trnH+psbK-psbI*	28	Pinaceae	64.29	(Ran, Wang et al. 2010)
2	*atpF-atpH+matK*	251	Land plants	64.0	(Fazekas, Burgess et al. 2008)
2	*psbA-trnH+rbcL*	251	Land plants	64.0	(Fazekas, Burgess et al. 2008)
3	*atpF-atpH+psbK-psbI+rbcL*	251	Land plants	63.0	(Fazekas, Burgess et al. 2008)
2	*matK+psbA-trnH*	251	Land plants	63.0	(Fazekas, Burgess et al. 2008)
3	*rbcL+rpoB+rpoC1*	251	Land plants	63.0	(Fazekas, Burgess et al. 2008)
2	*rpoB2+rpoC1*	48	Diverse plant genera	62.5	(Kress and Erickson 2007)
2	*atpF-atpH+trnH-psbA*	251	Land plants	61.0	(Fazekas, Burgess et al. 2008)
3	*matK+rpoB+rpoC1*	251	Land plants	61.0	(Fazekas, Burgess et al. 2008)
3	*atpF-atpH+psbA-trnH+psbK-psbI*	28	Pinaceae	60.71	(Ran, Wang et al. 2010)
3	*atpF-atpH+psbK-psbI+rpoB*	28	Pinaceae	60.71	(Ran, Wang et al. 2010)
3	*matK+psbA-trnH+rbcL*	28	Pinaceae	60.71	(Ran, Wang et al. 2010)
3	*matK+rbcL+rpoB*	28	Pinaceae	60.71	(Ran, Wang et al. 2010)
3	*psbA-trnH+psbK-psbI+rbcL*	28	Pinaceae	60.71	(Ran, Wang et al. 2010)
3	*psbK-psbI+rbcL+rpoB*	28	Pinaceae	60.71	(Ran, Wang et al. 2010)
2	*rbcL+rpoC1*	251	Land plants	60.0	(Fazekas, Burgess et al. 2008)
2	*psbK-psbI+rpoB*	251	Land plants	59.0	(Fazekas, Burgess et al. 2008)
2	*matK+rpoC1*	48	Diverse plant genera	58.3	(Kress and Erickson 2007)
2	*atpF-atpH+psbK-psbI*	251	Land plants	58.0	(Fazekas, Burgess et al. 2008)
3	*atpF-atpH+matK+rbcL*	28	Pinaceae	57.14	(Ran, Wang et al. 2010)
3	*matK+psbK-psbI+rpoB*	28	Pinaceae	57.14	(Ran, Wang et al. 2010)
2	*matK+rpoB2*	48	Diverse plant genera	56.3	(Kress and Erickson 2007)
2	ITS1+*matK*	48	Diverse plant genera	54.2	(Kress and Erickson 2007)

3	*atpF-atpH+matK+psbA-trnH*	28	Pinaceae	53.57	(Ran, Wang et al. 2010)
3	*atpF-atpH+matK+psbK-psbI*	28	Pinaceae	53.57	(Ran, Wang et al. 2010)
3	*atpF-atpH+matK+rpoB*	28	Pinaceae	53.57	(Ran, Wang et al. 2010)
3	*atpF-atpH+psbA-trnH+rbcL*	28	Pinaceae	53.57	(Ran, Wang et al. 2010)
3	*atpF-atpH+psbK-psbI+rbcL*	28	Pinaceae	53.57	(Ran, Wang et al. 2010)
3	*atpF-atpH+rbcL+rpoB*	28	Pinaceae	53.57	(Ran, Wang et al. 2010)
2	*matK+psbK-psbI*	28	Pinaceae	53.57	(Ran, Wang et al. 2010)
3	*matK+psbK-psbI+rbcL*	28	Pinaceae	53.57	(Ran, Wang et al. 2010)
2	*atpF-atpH+psbK-psbI*	28	Pinaceae	50.0	(Ran, Wang et al. 2010)
2	*psbA-trnH+psbK-psbI*	28	Pinaceae	50.0	(Ran, Wang et al. 2010)
2	*rpoB+rpoC1*	251	Land plants	50.0	(Fazekas, Burgess et al. 2008)
2	*atpF-atpH+matK*	28	Pinaceae	46.43	(Ran, Wang et al. 2010)
3	*atpF-atpH+psbA-trnH+rpoB*	28	Pinaceae	46.43	(Ran, Wang et al. 2010)
2	*matK+rbcL*	28	Pinaceae	46.43	(Ran, Wang et al. 2010)
2	*matK+rpoB*	28	Pinaceae	46.43	(Ran, Wang et al. 2010)
2	*matK+psbA-trnH*	28	Pinaceae	42.86	(Ran, Wang et al. 2010)
3	*matK+psbA-trnH+rpoB*	28	Pinaceae	42.86	(Ran, Wang et al. 2010)
3	*psbA-trnH+psbK-psbI+rpoB*	28	Pinaceae	42.86	(Ran, Wang et al. 2010)
3	*psbA-trnH+rbcL+rpoB*	28	Pinaceae	42.86	(Ran, Wang et al. 2010)
2	*psbK-psbI+rbcL*	28	Pinaceae	42.86	(Ran, Wang et al. 2010)
2	*psbA-trnH+rbcL*	28	Pinaceae	39.29	(Ran, Wang et al. 2010)
2	*psbA-trnH+rpoB*	28	Pinaceae	39.29	(Ran, Wang et al. 2010)
2	*atpF-atpH+psbA-trnH*	28	Pinaceae	35.71	(Ran, Wang et al. 2010)
2	*atpF-atpH+rbcL*	28	Pinaceae	35.71	(Ran, Wang et al. 2010)
2	*atpF-atpH+rpoB*	28	Pinaceae	35.71	(Ran, Wang et al. 2010)
2	*rbcL+rpoB*	28	Pinaceae	35.71	(Ran, Wang et al. 2010)

In summary, the SPInDel approach can be used for the identification of plant species. The theoretical work described here demonstrated that a high level of species discrimination is achievable by combining the length of hypervariable regions with indel variants.

Acknowledgements

CS is supported by the “Programa Ciências Sem Fronteiras” from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil. FP is supported by the Portuguese Foundation for Science and Technology (FCT), ERDF and Programa Operacional Potencial Humano (IF/01356/2012). CIIMAR was partially supported by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT and ERDF in the framework of the programme PT2020.

References

[1] Arenas M, Pereira F, Oliveira M, Pinto N, Lopes AM, Gomes V, et al. Forensic genetics and genomics: Much more than just a human affair. PLoS Genetics. 2017;13:e1006960.

[2] Bell KL, Burgess KS, Okamoto KC, Aranda R, Brosi BJ. Review and future prospects for DNA barcoding methods in forensic palynology. Forensic Science International: Genetics. 2016;21:110-6.

[3] Coyle HM. Forensic botany: principles and applications to criminal casework: CRC Press; 2004.

[4] Moreira F, Carneiro J, Pereira F. A proposal for standardization of transgenic reference sequences used in food forensics. Forensic Science International: Genetics. 2017;29:e26-e8.

[5] Ogden R, Linacre A. Wildlife forensic science: a review of genetic geographic origin assignment. Forensic Science International: Genetics. 2015;18:152-9.

[6] Zaya DN, Ashley MV. Plant genetics for forensic applications. Plant DNA Fingerprinting and Barcoding: Methods and Protocols. 2012:35-52.

[7] Linacre A, Tobe SS. An overview to the investigative approach to species testing in wildlife forensic science. Investigative genetics. 2011;2:2.

[8] Woolfe M, Primrose S. Food forensics: using DNA technology to combat misdescription and fraud. Trends in Biotechnology. 2004;22:222-6.

[9] Pereira F, Carneiro J, Amorim A. Identification of species with DNA-based technology: current progress and challenges. Recent patents on DNA & gene sequences. 2008;2:187-200.

[10] Hollingsworth, M. L., et al. (2009). "Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants." Mol Ecol Resour **9**(2): 439-457.

[11] Ferri G, Corradini B, Ferrari F, Santunione A, Palazzoli F, Alu M. Forensic botany II, DNA barcode for land plants: Which markers after the international agreement? Forensic Science International: Genetics. 2015;15:131-6.

[12] Pennisi E. Wanted: a barcode for plants. Science. 2007;318:190-1.

[13] Chase MW, Fay MF. Barcoding of plants and fungi. Science. 2009;325:682-3.

[14] Carneiro J, Pereira F, Amorim A. SPInDel: a multifunctional workbench for species identification using insertion/deletion variants. Molecular ecology resources. 2012;12:1190-5.

[15] Pereira F, Carneiro J, Matthiesen R, van Asch B, Pinto N, Gusmão L, et al. Identification of species by multiplex analysis of variable-length sequences. Nucleic acids research. 2010;38:e203-e.

[16] Alves C, Pereira R, Prieto L, Aler M, Amaral CR, Arévalo C, et al. Species identification in forensic samples using the SPInDel approach: A GHEP-ISFG inter-laboratory collaborative exercise. Forensic Science International: Genetics. 2017;28:219-24.

[17] Gonçalves J, Marks CA, Obendorf D, Amorim A, Pereira F. A multiplex PCR assay for identification of the red fox (Vulpes vulpes) using the mitochondrial ribosomal RNA genes. Conservation genetics resources. 2015;7:45-8.

[18] Wolfe KH, Li W-H, Sharp PM. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proceedings of the National Academy of Sciences. 1987;84:9054-8.

[19] Lynch M, Koskella B, Schaack S. Mutation pressure and the evolution of organelle genomic architecture. Science. 2006;311:1727-30.

[20] Pereira F, Carneiro J, Van Asch B. A guide for mitochondrial DNA analysis in non-human forensic investigations. Open Forensic Science Journal. 2010;3:33-44.

[21] Ford CS, Ayres KL, Toomey N, Haider N, Van Alphen Stahl J, Kelly LJ, et al. Selection of candidate coding DNA barcoding regions for use on land plants. Botanical Journal of the Linnean Society. 2009;159:1-11.

[22] Hollingsworth, P. M., et al. (2011). "Choosing and using a plant DNA barcode." PLoS One **6**(5): e19254.

[23] Olmstead RG, Palmer JD. Chloroplast DNA systematics: a review of methods and data analysis. American journal of botany. 1994:1205-24.

[24] Santos C, Pereira F. Design and evaluation of PCR primers for amplification of four chloroplast DNA regions in plants. Conservation Genetics Resources. 2017;9:9-12.

[25] Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, et al. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. American journal of botany. 2005;92:142-66.

[26] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792-7.

[27] Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28:1647-9.

[28] Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, et al. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res. 2007;35:e14.

[29] Yamane K, Yano K, Kawahara T. Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. DNA research. 2006;13:197-204.

[30] Kelchner SA. The evolution of non-coding chloroplast DNA and its application in plant systematics. Annals of the Missouri Botanical Garden. 2000:482-98.

[31] Graham SW, Reeves PA, Burns AC, Olmstead RG. Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. International Journal of Plant Sciences. 2000;161:S83-S96.

[32] Green BR. Chloroplast genomes of photosynthetic eukaryotes. The plant journal. 2011;66:34-44.

[33] Xiong A-S, Peng R-H, Zhuang J, Gao F, Zhu B, Fu X-Y, et al. Gene duplication, transfer, and evolution in the chloroplast genome. Biotechnology advances. 2009;27:340-7.

[34] Wallinger C, Juen A, Staudacher K, Schallhart N, Mitterrutzner E, Steiner E-M, et al. Rapid plant identification using species-and group-specific primers targeting chloroplast DNA. PLoS One. 2012;7:e29473.

[35] Kress WJ, Erickson DL. A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS one. 2007;2:e508.

[36] Yao H, Song J-Y, Ma X-Y, Liu C, Li Y, Xu H-X, et al. Identification of Dendrobium species by a candidate DNA barcode sequence: the chloroplast psbA-trnH intergenic region. Planta medica. 2009;75:667-9.

[37] Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences of the United States of America. 2005;102:8369-74.

[38] Pang X, Luo H, Sun C. Assessing the potential of candidate DNA barcodes for identifying non-flowering seed plants. Plant Biology. 2012;14:839-44.

[39] Lahaye RS, V.; Duthoit, S.; Maurin, O. and van der Bank, M. A test of psbK-psbI and atpF-atpH as potential plant DNA barcodes using the flora of the Kruger National Park as a model system (South Africa). Nature Precedings. 2008:21.

[40] Baraket G, Olfa S, Khaled C, Messaoud M, Mohamed M, Mokhtar T, et al. Chloroplast DNA analysis in Tunisian fig cultivars (Ficus carica L.): Sequence variations of the trnL-trnF intergenic spacer. Biochemical systematics and ecology. 2008;36:828-35.

[41] Ghada B, Ahmed BA, Khaled C, Olfa S, Messaoud M, Mokhtar T, et al. Molecular evolution of chloroplast DNA in fig (Ficus carica L.): Footprints of sweep selection and recent expansion. Biochemical Systematics and Ecology. 2010;38:563-75.

[42] Sigurgeirsson A, Szmidt AE. Phylogenetic and biogeographic implications of chloroplast DNA variation in Picea. Nordic Journal of Botany. 1993;13:233-46.

[43] Saddhe AA, Jamdade RA, Kumar K. Evaluation of multilocus marker efficacy for delineating mangrove species of West Coast India. PloS one. 2017;12:e0183245.

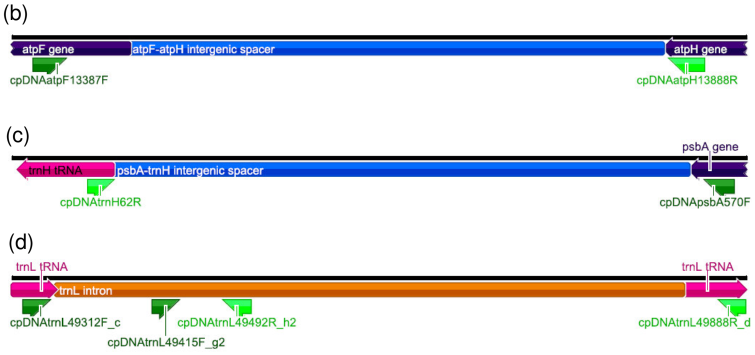
[44] Group, C. P. W., et al. (2009). "A DNA barcode for land plants." Proceedings of the National Academy of Sciences **106**(31): 12794-12797.

[46] Ran JH, Wang PP, Zhao HJ, Wang XQ. A test of seven candidate barcode regions from the plastome in Picea (Pinaceae). Journal of integrative plant biology. 2010;52:1109-26.

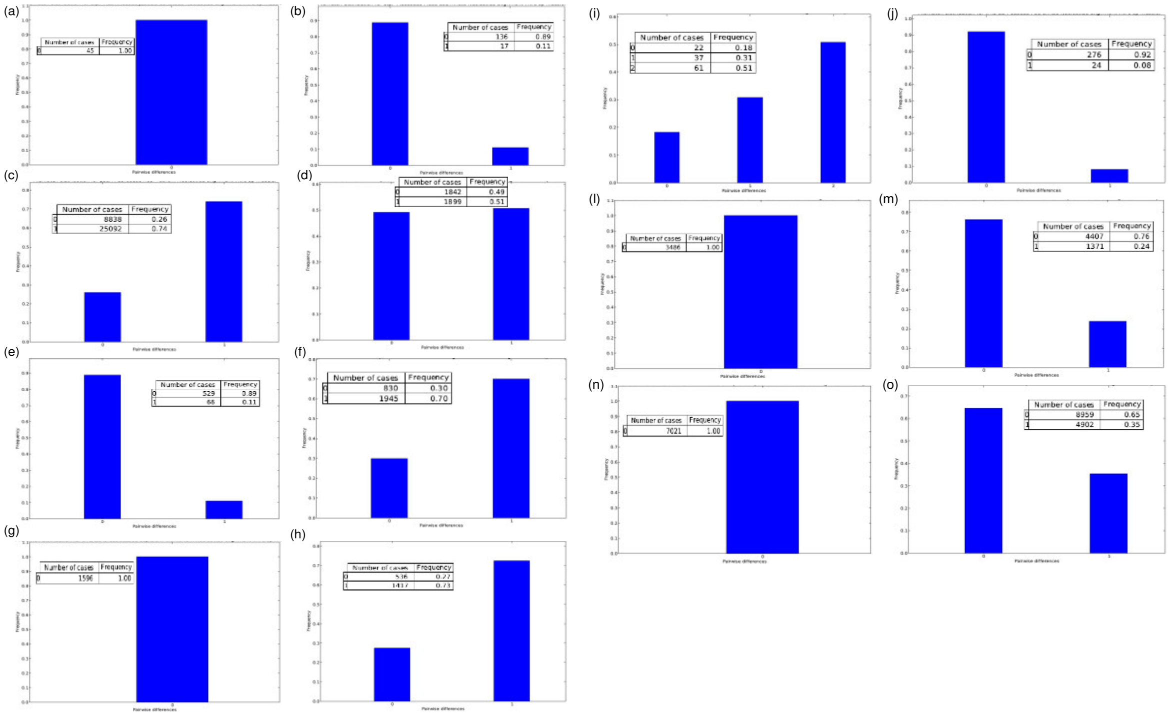
Supplementary Material

(a)

Genomic region	Length (bp)	Position
atpF-atpH	509	13387 - 13888
psbA-trnH	502	62 - 570
trnL CD	577	49312 - 49888
trnL GH	78	49415 - 49492



Supplementary Figure S1. Genomic regions selected for testing the SPInDel method. (a) Name, length and genomic location of the target regions in the *Nicotiana tabacum* cpDNA reference sequence (NC_001879.2). (b) Location of the SPInDel conserved regions (green arrows) in the cpDNA *atpF-atpH* genomic region. (c) Location of the SPInDel conserved regions (green arrows) in the cpDNA *psbA-trnH* genomic region. (d) Location of the SPInDel conserved regions (green arrows) in the cpDNA *trnL* (UAA) gene region.



Supplementary Figure S2. Mismatch distributions per species: (a) *atpF-atpH Justicia adhatoda* (Acantaceae) (b) *atpF-atpH Musa acuminata* (Musaceae) (c) *psbA-trnH Acer rubrum* (Aceraceae) (d) *psbA-trnH Onobrychis vicifolia* (Fabaceae) (e) *psbA-trnH Phalaris arundinaceae* (Poaceae) (f) *psbA-trnH Potentilla argentea* (Rosaceae) (g) *trnL* CD *Lepidium montanum* (Brassicaceae) (h) *trnL* CD *Silene latifolia* (Caryophyllaceae) (i) *trnL* CD *Ficus carica* (Moraceae) (j) *trnL* CD *Poa annua* (Poaceae) (l) *trnL* GH *Brochera holboelli* (Brassicaceae) (m) *trnL* GH *Ranunculus kuepferi* (Ranunculaceae) (n) *trnL* GH *Carapichea ipecacuanha* (Rubiaceae) (o) *trnL* GH *Populus balsamifera* (Salicaceae).

Supplementary Table S1. Terms associated with the SPInDel concept.

Term	Definition	Symbol	Formula
SPInDel conserved region	Regions with none or small variability at the sequence level used to delimit the hypervariable segments	e.g. A, B, C	
SPInDel hypervariable region	Regions containing multiple indels across species that potentially allow for differentiation by the determination of sequence length	e.g., AB, BC	
Number of SPInDel hypervariable regions		*n*	
Standard SPInDel profile	Set of fragment length of all contiguous SPInDel hypervariable regions observed in a sequence (AB length; BC length; CD length)		
Taxonomic group	Taxonomic group under investigation (e.g., family, genus)	G	
Number of sequences	Total number of sequences represented on group G	N	
Species-specific SPInDel profile	Profile that is found in one specie within a taxonomic group and can be defined by a unique combination of fragments lengths (e.g. 155-191-69-223)	*sp*	
Number of species-specific SPInDel profiles	The number of unique profiles found in a taxonomic group	N_{sp}	
Species-shared profile	Profile common to more than one species within a taxonomic group	*sh*	
Number of species-shared SPInDel profiles	Number of profiles shared between species at a group	$N_{(profile)\ sh}$	
Number of species with shared SPInDel profile	Number of species that have shared profile	$N_{(species)\ sh}$	
Total number of different profiles	The number of profiles found in a taxonomic group, i.e. the number of species-specific SPInDel profiles plus the number of species-shared SPInDel profiles	N_{dp}	$N_{dp}=N_{sp}+N_{(profile)sh}$
Frequency of species-specific SPInDel profiles	The frequency of unique profiles found in a taxonomic group	f_n^G or f_{sp}	$f_n^G = \frac{N_{sp}}{N}$
Average number of pairwise differences	Average number of differences in the length of hypervariable regions between two individual profiles	$\bar{p}_n^G$ or p	$\bar{p}\,{}_n^G = \frac{\sum_{k=1}^{N}\sum_{l>k}^{N} d_{kl}}{\frac{N(N-1)}{2}}$
Mismatch distribution	Frequency distribution of the number of SPInDel hypervariable regions that differ between all pairs of SPInDel profiles in a taxonomic group		
Average number of pairwise differences per locus	The average number of pairwise differences considering each locus		$\frac{\bar{p}_n^G}{n}$
Discrimination Power (%)	The percentage of species that present a unique profile on a particular group	DP	$DP = f_{sp}.100$

Supplementary Table S2. Number of individuals and species used in intra-specific analysis.

Region	Family	Species	Number of individuals
atpF-atpH	Acanthaceae	*Justicia adhatoda*	10
atpF-atpH	Musaceae	*Musa acuminata*	18
psbA-trnH	Aceraceae	*Acer rubrum*	261
psbA-trnH	Fabaceae	*Onobrychis viciifolia*	87
psbA-trnH	Poaceae	*Phalaris arundinaceae*	35
psbA-trnH	Rosaceae	*Potentilla argentea*	75
trnL CD	Brassicaceae	*Lepidium montanum*	57
trnL CD	Caryophyllaceae	*Silene latifolia*	63
trnL CD	Moraceae	*Ficus carica*	16
trnL CD	Poaceae	*Poa annua*	25
trnL GH	Brassicaceae	*Boechera holboelli*	84
trnL GH	Ranunculaceae	*Ranunculus kuepferi*	108
trnL GH	Rubiaceae	*Carapichea ipecacuanha*	119
trnL GH	Salicaceae	*Populus balsamifera*	167

Supplementary Table S3. General description of standard SPInDel profiles from the *atpF-atpH* cpDNA region.

Family	Number of species per family (N)	Number of species-specific profiles (N_{sp})	Frequency of species-specific profiles (f_{sp})	Number of species with shared profiles ($N_{(species)\ sh}$)	Number of species-shared profiles ($N_{(profile)\ sh}$)	Frequency of species-shared profiles (f_{sh})	Number of species-different profiles (N_{dp})	Frequency of species-different profiles (f_{dp})	Average number of pairwise differences ($\bar{p}_n^G$)	Discrimination power (%)
Apiaceae	11	11	1.00	0	0	0.00	11	1.00	1.00	100.00
Araceae	34	12	0.35	22	9	0.26	21	0.62	0.97	35.29
Araucariaceae	15	2	0.13	13	2	0.13	4	0.27	0.54	13.33
Asparagaceae	24	5	0.21	19	6	0.25	11	0.46	0.90	20.83
Asteraceae	62	21	0.34	41	11	0.18	32	0.52	0.95	33.87
Brassicaceae	14	9	0.64	5	2	0.14	11	0.79	0.96	64.29
Campanulaceae	33	5	0.15	28	6	0.18	11	0.33	0.86	15.15
Colchicaceae	12	8	0.67	4	2	0.17	10	0.83	0.97	66.67
Cupressaceae	17	9	0.53	8	3	0.18	12	0.71	0.94	52.94
Cyperaceae	28	10	0.36	18	4	0.14	14	0.50	0.89	35.71
Fissidentaceae	11	8	0.73	3	1	0.09	9	0.82	0.95	72.73
Iridaceae	22	8	0.36	14	3	0.14	11	0.50	0.85	36.36
Liliaceae	37	19	0.51	18	6	0.16	25	0.68	0.96	51.35
Melanthiaceae	19	7	0.37	12	5	0.26	12	0.63	0.94	36.84
Melastomataceae	23	15	0.65	8	4	0.17	19	0.83	0.98	65.22
Moraceae	67	5	0.07	62	11	0.16	16	0.24	0.87	7.46
Musaceae	36	3	0.08	33	7	0.19	10	0.28	0.88	8.33
Orchidaceae	21	11	0.52	10	5	0.24	16	0.76	0.98	52.38
Paniceae	28	3	0.11	25	5	0.18	8	0.29	0.79	10.71
Pinaceae	58	7	0.12	51	7	0.12	14	0.24	0.70	12.07
Plantaginaceae	13	8	0.62	5	1	0.08	9	0.69	0.87	61.54
Poaceae	181	21	0.12	160	25	0.14	46	0.25	0.95	11.60
Primulaceae	14	9	0.64	5	2	0.14	11	0.79	0.96	64.29
Ranunculaceae	19	11	0.58	8	3	0.16	14	0.74	0.95	57.89
Rosaceae	22	11	0.50	11	5	0.23	16	0.73	0.97	50.00
Salicaceae	11	9	0.82	2	1	0.09	10	0.91	0.98	81.82
Violaceae	48	15	0.31	33	9	0.19	24	0.50	0.95	31.25
Zamiaceae	64	2	0.03	62	6	0.09	8	0.12	0.84	3.13
Zingiberaceae	16	4	0.25	12	4	0.25	8	0.50	0.88	25.00

Supplementary Table S4. General description of standard SPInDel profiles from the *psbA-trnH* cpDNA region.

Family	Number of species per family (N)	Number of species-specific profiles (N_{sp})	Frequency of species-specific profiles (f_{sp})	Number of species with shared profiles ($N_{(species)\ sh}$)	Number of species-shared profiles ($N_{(profile)\ sh}$)	Frequency of species-shared profiles (f_{sh})	Number of species-different profiles (N_{dp})	Frequency of species-different profiles (f_{dp})	Average number of pairwise differences ($\bar{p}_n^G$)	Discrimination power (%)
Acanthaceae	15	11	0.73	4	2	0.13	13	0.87	0.98	73.33
Aceraceae	42	13	0.31	29	13	0.31	26	0.62	0.98	30.95
Adoxaceae	81	15	0.19	66	16	0.20	31	0.38	0.95	18.52
Alismataceae	42	7	0.17	35	5	0.12	12	0.29	0.69	16.67
Annonaceae	63	16	0.25	47	5	0.08	21	0.33	0.83	25.40
Apiaceae	99	7	0.07	92	16	0.16	23	0.23	0.94	7.07
Apocynaceae	58	26	0.45	32	12	0.21	38	0.66	0.98	44.83
Araliaceae	115	23	0.20	92	23	0.20	46	0.40	0.96	20.00
Arecaceae	24	18	0.75	6	3	0.12	21	0.88	0.99	75.00
Aspleniaceae	11	4	0.36	7	3	0.27	7	0.64	0.91	36.36
Asteraceae	275	40	0.15	235	48	0.17	88	0.32	0.97	14.55
Betulaceae	42	14	0.33	28	9	0.21	23	0.55	0.95	33.33
Boraginaceae	102	25	0.25	77	13	0.13	38	0.37	0.92	24.51
Brassicaceae	16	4	0.25	12	3	0.19	7	0.44	0.79	25.00
Burseraceae	62	9	0.15	53	6	0.10	15	0.24	0.69	14.52
Cactaceae	15	10	0.67	5	2	0.13	12	0.80	0.96	66.67
Caprifoliaceae	18	12	0.67	6	3	0.17	15	0.83	0.98	66.67
Celastraceae	26	14	0.54	12	5	0.19	19	0.73	0.97	53.85
Colchicaceae	54	19	0.35	35	7	0.13	26	0.48	0.93	35.19
Combretaceae	43	30	0.70	13	5	0.12	35	0.81	0.99	69.77
Convolvulaceae	12	10	0.83	2	1	0.08	11	0.92	0.98	83.33
Cornaceae	39	16	0.41	23	7	0.18	23	0.59	0.95	41.03
Curcubitaceae	196	25	0.13	171	30	0.15	55	0.28	0.96	12.76
Cupressaceae	17	15	0.88	2	1	0.06	16	0.94	0.99	88.24
Cyperaceae	42	4	0.10	38	2	0.05	6	0.14	0.27	9.52
Dicranaceae	11	4	0.36	7	3	0.27	7	0.64	0.91	36.36
Dioscoreaceae	51	15	0.29	36	11	0.22	26	0.51	0.96	29.41
Ephedraceae	21	0	0.00	21	2	0.10	2	0.10	0.32	0.00
Escalloniaceae	37	10	0.27	27	6	0.16	16	0.43	0.85	27.03
Fabaceae	358	75	0.21	283	66	0.18	141	0.39	0.98	20.95
Fagaceae	10	3	0.30	7	2	0.20	5	0.50	0.80	30.00
Fissidentaceae	12	1	0.08	11	3	0.25	4	0.33	0.71	8.33
Frullaniaceae	114	5	0.04	109	11	0.10	16	0.14	0.79	4.39
Gentianaceae	47	29	0.62	18	8	0.17	37	0.79	0.99	61.70
Gesneriaceae	213	31	0.15	182	45	0.21	76	0.36	0.98	14.55
Grossulariaceae	42	6	0.14	36	12	0.29	18	0.43	0.94	14.29

Hamamelidaceae	19	3	0.16	16	6	0.32	9	0.47	0.88	15.79
Hyacinthaceae	21	7	0.33	14	3	0.14	10	0.48	0.84	33.33
Hydrangeaceae	34	13	0.38	21	8	0.24	21	0.62	0.97	38.24
Hymenophyllaceae	14	14	1.00	0	0	0.00	14	1.00	1.00	100.00
Iridaceae	49	7	0.14	42	7	0.14	14	0.29	0.79	14.29
Jamesoniellaceae	30	5	0.17	25	2	0.07	7	0.23	0.42	16.67
Junglandaceae	18	1	0.06	17	2	0.11	3	0.17	0.58	5.56
Lamiaceae	84	47	0.56	37	12	0.14	59	0.70	0.98	55.95
Lauraceae	36	9	0.25	27	8	0.22	17	0.47	0.92	25.00
Lejeneaceae	77	8	0.10	69	14	0.18	22	0.29	0.93	10.39
Linaceae	16	4	0.25	12	3	0.19	7	0.44	0.79	25.00
Loasaceae	10	7	0.70	3	1	0.10	8	0.80	0.93	70.00
Loranthaceae	12	10	0.83	2	1	0.08	11	0.92	0.98	83.33
Magnoliaceae	23	10	0.43	13	5	0.22	15	0.65	0.95	43.48
Melastomataceae	20	9	0.45	11	5	0.25	14	0.70	0.96	45.00
Moraceae	84	16	0.19	68	17	0.20	33	0.39	0.93	19.05
Myrtaceae	44	19	0.43	25	10	0.23	29	0.66	0.98	43.18
Oleaceae	11	4	0.36	7	2	0.18	6	0.55	0.80	36.36
Orchidaceae	406	27	0.07	379	40	0.10	67	0.17	0.92	6.65
Orobanchaceae	24	12	0.50	12	5	0.21	17	0.71	0.97	50.00
Orthotrichaceae	32	5	0.16	27	3	0.09	8	0.25	0.68	15.63
Pallaviciniaceae	10	7	0.70	3	1	0.10	8	0.80	0.93	70.00
Passifloriaceae	31	13	0.42	18	7	0.23	20	0.65	0.96	41.94
Penaeaceae	29	5	0.17	24	5	0.17	10	0.34	0.76	17.24
Pinaceae	50	15	0.30	35	8	0.16	23	0.46	0.90	30.00
Piperaceae	21	12	0.57	9	4	0.19	16	0.76	0.97	57.14
Poaceae	425	13	0.03	412	36	0.08	49	0.12	0.95	3.06
Polygonaceae	49	23	0.47	26	7	0.14	30	0.61	0.95	46.94
Potamogetonaceae	34	11	0.32	23	6	0.18	17	0.50	0.93	32.35
Pottiaceae	15	3	0.20	12	2	0.13	5	0.33	0.71	20.00
Primulaceae	30	7	0.23	23	9	0.30	16	0.53	0.95	23.33
Pteridaceae	17	11	0.65	6	2	0.12	13	0.76	0.95	64.71
Ranunculaceae	114	25	0.22	89	24	0.21	49	0.43	0.97	21.93
Rosaceae	228	44	0.19	184	42	0.18	86	0.38	0.98	19.30
Rubiaceae	115	22	0.19	93	27	0.23	49	0.43	0.96	19.13
Salicaceae	11	7	0.64	4	2	0.18	9	0.82	0.96	63.64
Saxifragaceae	20	11	0.55	9	4	0.20	15	0.75	0.97	55.00
Solanaceae	160	33	0.21	127	18	0.11	51	0.32	0.93	20.63
Sphagnaceae	23	3	0.13	20	5	0.22	8	0.35	0.85	13.04
Tamaricaceae	15	7	0.47	8	3	0.20	10	0.67	0.92	46.67
Veroniceae	62	19	0.31	43	15	0.24	34	0.55	0.98	30.65
Violaceae	49	24	0.49	25	11	0.22	35	0.71	0.99	48.98
Vitaceae	115	19	0.17	96	19	0.17	38	0.33	0.94	16.52

Supplementary Table S5. General description of standard SPInDel profiles from the *trnL* CD cpDNA region.

Family	Number of species per family (N)	Number of species-specific profiles (N_{sp})	Frequency of species-specific profiles (f_{sp})	Number of species with shared profiles ($N_{(species)\ sh}$)	Number of species-shared profiles ($N_{(profile)\ sh}$)	Frequency of species-shared profiles (f_{sh})	Number of species-different profiles (N_{dp})	Frequency of species-different profiles (f_{dp})	Average number of pairwise differences ($\bar{\boldsymbol{p}}_{\boldsymbol{n}}^{\boldsymbol{G}}$)	Discrimination power (%)
Amaryllidaceae	18	1	0.06	17	1	0.06	2	0.11	0.11	5.56
Annonaceae	87	6	0.07	81	6	0.07	12	0.14	0.77	6.90
Apocynaceae	22	6	0.27	16	3	0.14	9	0.41	0.85	27.27
Araceae	63	31	0.49	32	11	0.17	42	0.67	1.70	49.21
Asteraceae	283	19	0.07	264	26	0.09	45	0.16	1.34	6.71
Boraginaceae	27	7	0.26	20	3	0.11	10	0.37	1.00	25.93
Brassicaceae	176	11	0.06	165	6	0.03	17	0.10	1.15	6.25
Burseraceae	20	6	0.30	14	3	0.15	9	0.45	1.13	30.00
Cactaceae	84	20	0.24	64	12	0.14	32	0.38	1.74	23.81
Caryophyllaceae	28	24	0.86	4	2	0.07	26	0.93	2.60	85.71
Cephalotaxaceae	13	4	0.31	9	1	0.08	5	0.38	0.81	30.77
Cyatheaceae	11	6	0.55	5	2	0.18	8	0.73	0.93	54.55
Ericaceae	10	10	1.00	0	0	0.00	10	1.00	2.51	100.00
Eriocaulaceae	33	15	0.45	18	3	0.09	18	0.55	1.40	45.45
Euphorbiaceae	39	14	0.36	25	9	0.23	23	0.59	1.62	35.90
Fabaceae	232	55	0.24	177	34	0.15	89	0.38	1.90	23.71
Gesneriaceae	21	10	0.48	11	4	0.19	14	0.67	1.55	47.62
Gnetaceae	13	11	0.85	2	1	0.08	12	0.92	1.73	84.62
Goodeniaceae	13	13	1.00	0	0	0.00	13	1.00	2.06	100.00
Iridaceae	15	10	0.67	5	2	0.13	12	0.80	2.20	66.67
Juncaceae	48	29	0.60	19	5	0.10	34	0.71	2.12	60.42
Lamiaceae	22	5	0.23	17	5	0.23	10	0.45	1.91	22.73
Liliaceae	15	10	0.67	5	2	0.13	12	0.80	1.76	66.67
Magnoliaceae	10	1	0.10	9	1	0.10	2	0.20	0.20	10.00
Malvaceae	24	17	0.71	7	3	0.12	20	0.83	2.40	70.83
Melanthiaceae	19	10	0.53	9	4	0.21	14	0.74	2.25	52.63
Oleaceae	13	4	0.31	9	2	0.15	6	0.46	1.03	30.77
Orchidaceae	40	22	0.55	18	5	0.12	27	0.68	2.22	55.00
Orobanchaceae	57	35	0.61	22	6	0.11	41	0.72	2.12	61.40
Pinaceae	55	7	0.13	48	7	0.13	14	0.25	1.10	12.73
Poaceae	397	61	0.15	336	48	0.12	109	0.27	1.66	15.37
Polygonaceae	12	9	0.75	3	1	0.08	10	0.83	2.00	75.00
Rosaceae	21	10	0.48	11	3	0.14	13	0.62	1.71	47.62
Rubiaceae	335	45	0.13	290	29	0.09	74	0.22	1.33	13.43
Rutaceae	17	8	0.47	9	3	0.18	11	0.65	1.49	47.06
Salicaceae	17	11	0.65	6	2	0.12	13	0.76	2.12	64.71
Saxifragaceae	12	12	1.00	0	0	0.00	12	1.00	2.41	100.00

Scrophulariaceae	49	11	0.22	38	5	0.10	16	0.33	1.10	22.45
Solanaceae	26	4	0.15	22	4	0.15	8	0.31	0.76	15.38
Stilbaceae	13	7	0.54	6	2	0.15	9	0.69	1.83	53.85
Taxaceae	24	11	0.46	13	2	0.08	13	0.54	1.87	45.83
Theaceae	19	2	0.11	17	5	0.26	7	0.37	1.51	10.53
Verbenaceae	38	4	0.11	34	2	0.05	6	0.16	0.30	10.53
Vitaceae	36	18	0.50	18	4	0.11	22	0.61	1.82	50.00

Supplementary Table S6. General description of standard SPInDel profiles from the *trnL* GH cpDNA region.

Family	Number of species per family (N)	Number of species-specific profiles (N_{sp})	Frequency of species-specific profiles (f_{sp})	Number of species with shared profiles ($N_{(species)\ sh}$)	Number of species-shared profiles ($N_{(profile)\ sh}$)	Frequency of species-shared profiles (f_{sh})	Number of species-different profiles (N_{dp})	Frequency of species-different profiles (f_{dp})	Average number of pairwise differences ($\bar{p}_n^G$)	Discrimination power (%)
Acanthaceae	231	2	0.01	229	12	0.05	14	0.06	0.78	0.87
Aceraceae	118	2	0.02	116	2	0.02	4	0.03	0.10	1.69
Actinidiaceae	37	2	0.05	35	3	0.08	5	0.14	0.30	5.41
Adoxaceae	14	5	0.36	9	1	0.07	6	0.43	0.60	35.71
Aizoaceae	218	10	0.05	208	12	0.06	22	0.10	0.88	4.59
Amaranthaceae	150	11	0.07	139	28	0.19	39	0.26	0.95	7.33
Amaryllidaceae	440	1	0.00	439	12	0.03	13	0.03	0.77	0.23
Anacardiaceae	140	1	0.01	139	8	0.06	9	0.06	0.45	0.71
Annonaceae	670	5	0.01	665	13	0.02	18	0.03	0.76	0.75
Apiaceae	188	6	0.03	182	9	0.05	15	0.08	0.52	3.19
Apocynaceae	810	7	0.01	803	12	0.01	19	0.02	0.49	0.86
Aquifoliaceae	108	1	0.01	107	1	0.01	2	0.02	0.02	0.93
Araceae	434	3	0.01	431	16	0.04	19	0.04	0.83	0.69
Araliaceae	321	0	0.00	321	3	0.01	3	0.01	0.10	0.00
Arecaceae	300	2	0.01	298	7	0.02	9	0.03	0.50	0.67
Aristolochiaceae	75	14	0.19	61	15	0.20	29	0.39	0.94	18.67
Asparagaceae	159	2	0.01	157	10	0.06	12	0.08	0.85	1.26
Asteliaceae	38	1	0.03	37	2	0.05	3	0.08	0.38	2.63
Asteraceae	1955	4	0.00	1951	13	0.01	17	0.01	0.54	0.20
Begoniaceae	97	2	0.02	95	4	0.04	6	0.06	0.28	2.06
Berberidaceae	64	7	0.11	57	4	0.06	11	0.17	0.48	10.94
Betulaceae	58	0	0.00	58	3	0.05	3	0.05	0.22	0.00
Bignoniaceae	121	3	0.02	118	7	0.06	10	0.08	0.79	2.48
Boraginaceae	405	1	0.00	404	19	0.05	20	0.50	0.77	0.25
Brassicaceae	1236	7	0.01	1229	10	0.01	17	0.01	0.50	0.57
Bromeliaceae	313	0	0.00	313	4	0.01	4	0.01	0.10	0.00
Burseraceae	69	2	0.03	67	2	0.03	4	0.06	0.16	2.90
Cactaceae	449	15	0.03	434	34	0.08	49	0.11	0.88	3.34
Calceolariaceae	12	1	0.08	11	2	0.17	3	0.25	0.59	8.33
Calycanthaceae	12	1	0.08	11	1	0.08	2	0.17	0.17	8.33
Campanulaceae	288	1	0.00	287	10	0.03	11	0.04	0.42	0.35
Canellaceae	14	0	0.00	14	1	0.07	1	0.07	0.00	0.00
Cannabaceae	22	3	0.14	19	2	0.09	5	0.23	0.41	13.64
Capparaceae	22	4	0.18	18	5	0.23	9	0.41	0.88	18.18
Caprifoliaceae	241	2	0.01	239	12	0.05	14	0.06	0.86	0.83
Caricaceae	31	0	0.00	31	1	0.03	1	0.03	0.00	0.00
Caryophyllaceae	189	14	0.07	175	22	0.12	36	0.19	0.91	7.41

Celastraceae	265	8	0.03	257	10	0.04	18	0.07	0.45	3.02
Cephalotaxaceae	13	0	0.00	13	1	0.08	1	0.08	0.00	0.00
Chloranthaceae	51	0	0.00	51	8	0.16	8	0.16	0.85	0.00
Chrysobalanaceae	10	2	0.20	8	2	0.20	4	0.40	0.73	20.00
Clethraceae	21	0	0.00	21	3	0.14	3	0.14	0.41	0.00
Clusiaceae	17	8	0.47	9	4	0.24	12	0.71	0.96	47.06
Colchicaceae	217	3	0.01	214	1	0.00	4	0.02	0.03	1.38
Convolvulaceae	225	10	0.04	215	20	0.09	30	0.13	0.89	4.44
Coriariaceae	10	2	0.20	8	1	0.10	3	0.30	0.38	20.00
Cornaceae	16	2	0.12	14	2	0.12	4	0.25	0.44	12.50
Costaceae	55	3	0.05	52	3	0.05	6	0.11	0.35	5.45
Crassulaceae	177	4	0.02	173	6	0.03	10	0.06	0.62	2.26
Cucurbitaceae	416	2	0.00	414	22	0.05	24	0.06	0.89	0.48
Cunoniaceae	30	6	0.20	24	4	0.13	10	0.33	0.82	20.00
Cupressaceae	133	0	0.00	133	4	0.03	4	0.03	0.16	0.00
Cyperaceae	591	26	0.04	565	43	0.07	69	0.12	0.78	4.40
Dioscoreaceae	62	6	0.10	56	8	0.13	14	0.23	0.86	9.68
Dipterocarpaceae	169	1	0.01	168	2	0.01	3	0.02	0.10	0.59
Ebenaceae	130	4	0.03	126	4	0.03	8	0.06	0.26	3.08
Elaeagnaceae	23	2	0.09	21	2	0.09	4	0.17	0.38	8.70
Elaeocarpaceae	65	2	0.03	63	3	0.05	5	0.08	0.52	3.08
Ephedraceae	18	1	0.06	17	1	0.06	2	0.11	0.11	5.56
Ericaceae	344	2	0.01	342	14	0.04	16	0.05	0.79	0.58
Euphorbiaceae	796	12	0.02	784	34	0.04	46	0.06	0.92	1.51
Fabaceae	2599	6	0.00	2593	32	0.01	38	0.01	0.83	0.23
Fagaceae	89	3	0.03	86	8	0.09	11	0.12	0.55	3.37
Gentianaceae	406	10	0.02	396	30	0.07	40	0.10	0.95	2.46
Geraniaceae	160	6	0.04	154	8	0.05	14	0.09	0.62	3.75
Gesneriaceae	556	6	0.01	550	9	0.02	15	0.03	0.34	1.08
Goodeniaceae	164	1	0.01	163	12	0.07	13	0.08	0.73	0.61
Haemodoraceae	51	4	0.08	47	8	0.16	12	0.24	0.85	7.84
Hamamelidaceae	47	5	0.11	42	7	0.15	12	0.26	0.84	10.64
Hyacinthaceae	260	3	0.01	257	15	0.06	18	0.07	0.65	1.15
Hydrangeaceae	14	1	0.07	13	1	0.07	2	0.14	0.14	7.14
Hypoxidaceae	47	4	0.09	43	6	0.13	10	0.21	0.75	8.51
Iridaceae	518	8	0.02	510	17	0.03	25	0.05	0.81	1.54
Isoetaceae	14	0	0.00	14	1	0.07	1	0.07	0.00	0.00
Juglandaceae	20	0	0.00	20	3	0.15	3	0.15	0.48	0.00
Juncaceae	90	4	0.04	86	3	0.03	7	0.08	0.50	4.44
Lamiaceae	707	8	0.01	699	17	0.02	25	0.04	0.87	1.13
Lardizabalaceae	18	4	0.22	14	4	0.22	8	0.44	0.87	22.22
Lauraceae	56	1	0.02	55	4	0.07	5	0.09	0.40	1.79
Lecythidaceae	110	2	0.02	108	11	0.10	13	0.12	0.86	1.82

Lentibulariaceae	18	3	0.17	15	4	0.22	7	0.39	0.84	16.67
Liliaceae	160	14	0.09	146	18	0.11	32	0.20	0.84	8.75
Limnanthaceae	19	0	0.00	19	2	0.11	2	0.11	0.35	0.00
Linaceae	59	0	0.00	59	9	0.15	9	0.15	0.84	0.00
Loasaceae	119	1	0.01	118	13	0.11	14	0.12	0.86	0.84
Loganiaceae	19	3	0.16	16	2	0.11	5	0.26	0.58	15.79
Loranthaceae	79	4	0.05	75	11	0.14	15	0.19	0.80	5.06
Lowiaceae	15	1	0.07	14	1	0.07	2	0.13	0.13	6.67
Lycopodiaceae	39	3	0.08	36	2	0.05	5	0.13	0.55	7.69
Lygodiaceae	14	0	0.00	14	1	0.07	1	0.07	0.00	0.00
Lythraceae	39	3	0.08	36	4	0.10	7	0.18	0.76	7.69
Magnoliaceae	48	1	0.02	47	2	0.04	3	0.06	0.12	2.08
Malpighiaceae	75	3	0.04	72	4	0.05	7	0.09	0.52	4.00
Malvaceae	214	8	0.04	206	19	0.09	27	0.13	0.92	3.74
Marantaceae	87	5	0.06	82	5	0.06	10	0.11	0.43	5.75
Maratticaceae	41	0	0.00	41	1	0.02	1	0.02	0.00	0.00
Marcgraviaceae	14	0	0.00	14	2	0.14	2	0.14	0.36	0.00
Melanthiaceae	83	3	0.04	80	10	0.12	13	0.16	0.83	3.61
Meliaceae	39	5	0.13	34	3	0.08	8	0.21	0.45	12.82
Melianthaceae	14	3	0.21	11	2	0.14	5	0.36	0.73	21.43
Menispermaceae	96	3	0.03	93	11	0.11	14	0.15	0.79	3.13
Moraceae	70	2	0.03	68	3	0.04	5	0.07	0.16	2.86
Musaceae	29	3	0.10	26	2	0.07	5	0.17	0.46	10.34
Myodocarpaceae	18	0	0.00	18	1	0.06	1	0.06	0.00	0.00
Myricaceae	29	1	0.03	28	1	0.03	2	0.07	0.07	3.45
Myristicaceae	14	1	0.07	13	1	0.07	2	0.14	0.14	7.14
Myrtaceae	123	2	0.02	121	6	0.05	8	0.07	0.49	1.63
Nartheciaceae	25	0	0.00	25	6	0.24	6	0.24	0.74	0.00
Nepenthaceae	11	1	0.09	10	1	0.09	2	0.18	0.18	9.09
Nothogaceae	12	0	0.00	12	1	0.08	1	0.08	0.00	0.00
Nymphaeaceae	53	2	0.04	51	1	0.02	3	0.06	0.07	3.77
Ochnaceae	11	1	0.09	10	2	0.18	3	0.27	0.56	9.09
Oleaceae	135	3	0.02	132	3	0.02	6	0.04	0.40	2.22
Onagraceae	200	6	0.03	194	9	0.04	15	0.08	0.78	3.00
Orchidaceae	1538	12	0.01	1526	34	0.02	46	0.03	0.76	0.78
Orobanchaceae	320	2	0.01	318	14	0.04	16	0.05	0.86	0.63
Osmundaceae	15	1	0.07	14	2	0.13	3	0.20	0.45	6.67
Oxalidaceae	246	6	0.02	240	4	0.02	10	0.04	0.14	2.44
Pandanaceae	36	3	0.08	33	3	0.08	6	0.17	0.61	8.33
Papaveraceae	115	8	0.07	107	7	0.06	15	0.13	0.78	6.96
Passifloraceae	170	6	0.04	164	12	0.07	18	0.11	0.78	3.53
Pentaphylacaceae	39	2	0.05	37	1	0.03	3	0.08	0.10	5.13
Phrymaceae	119	5	0.04	114	11	0.09	16	0.13	0.88	4.20

Piperaceae	61	3	0.05	58	2	0.03	5	0.08	0.24	4.92
Pittosporaceae	34	1	0.03	33	2	0.06	3	0.09	0.49	2.94
Plantaginaceae	451	5	0.01	446	12	0.03	17	0.04	0.79	1.11
Plumbaginaceae	26	4	0.15	22	7	0.27	11	0.42	0.92	15.38
Poaceae	2078	8	0.00	2070	23	0.01	31	0.01	0.74	0.38
Podocarpaceae	102	1	0.01	101	1	0.01	2	0.02	0.02	0.98
Podostemaceae	35	10	0.29	25	8	0.23	18	0.51	0.94	28.57
Polemoniaceae	219	4	0.02	215	11	0.05	15	0.07	0.86	1.83
Polygalaceae	275	5	0.02	270	21	0.08	26	0.09	0.87	1.82
Polygonaceae	215	3	0.01	212	8	0.04	11	0.05	0.80	1.40
Potamogetonaceae	49	3	0.06	46	3	0.06	6	0.12	0.27	6.12
Primulaceae	222	5	0.02	217	15	0.07	20	0.09	0.76	2.25
Proteaceae	156	3	0.02	153	3	0.02	6	0.04	0.11	1.92
Pteridaceae	11	1	0.09	10	1	0.09	2	0.18	0.18	9.09
Ranunculaceae	343	5	0.01	338	25	0.07	30	0.09	0.88	1.46
Restionaceae	217	4	0.02	213	6	0.03	10	0.05	0.49	1.84
Rhamnaceae	159	3	0.02	156	9	0.06	12	0.08	0.41	1.89
Rosaceae	766	9	0.01	757	11	0.01	20	0.03	0.67	1.17
Rubiaceae	1427	2	0.00	1425	22	0.02	24	0.02	0.75	0.14
Rutaceae	186	6	0.03	180	10	0.05	16	0.09	0.81	3.23
Salicaceae	124	2	0.02	122	5	0.04	7	0.06	0.52	1.61
Sapindaceae	207	1	0.00	206	1	0.00	2	0.01	0.01	0.48
Sapotaceae	84	0	0.00	84	3	0.04	3	0.04	0.22	0.00
Saxifragaceae	105	7	0.07	98	14	0.13	21	0.20	0.86	6.67
Schisandraceae	33	1	0.03	32	3	0.09	4	0.12	0.70	3.03
Scrophulariaceae	198	4	0.02	194	9	0.05	13	0.07	0.48	2.02
Simaroubaceae	15	4	0.27	11	2	0.13	6	0.40	0.70	26.67
Smilacaceae	12	1	0.08	11	1	0.08	2	0.17	0.17	8.33
Solanaceae	721	3	0.00	718	8	0.01	11	0.02	0.26	0.42
Stachyuraceae	18	1	0.06	17	1	0.06	2	0.11	0.11	5.56
Stilbaceae	15	2	0.13	13	4	0.27	6	0.40	0.84	13.33
Styracaceae	22	2	0.09	20	2	0.09	4	0.18	0.54	9.09
Symplocaceae	84	2	0.02	82	3	0.04	5	0.06	0.14	2.38
Taxaceae	12	1	0.08	11	3	0.25	4	0.33	0.71	8.33
Theaceae	82	1	0.01	81	3	0.04	4	0.05	0.40	1.22
Thesiaceae	57	2	0.04	55	3	0.05	5	0.09	0.20	3.51
Thymelaeaceae	202	7	0.03	195	12	0.06	19	0.09	0.80	3.47
Tofieldiaceae	22	4	0.18	18	2	0.09	6	0.27	0.48	18.18
Typhaceae	17	0	0.00	17	2	0.12	2	0.12	0.53	0.00
Ulmaceae	12	1	0.08	11	1	0.08	2	0.17	0.17	8.33
Urticaceae	117	10	0.09	107	10	0.09	20	0.17	0.65	8.55
Velloziaceae	17	1	0.06	16	4	0.24	5	0.29	0.65	5.88
Verbenaceae	161	3	0.02	158	3	0.02	6	0.04	0.14	1.86

Violaceae	192	6	0.03	186	11	0.06	17	0.09	0.85	3.13
Vitaceae	300	14	0.05	286	16	0.05	30	0.10	0.84	4.67
Winteraceae	20	0	0.00	20	1	0.05	1	0.05	0.00	0.00
Xanthorrhoeaceae	45	4	0.09	41	8	0.18	12	0.27	0.89	8.89
Zamiaceae	55	4	0.07	51	1	0.02	5	0.09	0.14	7.27
Zingiberaceae	120	5	0.04	115	6	0.05	11	0.09	0.74	4.17
Zygophyllaceae	125	6	0.05	119	28	0.22	34	0.27	0.97	4.80

Supplementary Table S7. SPInDel profiles from concatenated *atpF-atpH*, *psbA-trnH*, *trnL* CG*, trnL* GH and *trnL* HD regions

Species	***atpF* F - *atpH* R**	***psbA* F - *trnH* R**	***trnL* CG**	***trnL* GH**	***trnL* HD**
Acer pseudoplatanus	159	439	120	94	423
Avena fatua	562	590	120	77	240
Beckmannia syzigachne	342	591	120	91	443
Betula pendula	620	378	120	99	266
Cardiocrinum giganteum	589	417	120	102	430
Capsella bursa pastoris	559	343	120	86	415
Eleusine indica	554	588	120	95	454
Hordeum bulbosum	548	595	120	85	467
Hordeum pusillum	554	595	120	85	474
Hordeum vulgare	556	595	120	85	478
Lolium perenne	563	593	120	91	453
Phyllostachys nigra	555	594	120	90	445
Phleum pratense	343	590	120	86	447
Picea abies	430	577	121	92	390
Picea jezoensis	430	577	121	92	390
Picea koraiensis	430	577	121	92	390
Picea mariana	430	562	121	92	390
Pinus sylvestris	417	605	121	83	390
Poa annua	343	590	120	91	446
Poa compressa	343	590	120	91	441
Phalaris arundinacea	551	575	120	91	453
Podocarpus macrophyllus	439	696	125	77	353
Silene latifolia	553	300	122	87	467
Silene vulgaris	554	355	122	85	480
Solanum lycopersicum	502	513	123	78	411
Solanum nigrum	502	498	123	78	411
Taxus baccata	336	576	122	79	362
Taxus canadensis	346	573	122	79	362
Thuja koraiensis	409	518	121	59	384
Thuja occidentalis	409	510	121	59	384
Torreya nucifera	437	574	121	79	377
Trisetum sibiricum	553	590	119	91	240
Tsuga canadensis	406	543	121	83	394
Tsuga sieboldii	406	559	121	83	394
Verbena urticifolia	479	325	123	81	390
Viola dissecta	637	418	121	81	387
Viola albida	634	425	121	91	381
Viola chaerophylloides	635	425	121	91	381
Number of different fragment lengths	27	29	6	16	24

Supplementary Table S8. Species and sequence lengths for the concatenated regions *atpF-atpH*, *psbA-trnH* and *trnL* GH.

Species	*atpF* F-*atpH* R	*psbA* F-*trnH* R	*trnL* G - H
Acer negundo	158	455	94
Acer pseudoplatanus	159	427	94
Acer saccharinum	158	472	100
Acer saccharum	157	457	93
Acer spicatum	158	428	94
Acorus calamus	597	441	90
Acorus gramineus	529	467	90
Agrostis hyemalis	565	595	96
Aira caryophyllea	566	589	91
Alopecurus aequalis	335	589	91
Alopecurus pratensis	335	589	86
Amentotaxus argotaenia	438	545	78
Amphicarpaea bracteata	547	235	89
Anthoxanthum nitens	552	589	91
Aralia racemosa	476	457	84
Arrhenatherum elatius	551	595	91
Arthraxon hispidus	573	587	90
Arundinella hirta	558	587	90
Asparagus cochinchinensis	569	567	91
Avena fatua	562	589	86
Beckmannia syzigachne	342	590	91
Betula pendula	620	377	99
Briza minor	557	589	91
Capillipedium assimile	590	588	90
Capsella bursa pastoris	559	342	86
Cardiocrinum giganteum var. yunnanense	589	416	102
Celastrus scandens	413	474	89
Cenchrus americanus	566	586	90
Chamaecyparis obtusa	398	497	78
Chamaecyparis pisifera	386	514	78
Cinna latifolia	342	582	91
Colchicum montanum	526	408	61
Cornus sericea	491	417	89
Corylus cornuta	624	444	99
Cunninghamia lanceolata	420	546	78
Cynodon dactylon	554	587	89
Dactylis glomerata	343	603	95
Datura stramonium	502	467	78
Echinochloa crus galli	575	594	90
Echinochloa crus galli var. crus galli	575	594	90
Echinochloa oryzicola	575	594	90
Eleusine indica	554	587	95
Elymus ciliaris	556	602	90
Elymus longearistatus	413	598	90
Elymus repens	564	602	90
Eragrostis curvula	554	588	90
Festuca ovina	331	589	91
Ficus benguetensis	547	363	88
Ficus benjamina	556	376	88
Ficus erecta	548	385	88
Ficus microcarpa	554	396	88
Ficus pumila	556	358	88
Ficus septica	550	365	88
Ficus stenophylla	547	375	88

Ficus variegata Blume, 1825	547	388	88
Frangula alnus	544	410	103
Galium aparine	418	284	70
Galium mollugo	407	278	70
Gloriosa modesta	535	270	61
Hamamelis virginiana	551	395	103
Holcus lanatus	548	588	91
Hordeum bogdanii	554	594	85
Hordeum brachyantherum subsp. Californicum	554	594	85
Hordeum bulbosum	548	594	85
Hordeum chilense	554	594	85
Hordeum comosum	554	594	85
Hordeum cordobense	554	494	85
Hordeum erectifolium	554	594	85
Hordeum euclaston	554	594	85
Hordeum flexuosum	554	594	85
Hordeum intercedens	558	601	85
Hordeum marinum subsp. gussoneanum	553	594	85
Hordeum marinum subsp. marinum	541	594	85
Hordeum muticum	554	594	85
Hordeum pubiflorum	554	594	85
Hordeum pusillum	554	594	85
Hordeum roshevitzii	554	594	85
Hordeum stenostachys	554	594	85
Hordeum vulgare subsp. spontaneum	556	594	85
Hordeum vulgare subsp. vulgare	556	594	85
Hyacinthoides non scripta	516	601	90
Isachne globosa	548	580	89
Juglans cinerea	593	238	88
Juniperus communis var. saxatilis	407	424	78
Juniperus rigida	406	422	78
Juniperus virginiana	410	487	78
Lagenaria siceraria	424	186	64
Linum perenne	599	389	85
Lolium multiflorum	563	592	91
Lolium perenne	563	592	91
Magnolia grandiflora	573	395	85
Metasequoia glyptostroboides	423	536	78
Milium effusum	343	588	90
Miscanthus sinensis	580	588	91
Onixotis triquetra	506	440	61
Oplismenus undulatifolius var. japonicus	581	587	90
Oryzopsis asperifolia	550	567	77
Ostrya virginiana	634	452	99
Panicum bisulcatum	570	590	90
Panicum dichotomiflorum	358	587	90
Paspalum dilatatum	337	585	89
Passiflora incarnata	624	316	102
Passiflora quadrangularis	645	318	102
Persicaria amphibia	510	231	68
Persicaria hydropiper	516	420	68
Persicaria maculosa	515	304	68
Phalaris arundinacea	551	574	91
Phleum pratense	343	589	86
Phragmites australis	497	587	90
Phyllostachys nigra var. henonis	555	593	90
Picea abies	430	576	92
Plantago major	454	310	79

Platycladus orientalis	409	504	79
Poa annua	343	589	96
Poa compressa	343	589	91
Podocarpus macrophyllus	439	694	77
Populus alba	526	292	107
Populus balsamifera	543	289	94
Populus tremuloides	500	292	106
Potamogeton natans	524	333	125
Rubus occidentalis	581	336	90
Salix babylonica	573	253	94
Sasa palmata	555	593	90
Sequoiadendron giganteum	425	553	78
Setaria viridis	566	587	90
Silene latifolia	553	299	87
Silene vulgaris	554	343	85
Solanum dulcamara	503	498	78
Solanum lycopersicum	502	512	78
Solanum nigrum	502	497	78
Sorghum halepense	579	588	88
Spodiopogon sibiricus	580	588	90
Taxus baccata	336	576	79
Thuja koraiensis	409	518	58
Thuja occidentalis	409	510	58
Torreya nucifera	437	574	79
Trifolium pratense	508	428	89
Trisetum sibiricum	553	589	91
Typha angustifolia	543	669	87
Typha latifolia	543	669	87
Verbena urticifolia	479	321	81
Veronica officinalis	454	356	81
Veronica serpyllifolia	454	380	81
Viburnum opulus	505	415	78
Viola albida var. takahashii	636	425	92
Viola biflora	639	396	90
Viola brevistipulata var. minor	635	332	90
Viola chaerophylloides	635	424	91
Viola diamantiaca	632	417	93
Viola dissecta	637	417	81
Viola lactiflora	631	420	93
Viola mandshurica	631	416	92
Viola orientalis	636	396	90
Viola patrinii	653	416	92
Viola phalacrocarpa	632	414	92
Viola philippica	630	416	92
Viola raddeana	667	393	92
Viola rossii	672	409	93
Viola selkirkii	641	418	92
Viola seoulensis	630	420	92
Viola tenuicornis	631	415	92
Viola tokubuchiana f. variegata	636	419	93
Viola tokubuchiana var. takedana	636	429	94
Viola variegata	631	421	92
Viola verecunda	651	409	91
Viola violacea	636	410	91
Viola yazawana	643	431	91
Vitis riparia	584	353	89
Zizania latifolia	555	590	90
Zoysia japonica	558	587	90
Number of different fragment lengths	94	97	29

Study 2
Design and evaluation of PCR primers for amplification of four chloroplast DNA regions in plants

(*Accepted in Conservation Genet Resour, 2016*)

Design and evaluation of PCR primers for amplification of four chloroplast DNA regions in plants

Chiara Santos[1,2] and Filipe Pereira[1*]

[1] Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Avenida General Norton de Matos, s/n, 4450-208 Matosinhos, Portugal

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

* Correspondence: fpereirapt@gmail.com

Abstract

The high genetic diversity of plants can be a problem when developing molecular methods that require conserved DNA sequences among species. Several chloroplast DNA (cpDNA) regions have been used for the identification of plant DNA from broad taxonomic groups, but many species fail to amplify due to genetic variation at primer-binding sites. Here, we evaluated the conservation degree of four chloroplast DNA (cpDNA) regions commonly used in plant investigations (*atpF-atpH, psbA-trnH, trn*L CD and *trn*L GH). We propose new conserved PCR primers for the study of the most common plant families, designed using consensus sequences obtained from 28 multiple sequences alignments with over 11,000 reference sequences. The new primers were able to amplify all target regions in representative samples from the seven families. The conserved genomic regions and PCR primers can be used in diverse areas of plant research, including DNA barcoding, molecular ecology, metagenomics or phylogeny.

Keywords: Plants, Conserved genomic regions, cpDNA, PCR primers

It has been shown that ‘universal’ PCR primers can be successfully employed in the detection of plant DNA from broad taxonomic groups (Taberlet, Gielly et al. 1991, Taberlet, Coissac et al. 2007, Hollingsworth, Graham et al. 2011). However, it is often difficult to accommodate in a single target region all genetic variation present in divergent plant lineages and many species fail to amplify by PCR. Here, we report a re-evaluation of PCR primers used for amplification of four cpDNA regions commonly used in plant investigations.

We started by downloading from NCBI Entrez Nucleotide database (http://www.ncbi.nlm.nih.gov) all available sequences of the cpDNA regions *atpF-atpH*, *psbA-trnH* and two regions of the *trn*L (UAA) intron named *trn*L CD and *trn*L GH. These regions were selected by having conserved domains (used as primers-binding sites) flanking variable regions, being commonly used in phylogenetic and population genetics studies (Hollingsworth, Graham et al. 2011). We then build 28 multiple sequence alignments as previously described (Pereira, Carneiro et al. 2010) using one sequence per species grouped in seven plant families: Asteraceae, Brassicaceae, Iridaceae, Orchidaceae, Poaceae, Rosaceae and Salicaceae (alignments available at http://plantaligdb.portugene.com/). The consensus sequences were extracted and aligned for each cpDNA region and the most conserved regions were used for primer design. Twenty-two different primers were designed taking into account that five of the eight potential primer-binding sites were found highly conserved across families, meaning that the same primer could be used in different families (Table 1; Fig. S1). The average number of pairwise matches across the positions of the alignment (pairwise identity) in the primer-binding sites was higher than 92.7% in all cases, with 32 cases reaching 100% (Table S1). Near half of the target regions (n=27) had a percentage of identical sites higher than 90%, meaning that the consensus primers represented most species in the alignment.

The set of primers was tested using two species of each of the seven plant families (Fig.1; Tables 1, S2). Total DNA was extracted from fresh leaves using an adaption of the CTAB protocol (Doyle 1987) followed by a standard phenol: chloroform protocol. The primers pairs (Tables 1, S1) were tested by singleplex PCR using 1µL of extracted DNA (20-100 ng) and the PCR conditions previously described (Gonçalves, Marks et al. 2015). The primers successfully amplified the target regions in all tested species of each family (Figure 1). The amplified products have the expected length in all samples when considering the reference sequences. The differences in the amplification efficiency observed for different target regions suggest that some polymorphisms may be affecting the binding of primers, although without abolishing the amplification.

Table 1. List of PCR primers for amplification of four chloroplast DNA regions in seven plant families.

Target region	Primer-binding site	Number of different primers	Primer names	5' - 3' sequence	Primer length (nt)	Predicted Tm (°C)	Number of families
atpF-atpH	atpF	3	cpDNAatpF_ABIRS_F	GGTATTAAACCCGAAACTCCC	21	59.5	Asteraceae, Brassicaceae, Iridaceae, Rosaceae, Salicaceae
			cpDNAatpF_Orc_F	GGTATTAAACTCGAAACTCCCAG	23	60.9	Orchidaceae
			cpDNAatpF_Poa_F	GGTATTAAGCCCGAAACTGCC	21	61.2	Poaceae
	atpH	6	cpDNAatpH_Ast_R	GCACTTTTATTTGCTAATCCTTTTG	25	59.2	Asteraceae
			cpDNAatpH_Bra_R	CGCTTTTATTTGCGAATCCTTTTG	24	60.3	Brassicaceae
			cpDNAatpH_R	GCACTTTTATTTGCGAATCCTTTTG	25	60.9	Iridaceae, Salicaceae
			cpDNAatpH_Orc_R	GCTCTTTTATTTGCAAATCCTTTTG	25	59.2	Orchidaceae
			cpDNAatpH_Poa_R	GCTTTTATTTGCGAACCCTTTTG	23	59.2	Poaceae
			cpDNAatpH_Ros_R	CTCTTTTATTTGCGAATCCCTTTG	24	60.3	Rosaceae
psbA-trnH	psbA	8	cpDNApsbA_Ast_F	GAAGCTCCATCTACAAATGGATA	23	59.2	Asteraceae
			cpDNApsbA_Bra_F	CTGCTGTTGAGGCTCCATC	19	59.5	Brassicaceae
			cpDNApsbA_Iri_F_I	GCTGCTGTCGAAGTTCCATC	20	60.5	Iridaceae
			cpDNApsbA_Iri_F_II	TTCCCTTTAGACCTAGCTGCT	21	59.5	Iridaceae
			cpDNApsbA_Orc_F	TTCCCTCTAGATCTAGCTTCTG	22	60.1	Orchidaceae
			cpDNApsbA_Poa_F	TAGCTGCTCTTGAAGTTCCATC	22	60.1	Poaceae
			cpDNApsbA_Ros_F	TAGCTGCTGTTGAAGTTCCATC	22	60.1	Rosaceae
			cpDNApsbA_Sal_F	TAGACCTAGCTGCTGTCGAAG	21	61.2	Salicaceae
	trnH	1	cpDNAtrnH_R	CCACTTGGCTACATCCGCC	19	61.6	Asteraceae, Brassicaceae, Iridaceae, Orchidaceae, Poaceae, Rosaceae, Salicaceae
trnL CD	trnL C	1	cpDNAtrnLC_F	CGAAATCGGTAGACGCTACG	20	60.5	Asteraceae, Brassicaceae, Iridaceae, Orchidaceae, Poaceae, Rosaceae, Salicaceae
	trnL D	1	cpDNAtrnLD_R	GGGGATAGAGGGACTTGAAC	20	60.5	Asteraceae, Brassicaceae, Iridaceae, Orchidaceae, Poaceae, Rosaceae, Salicaceae
trnL GH	trnL G	1	cpDNAtrnLG_F	GGGCAATCCTGAGCCAAATC	20	60.5	Asteraceae, Brassicaceae, Iridaceae, Orchidaceae, Poaceae, Rosaceae, Salicaceae
	trnL H	1	cpDNAtrnLH_R	CATCGAGTCTCTGCACCTATC	21	61.2	Asteraceae, Brassicaceae, Iridaceae, Orchidaceae, Poaceae, Rosaceae, Salicaceae
	Total	22					

An additional band was observed in the two Orchidaceae samples for the *atpF-atpH* region, without affecting the DNA detection (Fig. 1). The length of the amplified products in each target regions varied among samples because of insertion/deletion polymorphisms, frequent in cpDNA non-coding regions (Hamilton, Braverman et al. 2003, Yang and Wang 2007).

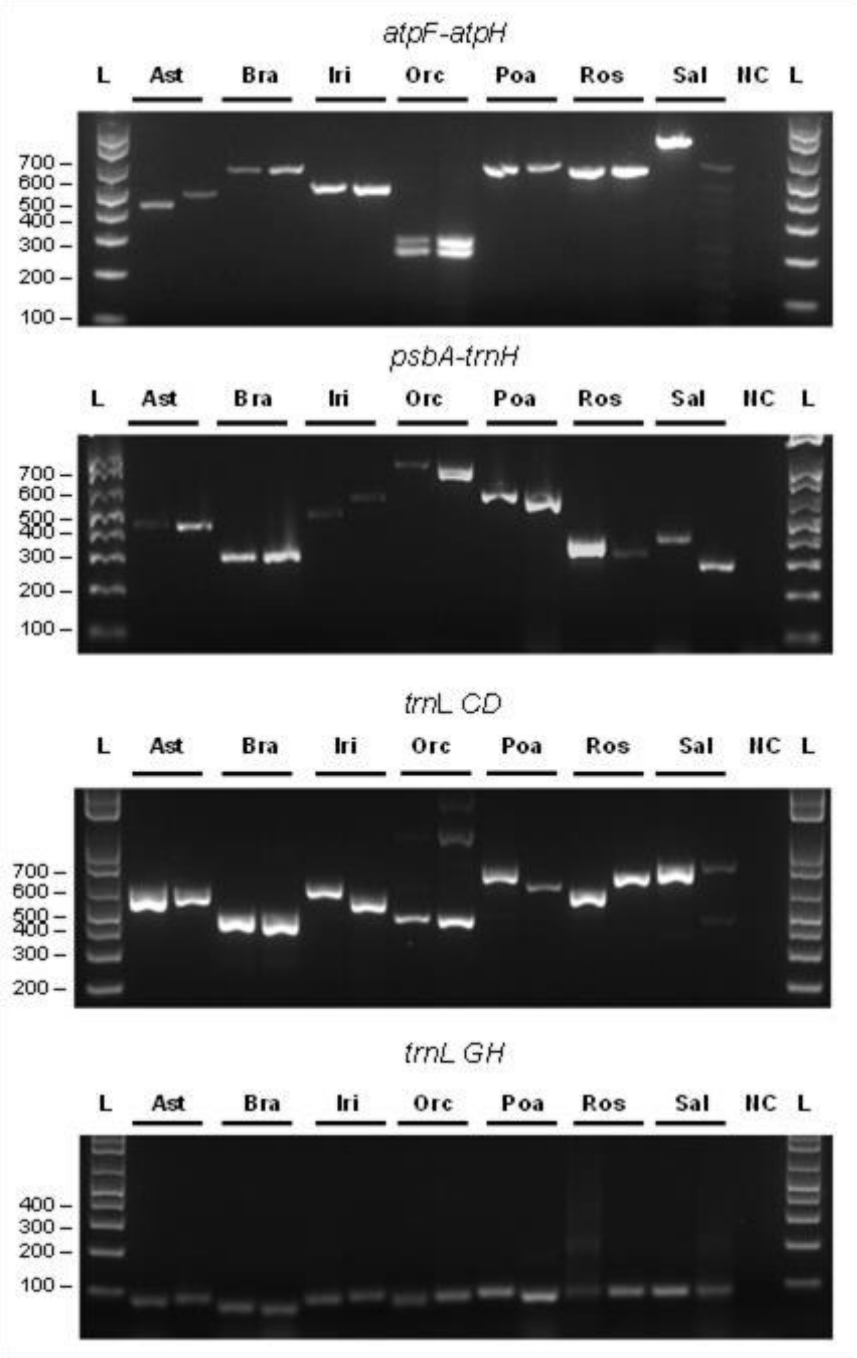

atpF-atpH
L Ast Bra Iri Orc Poa Ros Sal NC L
700 600 500 400 300 200 100
psbA-trnH
L Ast Bra Iri Orc Poa Ros Sal NC L
700 600 500 400 300 200 100
trnL CD
L Ast Bra Iri Orc Poa Ros Sal NC L
700 600 500 400 300 200
trnL GH
L Ast Bra Iri Orc Poa Ros Sal NC L
400 300 200 100

Fig. 1 Validation of the new conserved PCR primers for amplification of four chloroplast DNA (cpDNA) regions using two species of seven plant families: Asteraceae (Ast), Brassicaceae (Bra), Iridaceae (Iri), Orchidaceae (Orc), Poaceae (Poa), Rosaceae (Ros), and Salicaceae (Sal). *NC* negative control; *L*100-bp DNA ladder.

The conserved primers described here can be used to investigate the presence of these economically important plant families in varied types of samples, particularly in those where morphological characteristics are ambiguous. The targeting of high copy number cpDNA and the short length of some of the regions (e.g., *trn*L GH) facilitates the analysis of samples with low quality/quantity DNA, such as environmental samples, processed food products, animal gut contents, scats or forensic samples. Overall, the primers described here can facilitate the use of cpDNA markers to study a broad range of topics in plant biology.

Acknowledgements

The authors would like to thank Elisabete Couto and Luís Alves from 'Cantinho das Aromáticas' (www.cantinhodasaromaticas.pt), Paulo Almeida from Jardim Botânico da Universidade de Trás-os-Montes e Alto Douro (http://jb.utad.pt/) and Liliana Ramuge for providing samples. CS is supported by the "Programa Ciências Sem Fronteiras" reference BEX: 13186/13-2 from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil. FP is supported by the Portuguese Foundation for Science and Technology (FCT), European Regional Development Fund (ERDF) and Programa Operacional Potencial Humano. CIIMAR was partially supported by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT and ERDF in the framework of the programme PT2020.

Literature cited

Doyle, J. J. (1987). "A rapid DNA isolation procedure for small quantities of fresh leaf tissue." Phytochem. Bull. **19**: 11-15.

Gonçalves, J., et al. (2015). "A multiplex PCR assay for identification of the red fox (*Vulpes vulpes*) using the mitochondrial ribosomal RNA genes." Conservation Genetics Resources **7**(1): 45-48.

Hamilton, M. B., et al. (2003). "Patterns and relative rates of nucleotide and insertion/deletion evolution at six chloroplast intergenic regions in new world species of the Lecythidaceae." Mol Biol Evol **20**(10): 1710-1721.

Hollingsworth, P. M., et al. (2011). "Choosing and using a plant DNA barcode." PLoS One **6**(5): e19254.

Li, X., et al. (2015). "Plant DNA barcoding: from gene to genome." Biol Rev Camb Philos Soc **90**(1): 157-166.

Pereira, F., et al. (2010). "Identification of species by multiplex analysis of variable-length sequences." Nucleic Acids Res **38**(22): e203-e203.

Taberlet, P., et al. (2007). "Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding." Nucleic Acids Res **35**(3): e14.

Taberlet, P., et al. (1991). "Universal primers for amplification of three non-coding regions of chloroplast DNA." Plant molecular biology **17**(5): 1105-1109.

Yang, F.-S. and X.-Q. Wang (2007). "Extensive length variation in the cpDNA trnT-trnF region of hemiparasitic Pedicularis and its phylogenetic implications." Plant Systematics and Evolution **264**(3-4): 251-264.

Supplementary material

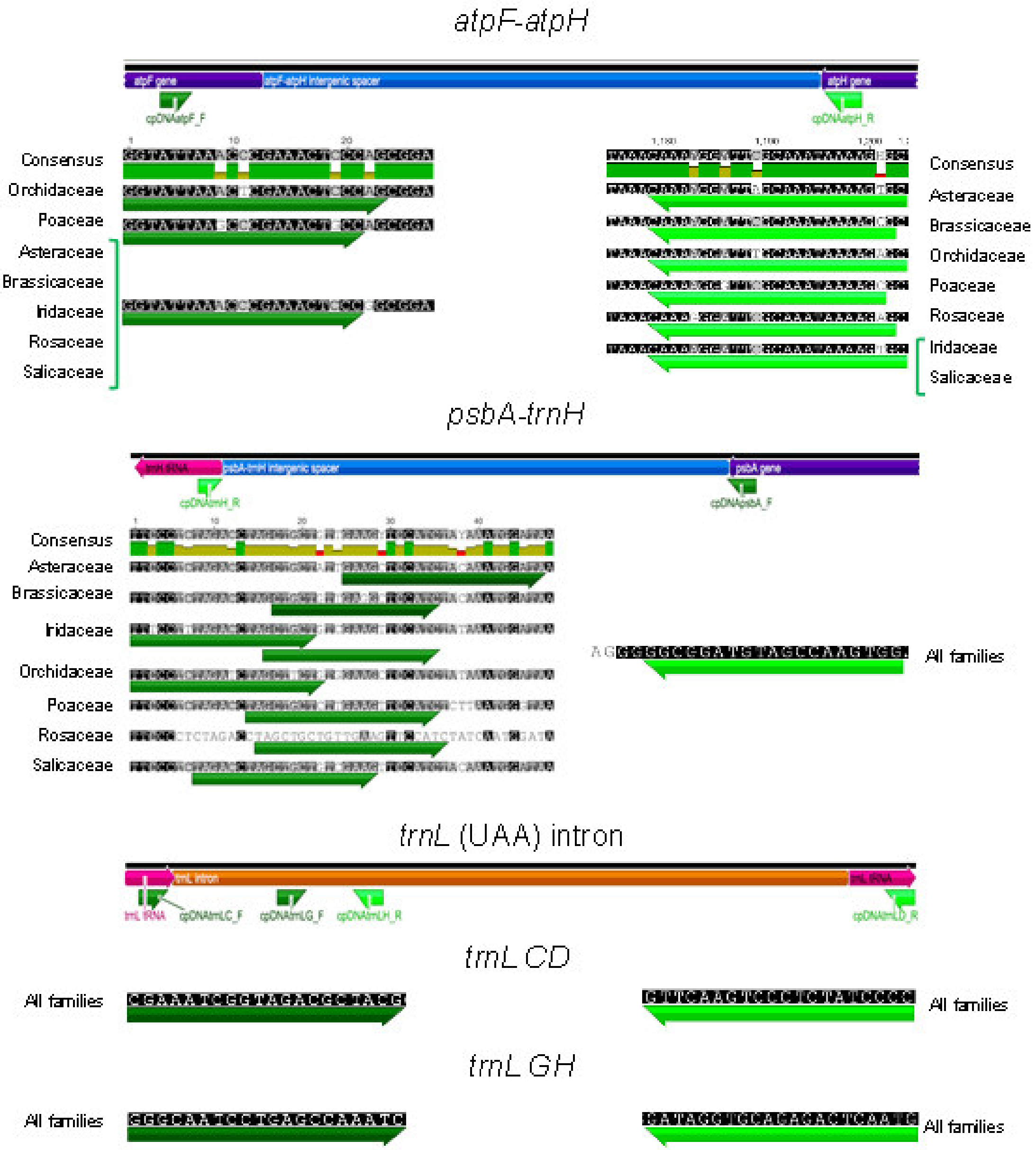
Supplementary table S1. List of conserved regions identified in the chloroplast DNA (cpDNA) of seven plant families. The table describes different measures of sequence conservation for 56 putative primer-binding sites (four cpDNA regions x two flanking regions x seven plant families). It should be noted that some of these primer-binding sites are equal in different families, meaning that 22 different primers are sufficient for 56 the binding regions.

Family	Target region	Primer-binding site	Number of species in alignment	Identical Sites	Pairwise Identity	Target region length			
						Minimum	Maximum	Mean	Range
Asteraceae	*atpF-atpH*	cpDNAatpF_F_Ast	63	21 (100%)	100%	424	509	478.7	85
		cpDNAatpH_R_Ast		21 (84%)	98%				
	psbA-trnH	cpDNApsbA_F_Ast	713	7 (30%)	99%	338	549	448	211
		cpDNAtrnH_R_Ast		9 (47%)	100%				
	trnL CD	cpDNAtrnLC_F_Ast	283	19 (95%)	100%	496	534	513.3	38
		cpDNAtrnLD_R_Ast		19 (95%)	100%				
	trnL GH	cpDNAtrnLG_F_Ast	1955	14 (70%)	100%	64	95	87.5	31
		cpDNAtrnLH_R_Ast		15 (71%)	100%				
Brassicaceae	*atpF-atpH*	cpDNAatpF_F_Bra	30	20 (95%)	98%	556	588	569.7	32
		cpDNAatpH_R_Bra		22 (92%)	99%				
	psbA-trnH	cpDNApsbA_F_Bra	81	9 (47%)	96%	215	437	310.3	222
		cpDNAtrnH_R_Bra		12 (63%)	98%				
	trnL CD	cpDNAtrnLC_F_Bra	176	19 (95%)	100%	383	596	405.9	213
		cpDNAtrnLD_R_Bra		16 (80%)	100%				
	trnL GH	cpDNAtrnLG_F_Bra	1236	15 (75%)	100%	62	94	84.2	32
		cpDNAtrnLH_R_Bra		13 (62%)	99%				
Iridaceae	*atpF-atpH*	cpDNAatpF_F_Iri	30	21 (100%)	100%	313	503	477.5	190
		cpDNAatpH_R_Iri		24 (96%)	98%				
	psbA-trnH	cpDNApsbA_F_Iri_I	282	9 (45%)	99%	560	604	578.6	44
		cpDNAtrnH_R_Iri		18 (95%)	100%				
	psbA-trnH	cpDNApsbA_F_Iri_II		13 (62%)	96%	575	619	593.6	44
		cpDNAtrnH_R_Iri		18 (95%)	100%				
	trnL CD	cpDNAtrnLC_F_Iri	15	20 (100%)	100%	547	586	570.7	39
		cpDNAtrnLD_R_Iri		20 (100%)	100%				
	trnL GH	cpDNAtrnLG_F_Iri	518	17 (85%)	100%	72	100	86.4	28
		cpDNAtrnLH_R_Iri		17 (81%)	99%				
Orchidaceae	*atpF-atpH*	cpDNAatpF_F_Orc	105	11 (48%)	93%	246	678	320.6	432

		cpDNAatpH_R_Orc		17 (68%)	98%				
	psbA-trnH	cpDNApsbA_F_Orc	84	11 (50%)	97%	717	1016	827.7	299
		cpDNAtrnH_R_Orc		14 (74%)	99%				
	trnL CD	cpDNAtrnLC_F_Orc	40	20 (100%)	100%	321	867	536.6	546
		cpDNAtrnLD_R_Orc		20 (100%)	100%				
	trnL GH	cpDNAtrnLG_F_Orc	1538	14 (70%)	100%	50	105	81.8	55
		cpDNAtrnLH_R_Orc		13 (63%)	98%				
Poaceae	*atpF-atpH*	cpDNAatpF_F_Poa	203	17 (81%)	97%	344	632	559.9	288
		cpDNAatpH_R_Poa		18 (78%)	99%				
	psbA-trnH	cpDNApsbA_F_Poa	212	15 (68%)	97%	575	660	594.0	85
		cpDNAtrnH_R_Poa		19 (100%)	100%				
	trnL CD	cpDNAtrnLC_F_Poa	397	18 (90%)	100%	396	671	600.2	275
		cpDNAtrnLD_R_Poa		18 (90%)	100%				
	trnL GH	cpDNAtrnLG_F_Poa	2078	15 (75%)	100%	59	109	89.7	50
		cpDNAtrnLH_R_Poa		17 (81%)	100%				
Rosaceae	*atpF-atpH*	cpDNAatpF_F_Ros	34	17 (81%)	96%	576	642	600.8	66
		cpDNAatpH_R_Ros		19 (79%)	95%				
	psbA-trnH	cpDNApsbA_F_Ros	296	13 (59%)	97%	234	589	346.4	355
		cpDNAtrnH_R_Ros		12 (63%)	100%				
	trnL CD	cpDNAtrnLC_F_Ros	21	20 (100%)	100%	494	646	588.3	152
		cpDNAtrnLD_R_Ros		14 (70%)	97%				
	trnL GH	cpDNAtrnLG_F_Ros	766	20 (100%)	100%	76	104	89.0	28
		cpDNAtrnLH_R_Ros		17 (81%)	100%				
Salicaceae	*atpF-atpH*	cpDNAatpF_F_Sal	13	21 (100%)	100%	517	654	574.3	137
		cpDNAatpH_R_Sal		21 (100%)	100%				
	psbA-trnH	cpDNApsbA_F_Sal	22	20 (95%)	99%	196	412	295.3	216
		cpDNAtrnH_R_Sal		19 (100%)	100%				
	trnL CD	cpDNAtrnLC_F_Sal	17	20 (100%)	100%	616	672	655.0	56
		cpDNAtrnLD_R_Sal		20 (100%)	100%				
	trnL GH	cpDNAtrnLG_F_Sal	124	20 (100%)	100%	93	111	96.9	18
		cpDNAtrnLH_R_Sal		19 (91%)	99%				

Supplementary table S2. List of samples used in this work.

Species	**Family**
Chrysanthemum leucanthemum	Asteraceae
Achillea ageratum	Asteraceae
Brasica rapa	Brassicaceae
Brassica oleracea	Brassicaceae
Crocus sativus	Iridaceae
Iris germa nica	Iridaceae
Orchis sp.	Orchidaceae
Phalaenopsis sp.	Orchidaceae
Zoysia sp.	Poaceae
Cymbopogon citratus	Poaceae
Fragaria sp.	Rosaceae
Malus sp.	Rosaceae
Salix babylonica	Salicaceae
Salix atrocinerea	Salicaceae



Supplementary figure S1. Schematic representation of chloroplast DNA (cpDNA) regions [*atpF-atpH*, *psbA-trnH* and *trnL* (UAA) intron] analysed in this work. The green arrow indicate the conserved regions where PCR primers were designed.

Study 3
PlantAligDB: A Database of Nucleotide Sequence Alignment for Plants

(*Submitted for publication*)

PlantAligDB: A Database of Nucleotide Sequence Alignments for Plants

Chiara Santos[1,2], João Carneiro[1,2] and Filipe Pereira[1*]

[1] Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Terminal de Cruzeiros do Porto de Leixões; Avenida General Norton de Matos, S/N 4450-208 Matosinhos – Portugal

[2] These authors contributed equally to this work.

[*]Corresponding author:

F. Pereira; E-mail fpereirapt@gmail.com; Tel (+351) 22 340 18 05; Fax (+351) 223390608.

Abstract

In recent years, a large number of nucleotide sequences have become available for plant species by the advent of massive parallel sequencing. The use of genomic data has been important for agriculture, food science, medicine or ecology. Despite the increasing amount of data, nucleotide sequences are usually available in public databases as isolated records with some descriptive information. Researchers interested in studying a specific plant family are forced to do multiple searches, sequence downloads, data curation and sequence alignments. This process is time-consuming and requires expensive computational resources and knowledge. In order to help researches overcoming these problems, we have built a comprehensive on-line resource of curated nucleotide sequence alignments for plant research, named PlantAligDB (available at http://plantaligdb. portugene.com). The latest release incorporates 514 alignments with a total of 66,052 sequences from six important genomic regions: *atpF-atpH*, *psbA-trnH*, *trn*L, *rbcL, matK* and ITS. The alignments represent 223 plant families from a variety of taxonomic groups. The users can quickly search the database, download and visualize the curated alignments and phylogenetic trees using dynamic browser-based applications. Different measures of genetic diversity are also available for each plant family. Overall, the PlantAligDB provides a complete, quality checked and regularly updated collection of alignments that can be used in taxonomic, DNA barcoding, molecular genetics, phylogenetic and evolutionary studies.

Keywords: DNA sequences, Multiple sequence alignments, Plant families

Introduction

The recent development of high-throughput sequencing technologies has increased significantly the number of nucleotide sequences available in public databases (Feuillet, Leach et al. 2011, Egan, Schlueter et al. 2012). Complete genome sequences are now accessible in public databases (e.g., EnsemblPlants) for the analysis and visualisation of genomic data for an ever-growing number of plants, such as *Beta vulgaris*, *Prunus persica* and *Citrus sinensis*, among many others. Sequences from individual genes or gene regions have also been deposit in public databases as a result of international initiatives. For instance, the DNA barcoding project has released thousands of sequences aiming at species identification and taxonomic classification of plants, mostly from the chloroplast DNA (cpDNA) protein-coding genes *rbcL* and *matK* (Group, Hollingsworth et al. 2009, Hollingsworth, Graham et al. 2011). The plastid *trn*L (UAA) intron is another good example of a cpDNA region highly represented in sequence databanks (Taberlet, Coissac et al. 2007).

Several web-based databases are available for plant genome sequences, usually dedicated to a single species or a genomic feature [e.g., (Meyer, Nagel et al. 2005, Lai, Berkman et al. 2012, Sakai, Lee et al. 2013, Numa and Itoh 2014)]. However, most nucleotide sequences are accessible in public databases as isolated records with simple descriptive information (taxonomy, geography, publications, etc.). For instance, the NCBI Entrez Nucleotide database (http://www.ncbi.nlm. nih.gov) and the BOLD - The Barcode of Life Data System (www.barcodinglife.org) (Ratnasingham and Hebert 2007) are useful repositories with descriptive information for sequence or species. Nevertheless, researchers interested in studying a specific plant family are forced to do multiple searches and sequence downloads of genetic data for their investigations. Moreover, the available sequences are not aligned and researches are forced to do their own alignments. The multiple sequence alignment step is critical because it determines the accuracy of the subsequence analyses, such as phylogenetic inference, identification of conserved motifs, function prediction, etc. Building accurate sequence alignments involves many steps, including the conversion of sequence files, running alignment algorithms in local computers or webservers, selection of best alignment parameters, and manual fine-tuning of the alignment. This process is laborious and requires costly computational resources, which are not always available.

We describe here an on-line database (PlantAligDB, available at http://plantaligdb.portugene.com) with a comprehensive, manually curated and regularly updated collection of alignments from diverse plant families (Figure 1). The PlantAligDB can help researchers designing accurate methods for plant identification (*matK* and *rbcL* are used in DNA barcoding projects) whether by identification of conserved motives that enable the design of primers or as a reference database for phylogenetic studies, allowing the construction of reference phylogenetic trees [e.g., genomic regions *atpF*-atpH (Domenech and Alapetite 2014), *psbA-trnH* (Dong, Liu et al. 2012) and ITS (Karehed, Groeninckx et al. 2008)]. Moreover, it provides useful data to understand the genetic diversity of the selected genomic regions.

Data curation

We retrieved all nucleotide sequences of different genomic regions from the NCBI Entrez Nucleotide database (http://www.ncbi.nlm.nih.gov) using the Geneious software (Drummond AJ 2009). Different combinations of search terms (e.g. '*gene name*'; viridiplantae'; 'chloroplast'; 'gene'; 'complete') and a maximum limit of 5000 bp as sequence length were used in searches to retrieve the largest number of sequences. After a preliminary curation of the data, we selected five cpDNA regions and one nuclear DNA region, which were the most represented in the NCBI database, commonly used in phylogenetic studies for being relevant and informative. The six genomic regions were named according to the gene regions where they are located: *atpF-atpH (*ATPase I subunit – ATPase III subunit*)*, *psbA-trnH* [Photosystem II 32 kDa protein – tRNA-His (GUG)], *trn*L [*tRNA*-Leu (UAA)], *rbcL* (rubisco large subunit), *matK* (maturase K) and ITS (internal transcribed spacer). We then removed from the datasets all redundant sequences belonging to the same species and sequences without a clear species assignment. We also reverse complement the sequences that were found in the opposite direction. The sequence orientation for each region is that of the most commonly found in the NCBI database. Therefore, the orientation of the *trnL (UAA)*, *atpF-atpH* and *rbcL* regions are the same of that used in the reference cpDNA sequence of *Nicotiana tabacum* (NC_001879.2), while the opposite orientation is used for regions *psbA-trnH* and *matK.* The target region named ITS in our database includes the internal transcribed spacer 1, 5.8S rRNA and internal transcribed spacer 2 section of the nuclear ribosomal DNA.

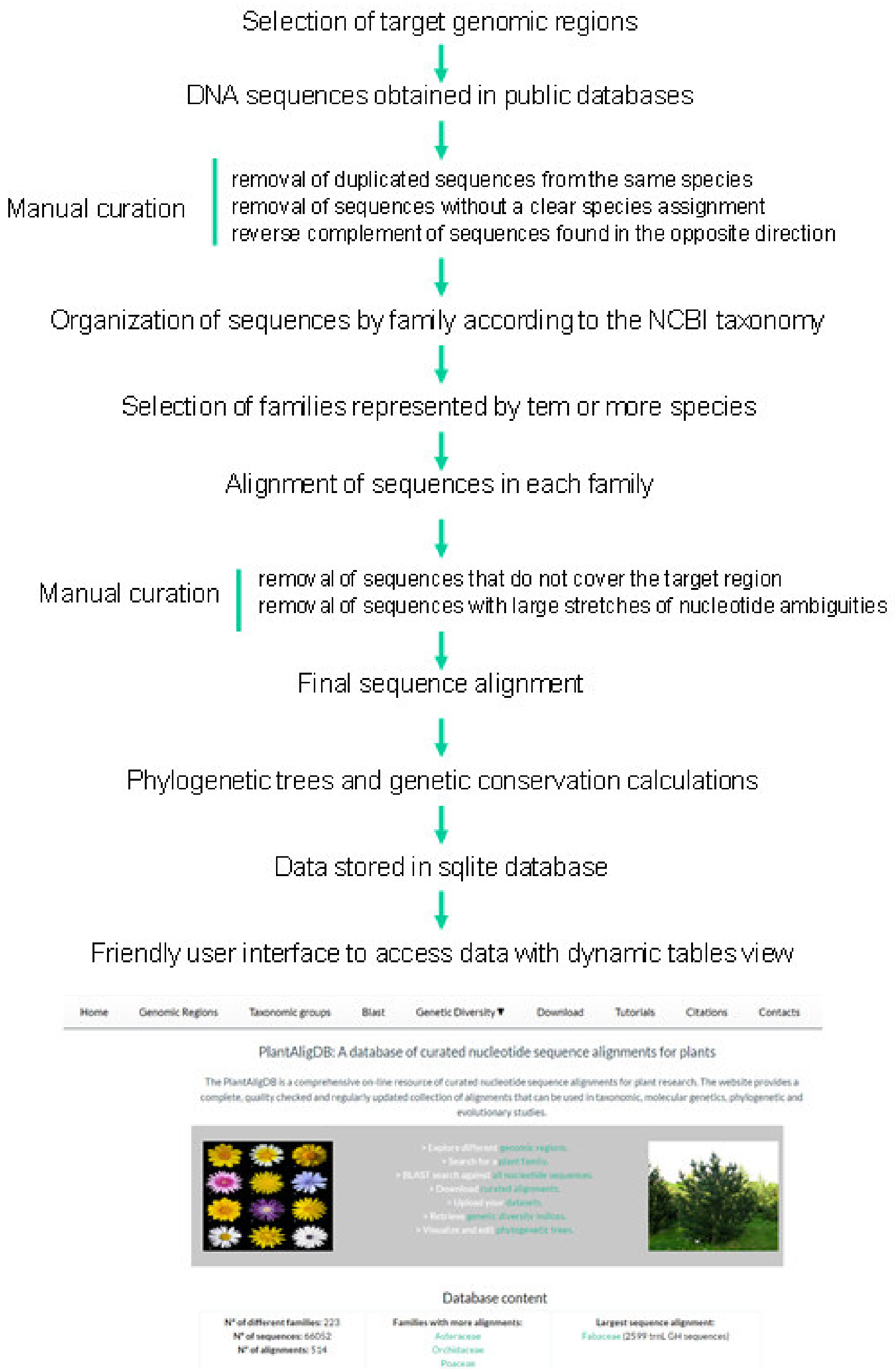
Because a high number of sequences were detected for the *trn*L (UAA) region (more than 50,000 hits), we used the external regions named "C" and "D" and the internal regions named "G" and "H" by (Taberlet, Coissac et al. 2007) as queries in the NCBI Basic Local Alignment Search Tool (BLAST; http://blast.ncbi.nlm.nih.gov/). The search was made against the nucleotide collection (nr/nt) of Tracheophyta (vascular plants) using the Biopython package (www.biopython.org) with an expected threshold of 1000 and a minimum word size of 16. Therefore, our database includes two datasets for the *trnL* (UAA) genomic region: the '*trnL* CD' target region with a length of 577 bp in *N. tabacum,* and the *'trnL* GH' with a length of 78 bp in *N. tabacum*, located inside the *trnL* CD region. All information regarding the selected target regions can be found in the *Genomic Regions* section of the PlantAligDB.

The nucleotide sequences of the six regions were organized by family according to the NCBI taxonomy and were aligned (each region and family in separated alignments) using the default parameters of the MUSCLE software (Edgar 2004) running in the Geneious software. The alignment was repeated in some families after excluding sequences that do not cover the entire region of interest and that had large stretches of nucleotide ambiguities. We only used alignments with ten or more species per family to build the PlantAligDB. Some species sequences were lost in this process filter. The neighbour joining phylogenetic tree of each region-family were calculated using Tamura-Nei model using Geneious Tree Builder. The methodology was built in Armadillo Workflow (http://www.bioinfo.uqam.ca/ armadillo/) to automate the update process of the database. The latest release update of June 2017 incorporates 514 alignments and phylogenetic trees, from 223 plant families.

Database organization

Basic structure

The PlantAligDB database is divided in nine sections (Figure 1): 1) *Home*, provides a brief description of what can be done in the database; 2) *Genomic regions*, describes the regions used in the database; 3) *Taxonomic groups*, the table containing the plant family/region alignments and phylogenetic trees; 4) *BLAST*, search the database using a query sequence by means of the BLASTN algorithm (Altschul, Gish et al. 1990); 5) *Genetic diversity*, describes the percentage of identical sites and pairwise identity values for each alignment; 6) *Download*, provides hyperlinks to download the curated alignments; 7) *Tutorials*, contains information about how the database was built, and how to use it; 8) *Citations*; 9) *Contacts*.


Selection of target genomic regions
DNA sequences obtained in public databases
Manual curation
removal of duplicated sequences from the same species
removal of sequences without a clear species assignment
reverse complement of sequences found in the opposite direction
Organization of sequences by family according to the NCBI taxonomy
Selection of families represented by tem or more species
Alignment of sequences in each family
Manual curation
removal of sequences that do not cover the target region
removal of sequences with large stretches of nucleotide ambiguities
Final sequence alignment
Phylogenetic trees and genetic conservation calculations
Data stored in sqlite database
Friendly user interface to access data with dynamic tables view
Home
Genomic Regions
Taxonomic groups
Blast
Genetic Diversity ▾
Download
Tutorials
Citations
Contacts

PlantAligDB: A database of curated nucleotide sequence alignments for plants

The PlantAligDB is a comprehensive on-line resource of curated nucleotide sequence alignments for plant research. The website provides a complete, quality checked and regularly updated collection of alignments that can be used in taxonomic, molecular genetics, phylogenetic and evolutionary studies.

- Explore different genomic regions.
- Search for a plant family.
- BLAST search against all nucleotide sequences.
- Download curated alignments.
- Upload your datasets.
- Retrieve genetic diversity indices.
- Visualize and edit phylogenetic trees.

Database content

N° of different families: 223
N° of sequences: 66052
N° of alignments: 514

Families with more alignments:
Asteraceae
Orchidaceae
Poaceae

Largest sequence alignment:
Fabaceae (2599 trnL GH sequences)

Figure 1. Workflow used to generate the curated alignments, phylogenetic trees, and genetic conservation values stored in PlantAligDB.

Taxonomic groups

The database is being regularly updated by our team and currently includes 514 alignments and phylogenetic trees from seven target regions: *atpF-atpH*, *psbA-trnH*, *trnL* CD, *trnL* GH, *rbcL*, *matK* and ITS (Table 1). Sequence alignments are provided for 223 different plant families. Currently, the *trnL* GH region has the largest number of sequences (n = 34,674). The *Fabaceae* family has the largest number of aligned species in a target region, with 2599 sequences for the *trnL* GH region. When considering all regions together, the *Fabaceae* (n = 4714), *Poaceae* (n = 4494) and *Asteraceae* (n = 4459) families are those with more sequences. The alignments for each plant family can be accessed through a dynamic table in the *Taxonomic groups* section of the database by following a hyperlink with the number of species included in each alignment (http://plantaligdb.portugene.com/cgi-bin/PlantAligDB_taxonomicgroups.cgi). The users are able to quickly search and locate a queried feature, order each column using the ascendant or descendent mode, filter the information, download the curated datasets, among other features. The multiple sequence alignment and phylogenetic tree can be visualized by clicking in the number of species present in the alignment.

Table 1. Summary of data currently available in the PlantAligDB.

Target region	**Genome**	**Type**	**Length (bp) in *Nicotiana tabacum***	**Number of alignments**	**Number of sequences**
atpF-atpH	cpDNA	Inter-genic spacer	502	31	1025
psbA-trnH	cpDNA	Inter-genic spacer	509	79	4852
***trnL* CD**	cpDNA	Intron	577	44	2527
***trnL* GH**	cpDNA	Intron	78	173	34674
rbcL	cpDNA	Protein-coding gene	1434	39	1748
matK	cpDNA	Protein-coding gene	1530	113	11341
ITS	nuDNA	Transcribed spacers and 5.8S gene	678	35	9885
			Total	514	66052

Genetic diversity

The database includes two measures of sequence conservation for each alignment: *percentage of identical sites* (PIS), calculated by dividing the number of identical positions in the alignment for an oligonucleotide by its length and the *percentage of pairwise identity* (PPI), calculated by counting the average number of pairwise matches across the positions of the alignment, divided by the total number of pairwise comparisons. Both sequence conservation measures are not intended to be used for comparison of different families and/or regions, since the number of sequences in each alignment can be very different. The PIS values in our current dataset vary from 0.16% to 99.07% (Table 2). The *matK* was the region with the lowest PIS value (0.16%) [Figure 2. f)], while the *trnL* GH was the region with the highest PIS value (99.07%), as can be seen in Figure 2 d) and Table 2. The *rbcL* was the most conserved region [Figure 2 e)] with an average of 78.48%, while the ITS was less conserved with an average of 28.84%

(Table 2). Our results are in accordance with earlier studies where *atpF-atpH* and *psbA-trnH* were found to be more variables than *matK* (Lahaye 2008). The *trnL* CD regions showed values slightly more conserved than *atpF-atpH* [Figure 2 c) and a)]. The lowest PPI value (69.96%) was found in *psbA-trnH* region [Figure 2 b)] and the highest was 100% in *trnL* GH, as shown in Figure 2 d) and Table 3. The ITS region was less conserved with an average of 87.21% [Table 3 and Figure 2 g)], while the *trnL* GH was the most conserved with an average of 97.09% [Table 3 d)].

Table 2. Average percentage of identical sites (PIS) values in all plant families organized by genomic region.

PIS	*atpF-atpH*	*psbA-trnH*	*trnL* CD	*trnL* GH	*rbcL*	*matK*	ITS
Mean	55.05	39.1	61.13	58.62	78.48	65	28.84
Max	85.95	97.42	96.46	99.07	95.24	97.3	78.45
Min	15.19	2.3	21	6.43	27.47	0.16	1.13

Table 3. Average percentage of pairwise identity (PPI) values in all plant families organized by genomic region.

PPI	*atpF-atpH*	*psbA-trnH*	*trnL* CD	*trnL* GH	*rbcL*	*matK*	ITS
Mean	95.49	94.02	96.34	97.09	96.94	95.23	87.21
Max	99.37	99.94	99.6	100	99.48	99.62	97.75
Min	87.59	69.96	90.54	88.54	81.18	78.72	70.79

Sequence alignments and phylogenetic trees

The alignments stored in the database can be visualized using a dynamic browser-based application named Wasabi (http://wasabiapp.org/) (Veidenberg, Medlar et al. 2016) with multiple options for the visualization and analysis of sequence data and phylogenetic trees. This resource is particularly useful to help researchers selecting the most appropriated genomic regions for their investigations. The users can zoom in and out the selected regions of the alignments, collapse regions with gaps, alternate between column and row selection, remove or add sequences, realign sequences, and export the sequence data in the FASTA format. If an Wasabi account is created, the user can re-align specific alignments with PAGAN (Loytynoja, Vilella et al. 2012) and PRANK (Löytynoja 2014). The user can merge different alignments from same region and different families using the PAGAN application. The download of the complete database of curated alignments is accessible in the *Download* section of the database.

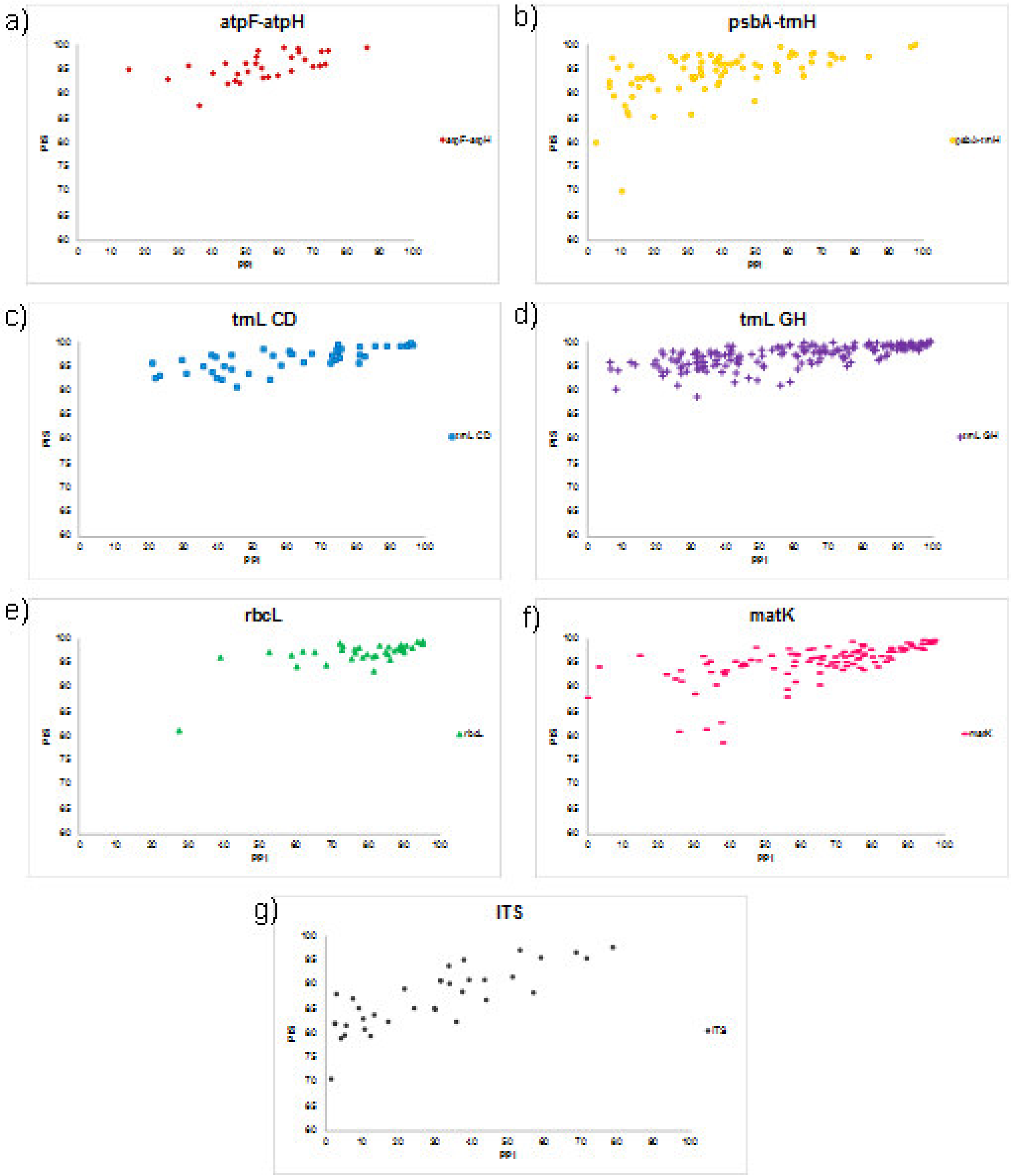


Figure 2. Graphic representation of the measures of sequence conservation PPI and PIS for each region-family alignment: a) *atpF-atpH* region, b) *psbA-trnH* region, c) *trnL* CD region, d) *trnL* GH region, e) *rbcL* region, f) *matK* region and g) ITS region.

How to use the PlantAligDB

To explore a genomic region, a researcher must start by accessing the ‘Genomic Regions’ section in the menu bar. For example, by selecting *psbA-trnH*, the user will find a brief description of the genomic region and the name, size and position of that region in the reference genome, and the families with available alignments. The user should select the ‘Taxonomic Groups’ section in the menu bar to search for a specific plant family. Then, through the search tool on the right side of the page, the user can type the name of the family and a dynamic table will be displayed with the number of sequences for each region. The user can also access a hyperlink with the description of the family and taxonomic tree using the resource Tree of Life Web Project (http://tolweb.org). Clicking on the number of sequences, the database is redirected to the Wasabi tool. The user can create a Wasabi account by providing an e-mail or choosing a temporary account, which allows to realign sequences with PAGAN or PRANK. The user can merge a PAGAN realignment with alignments of other families on the same region, by selecting a file on his local computer in the “alignment extension” option.

Availability and design

The PlantAligDB is freely available at http://plantaligdb.portugene.com and is optimized for the major web browsers (Internet Explorer, Firefox, Safari, and Chrome). The SQLite local database is used for data storage and runs on an Apache web server. The dynamic HTML pages were implemented using CGI-Perl and JavaScript and the dataset table views were generated using the JQuery plugin DataTables v1.9.4 (http://datatables.net/). The PlantAligDB visualization tables are generated automatically. The process of database update is optimized for large datasets. There are no access restrictions for academic and commercial use.

Literature cited

Altschul, S. F., et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Domenech, B. A.-L., C. B.; Baker, W. J.; and E. P. Alapetite, J.; and Nadot, S. (2014). "A phylogenetic analysis of palm subtribe Archontophoenicinae (Aracaceae) based on 14 DNA regions." Botanical Journal of the Linnean Society **175**: 469-481.

Dong, W., et al. (2012). "Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding." PLoS One **7**(4): e35071.

Drummond AJ, A. B., Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A (2009). "Geneious Pro 4.8.2.".

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Egan, A. N., et al. (2012). "Applications of next-generation sequencing in plant biology." Am J Bot **99**(2): 175-185.

Feuillet, C., et al. (2011). "Crop genome sequencing: lessons and rationales." Trends Plant Sci **16**(2): 77-88.

Group, C. P. W., et al. (2009). "A DNA barcode for land plants." Proceedings of the National Academy of Sciences **106**(31): 12794-12797.

Hollingsworth, P. M., et al. (2011). "Choosing and using a plant DNA barcode." PLoS One **6**(5): e19254.

Karehed, J., et al. (2008). "The phylogenetic utility of chloroplast and nuclear DNA markers and the phylogeny of the Rubiaceae tribe Spermacoceae." Mol Phylogenet Evol **49**(3): 843-866.

Lahaye, R. S., V.; Duthoit, S.; Maurin, O. and van der Bank, M. (2008). "A test of psbK-psbI and atpF-atpH as potential plant DNA barcodes using the flora of the Kruger National Park as a model system (South Africa)." Nature Precedings: 21.

Lai, K., et al. (2012). "WheatGenome.info: an integrated database and portal for wheat genome information." Plant Cell Physiol **53**(2): e2.

Loytynoja, A. (2014). "Phylogeny-aware alignment with PRANK." Methods Mol Biol **1079**: 155-170.

Loytynoja, A., et al. (2012). "Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm." Bioinformatics **28**(13): 1684-1691.

Meyer, S., et al. (2005). "PoMaMo--a comprehensive database for potato genome data." Nucleic Acids Res **33**(Database issue): D666-670.

Numa, H. and T. Itoh (2014). "MEGANTE: a web-based system for integrated plant genome annotation." Plant Cell Physiol **55**(1): e2.

Ratnasingham, S. and P. D. Hebert (2007). "bold: The Barcode of Life Data System (http://www.barcodinglife.org)." Mol Ecol Notes **7**(3): 355-364.

Sakai, H., et al. (2013). "Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics." Plant Cell Physiol **54**(2): e6.

Taberlet, P., et al. (2007). "Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding." Nucleic Acids Res **35**(3): e14.

Veidenberg, A., et al. (2016). "Wasabi: An Integrated Platform for Evolutionary Sequence Analysis and Data Visualization." Mol Biol Evol **33**(4): 1126-1130.

CHAPTER IV

DISCUSSION, CONCLUSIONS AND FUTURE PERSPECTIVES

DISCUSSION, CONCLUSIONS AND FUTURE PERSPECTIVES

Plants are used in the most diverse ways being extremely important resources. However, it is precisely its huge diversity and abundance that makes it difficult to characterize and identify these organisms. Modern techniques based on nucleic acids have been employed in the discrimination of plant species. Although the DNA barcoding aprroach is well established for animals using the COI region, has provided less reliable results in plants because of the lack of a single standard region. Failure in identifying plant species through DNA barcoding does not lie in the technique, but from the variability of the plants and the absence of a universally accepted concept of species, therefore, DNA barcoding is not the most appropriate technique to identify this taxonomic group (Moritz and Cicero 2004, Frézal and Leblois 2008).

In the first described work, we have tested the utility of the SPInDel method to identify plant species. The SPInDel approach is simple and intuitive, has already been used for identification of humans, common domestic animal and red fox (Gonçalves, Marks et al. 2015, Alves, Pereira et al. 2017). We have selected appropriate genomic regions for the SPInDel concept (hypervariable regions delimited by conserved segments) (Figure 1) by using the most coomonly described regions in previous works. The utility of each region depends on several factors, including the distance of the related species in the target group, the purpose of the research, the methodology to be used, among others (Semerikova and Semerikov 2014, Saddhe, Jamdade et al. 2017). We found that diverse plant species could be identified despide the use of the SPInDel method if the convenient hypervariable regions are chosen. Our approach analysed multiple loci at the same time, avoiding the wrong assignment of species due to missing data or unexpected allelic variants in a single region. This important advantage is highlighted in the mismatch distributions, where most of the profiles diverge by several fragment lengths. In the concatenation of *atpF-atpH*, *psbA-trnH* and *trnL* CD, 462 of the cases (66% of species) differ in all five hypervariable regions (*atpF-atpH*, *psbA-trnH*, *trnL* C-G, *trnL* G-H and *trnL* H-D). In the concatenation of *atpF-atpH*, *psbA-trnH* and *trnL* GH regions, 12988 of the cases (90% of species) differ for the three-hypervariable regions analysed [Study 1: Figure 4 b), Figure 5 b)]. In cases where one (or more) hypervariable region(s) have the same length for two species, or fail to amplify by intra-species polymorphisms, a correct identification is still possible based on the information from the remaining target regions. For example, *Hordeum bulbosum* and *H. pusillum* had the same length for *psbA* F – *trnH* R (595), *trnL* CG (120) and *trnL* GH (85) but were distinct for *atpF* F – *atpH* R (548 – 554) and *trnL* HD (467 – 474) (Study 1: Table S7). Previous investigations showed that equal SPInDel profiles in different species are very rare and that species from other classes have very different electrophoretic profiles (Gonçalves, Marks et al. 2015). The use of multiple loci also decreases the likelihood of false negatives because the probability of three target regions all failing to amplify by PCR due to polymorphisms is low, and certainly much lower than in methods using a single pair of PCR primers. The occurrence of false-positives caused by intra-species polymorphism is unlikely using our approach because most species diverge by several target regions (Study 1: Table S7, S8).

The Discrimination Power (DP) of each region was estimated in the SPInDel workbench (Carneiro, Pereira et al. 2012) (Figure 2), showing that three of the analysed regions had intermediate values of discrimination (Study 1: Table 2). However, the DP increases dramatically when two or more regions are combined (Study 1: Table 4). The frequency of species-different profiles (f_{dp}) also increases considerably from ~50% to <90%. The frequency of species-shared profiles (f_{sh}) was lower in the *trnL* CD and the *trnL* GH [Study 1: Figure 2 b), Table 2]. The DP of the SPInDel approach varies greatly in families with fewer than 500 individuals [Study 1: Figure 3 a)], and was linear with the increase of the frequency of the different profiles (f_{dp}), so that families with high values of f_{dp} also had a high DP [Study 1: Figure 3 b)]. Since the genomic region may be more or less effective for family-based identification, it is therefore important to have specific primers for families that amplify the highly informative regions. For example, we should take into account the seven plant families (Asteraceae, Brassicaceae, Iridaceae, Orchidaceae, Poaceae, Rosaceae and Salicaceae) represented in the four regions analysed (*atpF-atpH*, *psbA-trnH*, *trnL* CD and *trnL* GH) (Study 1: Figure S1). For Brassicaceae family, *atpF-atpH* region was more informative, the DP was 64.29%; for Iridaceae family the DP was higher in *trnL* CD (66.67%) region, for Salicaceae family, the region *atpF-atpH* was more informative with DP of 81.82% [Table 2 and Study 1: Figure 3 b), Tables S3, S4, S5, S6].

Table 2. The Discrimination power (%) of the seven plant families analysed for four target regions.

Family	*atpF-atpH*	*psbA-trnH*	*trnL* CD	*trnL* GH
Asteraceae	33.87	14.55	6.71	0.20
Brassicaceae	64.29	25.00	6.25	0.57
Iridaceae	36.36	14.29	66.67	1.54
Orchidaceae	52.38	6.65	55.00	0.78
Poaceae	11.60	3.06	15.37	0.38
Rosaceae	50.00	19.30	47.62	1.17
Salicaceae	81.82	63.64	64.71	1.61

There is no simple formula capable of predicting how many markers must be analysed to ensure the reliable identification of the species because the rates of molecular evolution vary between the different segments of the genome and through the taxa (Hebert, Cywinska et al. 2003, Narayan, Dodd et al. 2015). However, our concatenation process was very efficient. The combination of three regions of the cpDNA (*atpH-atpH*, *psbA-trnH* and *trnL* GH) distinguished more than 84% of the 170 species analysed (Study 1: Table 4 and S8). The frequency of shared profiles decreases (f_{sh}) drastically when regions are combined [Study 1: Figure 2 b)]. Separately, *atpF-atpH*, *psbA-trnH* and *trnL* CD showed f_{sh} of 0.16, 0.17 and 0.12 respectively (Study 1: Table 2), whereas when they were combined, the frequency of the shared profiles was reduced to 0.03 (Study 1: Table 4, Figure 2). Our results show that SPInDel achieved greater discrimination power when combining hypervariable regions than barcoding studies (Study 1: Table 5) (Hebert, Cywinska et al. 2003, Chase, Cowan et al. 2007, Kress and Erickson 2007, Sass, Little et al. 2007, Lahaye 2008) suggesting that our method can be a valuable tool for the plant species identification. Ran, Wang et al. (2010) combined the *atpF-atpH* and *psbA-trnH* regions with *psbK-psbI* and obtained a discrimination power of 60.71%. Fazekas, Burgess et al. (2008) combined these same regions for a larger number of samples and obtained 66% of discriminated species (Study 1: Table 5). Our concatenation of *atpF-atpH*, *psbA-trnH* and *trnL* CD distinguished more than 92% of the 38 species analysed [Study 1: Figure 2 a), Table 4, S7].

The concatenation of several individual markers improves the efficiency of plant species identification (Seberg and Petersen 2009, Dong, Liu et al. 2012). If the informative regions are combined correctly, high values of discrimination are reached with two or three markers [Study 1: Figure 4 a)]. In first study, we reviewed the discriminatory capacity of plant species identification obtained in other studies through the different combination of genomic regions (*atpF-atpH*, ITS, *matK*, *psbA-trnH*, *psbK-psbI*, *rbcL*, *rpoB*, *rpoC1* and *trnL*) to compare with the discriminatory power obtained in our combinations. We observed that the use of many markers, in addition to making the analyses more expensive and laborious, do not show any advantages, they can even reduce the number of correctly identified species. The greatest discriminating power is obtained by combining two or three markers, as indicated in our analyses (Figure 5 and Study 1: Table 5). Although our concatenation process proved to be superior to other combinations, other markers (e.g. *matK* and ITS) can be tested in the future to obtain greater discriminatory power.

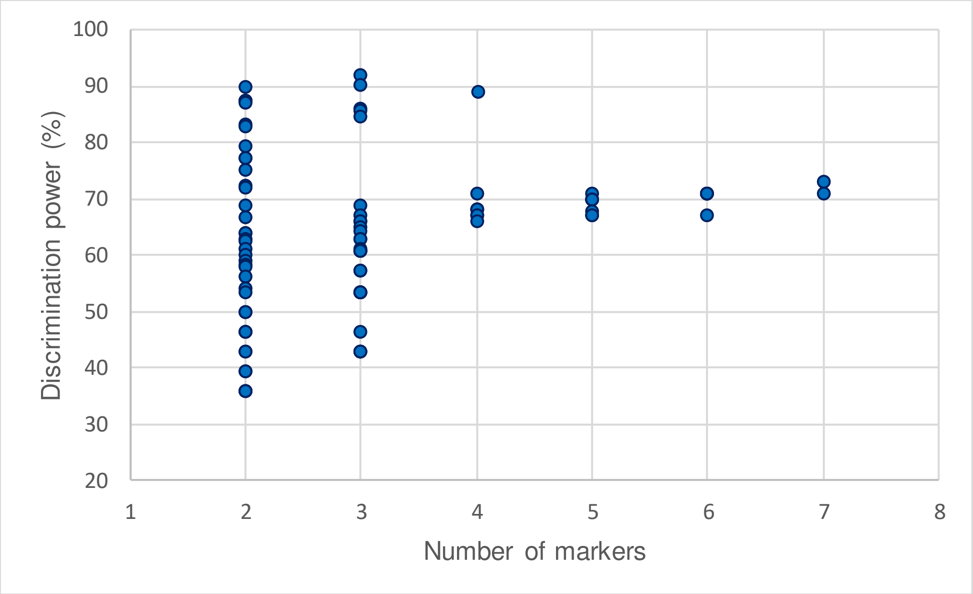


Figure 5. The discriminatory power (%) of different approaches for the identification of plant species by different combinations of markers (*atpF-atpH*, ITS, *matK*, *psbA-trnH*, *psbK-psbI*, *rbcL*, *rpoB*, *ropC1* and *trnL*).

The accurate identification of an organism depends on having a low intra species variation when compared with the one found between species. In any case, well-sampled data sets are still necessary to prove that intraspecific variation and interspecific divergence in cpDNA variable-length regions do not overlap for various taxonomic groups. Therefore, the intraspecific diversity of four genomic regions was analysed by aligning the largest number of sequences available for a set of 14 different species. Four species for *trnL* GH, *trnL* CD and *psbA-trnH*, and two species for *atpF-atpH* (Study 1: Figure S2). We found low intraspecific diversity in the dataset analysed. The *trnL* GH was the least divergent region as previously reported (Taberlet, Coissac et al. 2007, Tsai, Chiang et al. 2012). The lowest average values of intraspecific diversity were found in *psbA-trnH* (0.86%) and *trnL* GH (0.23%) (Study 1: Table 3). The range of intra- and interspecific diversities was previously analysed for the *psbA-trnH*, ITS2, *matK*, *rbcL*, *ycf5* and *rpoC1* markers, the authors concluded that only *psbA-trnH* and ITS2 have adequate gaps of intra- and interspecific variation (Chen, Yao et al. 2010). In Table 3, we describe the intraspecific diversity values found for the set of 14 species analysed and the interspecific diversity values found in the families to which the species of the intraspecific group belong (Study 1: Table 3 and Tables S3, S4, S5 and S6). The greatest range of variation was found in *trnL* CD for the family Caryophyllaceae (6.35 – 85.71%). The *psbA-trnH* region also show gaps of intermediate variation in the families (Table 3). The intraspecific diversity of other families with larger sample sizes or other genomic regions (e.g. *matK*, *rbcL* and ITS) should be determined in future studies.

Table 3. Intraspecific and interspecific diversity in some plant families for four cpDNA genomic regions.

Region	Family	Intraspecific diversity	Interspecific diversity
atpF-atpH	Acanthaceae	0.00	-
	Musaceae	5.56	8.33
psbA-trnH	Aceraceae	1.15	30.95
	Fabaceae	1.15	20.95
	Poaceae	0.00	3.06
	Rosaceae	1.13	19.30
***trnL* CD**	Brassicaceae	0.00	6.25
	Caryophyllaceae	6.35	85.71
	Moraceae	43.75	-
	Poaceae	4.00	15.37
***trnL* GH**	Brassicaceae	0.00	0.57
	Ranunculaceae	0.93	1.46
	Rubiaceae	0.00	0.14
	Salicaceae	0.00	1.61

We have shown that plant species can be conveniently and inexpensively identified through the SPInDel approach. Our method can easily be replicated in other laboratories thanks to the standardized methodology that allows the comparison of results. It has already been shown that the SPInDel concept is suitable for the identification of processed and mixed samples (Pereira, Carneiro et al. 2010, Alves, Pereira et al. 2017). Therefore, as future work, we suggest the use of samples such as teas, flours or herbal products. The cpDNA is present in several copies in each cell which represents an advantage over methods based on the nuclear genome. Our system analyses a short region with less than 150 bp (*trnL* GH), which can be useful in analysing samples with small amounts and/or degraded DNA.

In the first study described in this dissertation, we identify conserved regions that serve as anchors in the SPInDel approach (Figure 2). These short conserved regions can be used as primers binding sites that allow the amplification of hypervariable regions of cpDNA. Therefore, we developed an effective set of specific primers carefully designed and tested for relevant groups of plants. The set of primers developed by us

(Study 2: Table 1) will facilitate the work of researchers focusing on the most commonly used plant families. The families selected by us in study 2 are briefly described next.

Asteraceae is aimportant family of plants, with more than 23,000 species distributed globally. Mostly members in are herbaceous, but some species displayed high variation in morphological, physiological, and biochemical traits. Asteraceae has a notable economic and ecological significance, because includes important members like chicory, lettuce, sunflower, artichoke and calendula (Timme, Kuehl et al. 2007, Curci, De Paola et al. 2015, Wang, Cui et al. 2015). Brassicaceae (or Cruciferae) is a plant family with nearly 3700 species that has a taxonomy long been controversial because the generic boundaries between species are often poorly delimited. In addition to the model organism *Arabidopsis thaliana* that belongs to the Brassicaceae family important species such as cabbage, broccoli, turnip and mustard are also part of this group (Franzke, Lysak et al. 2011). Iridaceae is a large and diverse family that displays an unusually wide range of leaf anatomical characters, and this is consistent with its morphological diversity (Rudall 1994). *Crocus sativus*, commonly known as saffron, belongs to Iridaceae family, known for its aroma, colour and medicinal properties and is regarded as the most costly spice in the world (Hussain, Haq et al. 2014). Orchidaceae is one of the largest and most diverse Angiosperms family, with more than 20,000 species, some of them with quite commercial interest (Cafasso, Widmer et al. 2004, Su, Chao et al. 2013). Most orchids genomes are large size and complexes which tend to hamper genomic approaches (Chase, Cameron et al. 2003). Poaceae (grass family) has particular morphological and anatomical characteristics that generate controversy as to the phylogenetic origin of the species. Extremely important species such as rice, corn, oats, wheat, rye, barley and bamboo belong to the Poaceae family (Doyle, Davis et al. 1992). Rosaceae is a diverse family with about 3000 species, among them important fruit-producing crops: apple, pear, raspberries/blackberries, strawberries and stone fruits such as peach/nectarine, apricot, plum, cherry and almond. Rosaceae also contains a wide variety of ornamental plants including roses, flowering cherry, crab-apple, quince and *Prunus* genera used to wood production (Jung, Staton et al. 2007, Khan and Shinwari 2016). The Salicaceae family includes fast-growing hardwood species, important to forest industry like willow, aspen, cottonwood and poplar (Devantier, Moffatt et al. 1993). The number of species for this family is uncertainty due widely distributed of some individuals hampering access to material and because of the widespread hybridisation and great polymorphism of many species, which makes it difficult to find taxonomically reliable characters for species identification (Karrenberg, Edwards et al. 2002).

Faced with diversity and importance, the molecular identification tools of these families become necessary. In the third study described in this dissertation, we constructed a database that intuitively collects nucleotide sequences of plant species aligned by families to the main genomic regions used in species identification and taxonomic studies. In the section *Taxonomic groups* of PlantAligDB (http://plantaligdb.portugene.com/cgi-bin/PlantAligDB_taxonomicgroups.cgi) (Figure 6) we provided detailed plant families information.

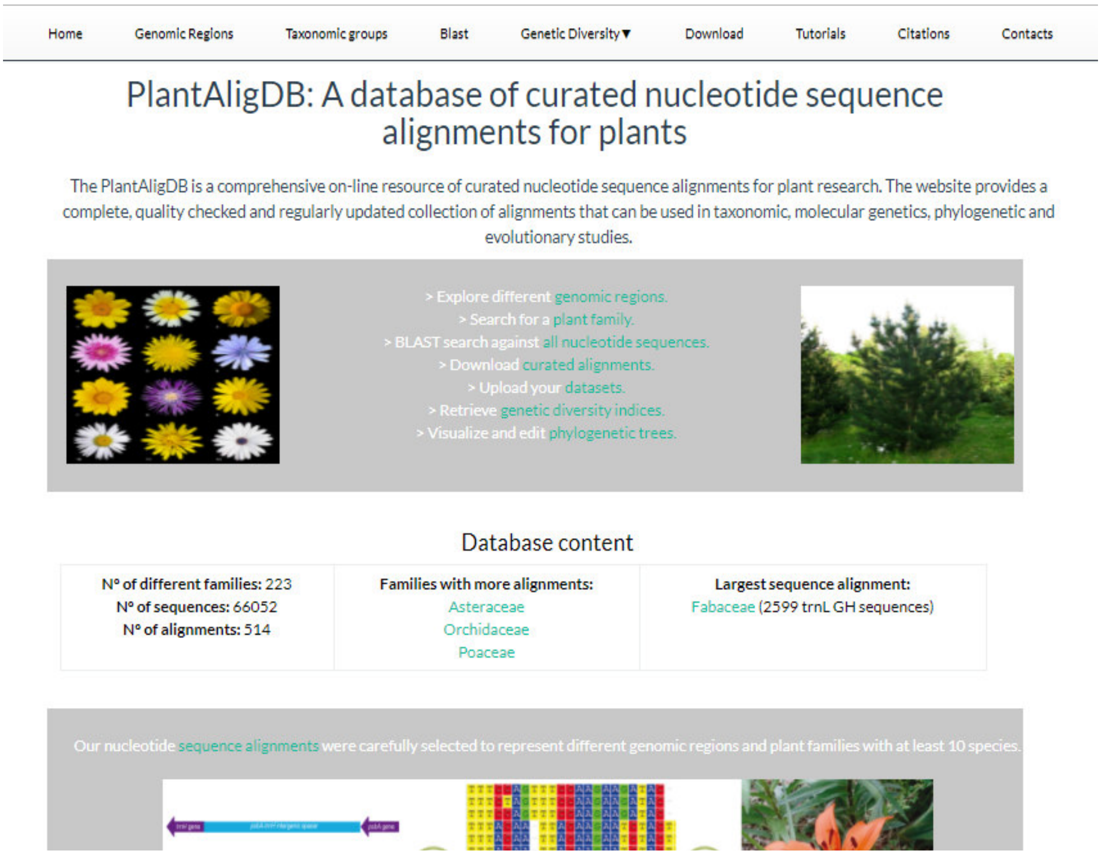
The use of a set of family-specific primers allows more accurate and precise quantification of the cpDNA present in the samples, even when contaminated with human DNA or other animals. The design of PCR primers in highly conserved regions significantly increases the probability of successful amplifications in highly divergent species. This approach is useful for detecting and amplifying length polymorphisms, discriminating between different plant species and are valid for identifying variations in DNA sequences (Pereira, Carneiro et al. 2010, Yang, Kung et al. 2015). Many highly conserved regions exist simultaneously in cpDNA, such as tRNA genes, which had similar conserved in structure, content and location. These regions therefore provide suitable targets for designing conserved PCR primers (Figure 2). Sufficiently conserved regions have been selected allowing the same primer to be used for several families, for example, cpDNAatpF_ABIRS_F primer have been used for families Asteraceae, Brassicaceae, Iridaceae, Rosaceae and Salicaceae or cpDNAtrnH_R primer that can be used for the seven families (Study 2: Table 1, Figure S1).

Small structural variations in cpDNA did not affect the usefulness of the universal primers because they corresponded to highly conserved regions of the genome and were designed from alignment of various sequences (Figure 4). The number of aligned sequences varies from 13 in *atpF-atpH* for Salicaceae, to 2078 in *trnL* GH for Poaceae. All families had high values of Pairwise Identity in the sites where the primers were designed, most of them with 100% in both forward and reverse primers. The greatest variation in the target region length was found in Orchidaceae sequences (546bp) for *trnL* CD genomic region (Study 2: Table S1). It has been reported that species of this family have wide length variation of the chloroplast genome due to the presence of indels (Jheng, Chen et al. 2012, Yang, Tang et al. 2013, Peyachoknagul, Mongkolsiriwatana et al. 2014).

In order to evaluate the utility of the universal primers, we have tested DNA extractions from different plant tissues. We obtain negative PCR amplifications when using DNA extracted from petals and dry products. There is evidence that organelle DNA can behave differently in different materials, both quantitatively and structurally (Golczyk, Greiner et al. 2014). We extracted DNA from 14 fresh or frozen leaves, two different species for each of the seven families, using the CTAB method. The extraction method proved effective even when frozen leaf samples were used. The family individuals could be differentiated by the difference in size of the fragments generated by the indels (Study 2: Figure 1).

We expect that these universal primers will effectively increase the efficiency and feasibility of complete cpDNA sequencing (Study 2: Table 1). These primers can improve the phylogenetic resolution and aid in identifying of plant species, especially at the taxonomic level below family. Determining the sequencing reliability of complete genomes is crucial for phylogenetic studies, and it is directly related to the reliability of the primers (Yang, Li et al. 2014). Our set of primers can be used to identify the presence of species from these important families in mixed samples. Future work can will include the development of larger sets of primers for other interest families of plants and using the conserved regions that we have identified in the first study or other important genomic regions such as *matK*, *rbcL* and ITS how we identified in third study (Study 3: Table 1).

In addition to the existence of universal PCR primers, a successful identification system requires also the existence of reference sequence databases. Therefore, we decided to use the sequences obtained in the first study, and search for new ones for others genomic regions, to build the first database with manually curated alignments of nucleotide sequences of plants families for several genomic regions. The PlantAligDB integrates the existing information of cpDNA genomic regions in different families with an intuitive interface and research tools, to facilitate the work of researchers (http://plantaligdb.portugene.com/cgi-bin/PlantAligDB_home.cgi) (Figure 6). It can be used by researchers to develop analysis and share results with collaborators or local storage, as there is availability of unloading the alignments. The goal is collect and maintain relevant information about plant families, for genomic regions used in species identification, and present it in an easily accessible format (Study 3). Table 1 summarise the current statistics available in the PlantAligDB.



PlantAligDB: A database of curated nucleotide sequence alignments for plants

The PlantAligDB is a comprehensive on-line resource of curated nucleotide sequence alignments for plant research. The website provides a complete, quality checked and regularly updated collection of alignments that can be used in taxonomic, molecular genetics, phylogenetic and evolutionary studies.

> Explore different genomic regions.
> Search for a plant family.
> BLAST search against all nucleotide sequences.
> Download curated alignments.
> Upload your datasets.
> Retrieve genetic diversity indices.
> Visualize and edit phylogenetic trees.

Database content

Nº of different families: 223 **Nº of sequences:** 66052 **Nº of alignments:** 514	**Families with more alignments:** Asteraceae Orchidaceae Poaceae	**Largest sequence alignment:** Fabaceae (2599 trnL GH sequences)

Our nucleotide sequence alignments were carefully selected to represent different genomic regions and plant families with at least 10 species.

Figure 6. Home page of PlantAligDB.

Our alignments were performed with 10 or more species per family (Figure 4), in order to avoid biases in the analyses. The alignments were checked manually to ensure proper data layout and eliminate inconsistencies that may lead to incorrect conclusions (Sakai, Lee et al. 2013). The PlantAligDB can help researchers to develop new methods of identifying plant species and be used as reference database for phylogenetic studies. The database also describes the most conserved and variable regions of the main genomic regions used in plant species identification (*atpF-atpH*, *psbA-trnH*, *trnL* CD, *trnL* GH, *matK*, *rbcL* and ITS) (Figure 7) for those researchers interested in designing new primers. Our database is embracing because it brings together the contents of different plant families, unlike, for example, GDR database that is dedicated a single family (Rosaceae) (Jung, Staton et al. 2007). Although there are other databases dedicated to plants that provide molecular and taxonomic information most of which are limited to interest species like rice (Sakai, Lee et al. 2013), wheat (Lai, Berkman et al. 2012) and potato (Meyer, Nagel et al. 2005). In PlantAligDB is available molecular and taxonomic information of thousands of species once it has been built with more than 66,000 sequences. Our dataset is restricted to the main genomic regions of the cpDNA and one region of the nuclear genome (Figure 7), but other regions may be added in the future.

The data in PlantAligDB provides an overview of the diversity of 223 different plant families (Study 3: Figure 2), and may serve as a reference database in the search for unknown sequences through the BLAST tool included. PlantAligDB has an efficient integration of the multiple data available thanks to research tools that facilitate the analysis of several families quickly.

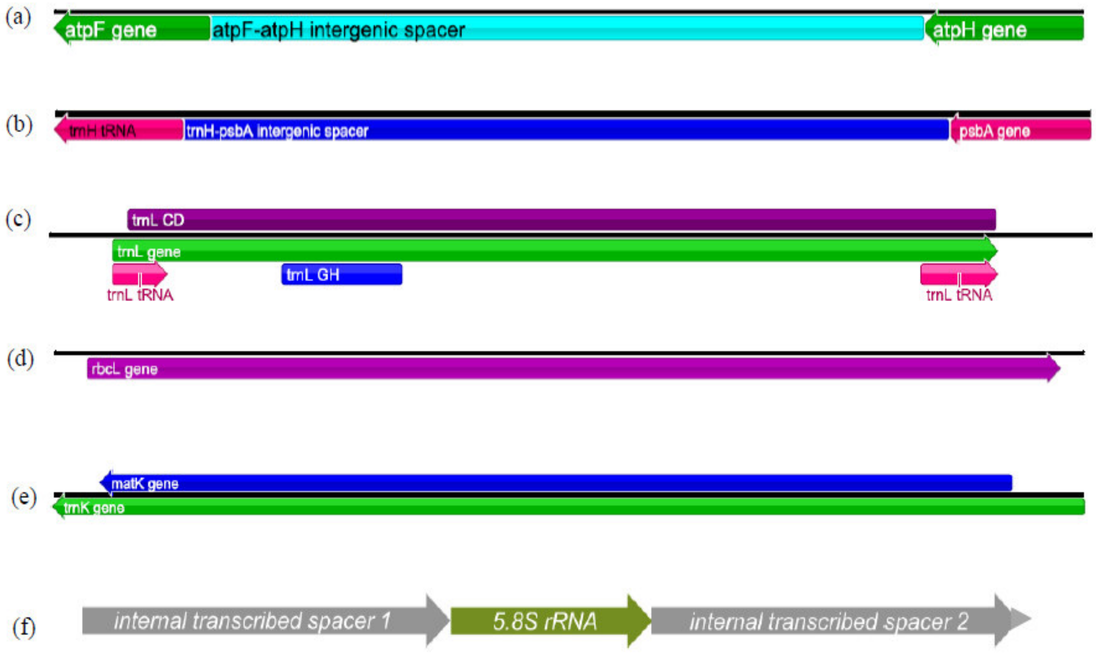


Figure 7. Schematic representation of genomic regions analysed in PlantAligDB a) *atpF-atpH* intergenic region, b) *psbA-trnH* intergenic region, c) *trnL* (UAA) intron (CD) and P6 loop (GH), d) *rbcL* gene, e) *matK* gene and f) ITS (internal transcribed spacer).

The PlantAligDB provides information on the genetic diversity of the regions through two conservation measures, the Percentage of identical sites (PIS) and Percentage of pairwise identity (PPI). These data can be found in the section *Genetic diversity*, where a dynamic table present the measures of sequence conservation for region in each family (Study 3: Figure 2, Tables 2 and 3). The ITS [Figure 7 f)] was the region with the lowest conservation values, with PIS mean of 28.84 and PPI mean of 87.21. This marker hold few identical sites in the sequences aligned (Table 5). *rbcL* [Figure 7 d)] was the region with highest mean PIS (78.48) and *trnL* GH [Figure 7 c)] showed the highest conserved values with mean PPI of (97.09) reaching 100% in PPI for two families (Calycanthaceae and Limnanthaceae) (Table 4). This results allow an overview of how conserved are the regions across the families (Study 3: Figure 2). Through the PIS measure we can conclude that in our set of data, regions from the most diverse to the most conserved was ITS > *psbA-trnH* > *atpF-atpH* > *trnL* GH > *trnL* CD > *matK* > *rbcL* (Figure 7).

Table 4. The families with highest and lowest PIS and PPI conservation measure for targets regions analysed in PlantAligDB.

Marker	Family	PIS	Mean	Family	PPI	Mean
atpF-atpH	Poaceae	15.19	55.05	Ranunculaceae	87.59	95.49
	Araucariaceae	85.95		Araucariaceae	99.37	
psbA-trnH	Fabaceae	23	39.01	Hymenophyllaceae	69.96	94.02
	Ephedraceae	97.42		Ephedroceae	99.94	
***trnL* CD**	Poaceae	21	61.13	Polyganaceae	90.54	96.34
	Magnoliaceae	96.46		Theaceae	99.6	
***trnL* GH**	Orchidaceae	6.43	58.62	Convolvulaceae	88.58	97.09
	Calycanthaceae	99.07		Calycanthaceae/ Limnanthaceae	100	
rbcL	Orobanchaceae	27.47	78.48	Orobanchaceae	81.18	96.94
	Riperaceae	95.24		Fagaceae	99.48	
matK	Fabaceae	0.16	65	Characeae	78.72	95.23
	Betulaceae	97.03		Berberidaceae	99.62	
ITS	Amaryllidaceae	1.13	28.84	Amaryllidaceae	70.79	87.21
	Lamiaceae	78.45		Lamiaceae	97.75	

The PlantAligDB is by far the largest set of alignments for plant families currently available and presents diverse information for each genomic region. Moreover, the database allows multiple types of interactions with the datasets so users can have a fast characterization of genomic regions (Figure 7) and plant families. For the first time, important plant families and genomic regions can be analysed using a single platform. We believe that PlantAligDB will be a useful tool to help researchers gain greater knowledge about important plant families and genomic regions. PlantAligDB will also be useful for researchers interested in designing accurate methods for the identification and screening of plant species in different contexts by providing detailed information on deleted or duplicated genomic regions. Finally, our database is an easily accessible platform for those who might want to explore the organization and the general conservation level assign of the main genomic regions. The disposition of this information within one place, together with links to external resources, greatly facilitates researchers who wish to use this information to improve the study of plant species (Lai, Berkman et al. 2012, Damas, Carneiro et al. 2013).

In this dissertation, we focused on cpDNA genetic markers (Figure 8) by several reasons. The cpDNA has highly informative noncoding regions rich in indels, which we used in the work described in manuscript 1, often including conserved domains, used to design universal PCR primers as described in manuscript 2 (Figure 1). The cpDNA has also several reference genomes sequences available, which we use to build the PlantAligDB (Figure 6) described in manuscript 3. The cpDNA has also a small size and high copy per cell which, increases the possibility of obtaining material from low quality and quantity DNA samples (Ronning, Rudi et al. 2005, Pereira, Carneiro et al. 2010, Lin, Lin et al. 2015, Thomsen and Willerslev 2015, De Castro, Comparone et al. 2017).

This dissertation describes three original research works that contribute with new molecular and bioinformatics tools to study the most relevant families of plants. The main objectives were achieved, since we were able to demonstrate that the identification of plant species can be achieved using variable length chloroplast DNA sequences. We have also designed and successfully tested conserved PCR primers for amplification of informative chloroplast DNA regions in the most relevant plants families. The PlantAligDB is available and provides a comprehensive free on-line resource of curated nucleotide sequence alignments for plant research. We have successfully carried out several important steps during this research, including efficient DNA extractions from leave tissues, designed a set of conserved PCR primers, performed PCR amplifications and carried our diverse bioinformatics analyses and built a web-based workbench.

CHAPTER V

REFERENCES

REFERENCES

Abe, T., et al. (2014). "tRNADB-CE: tRNA gene database well-timed in the era of big sequence data." Front Genet **5**: 114.

Aburjai, T. and F. M. Natsheh (2003). "Plants used in cosmetics." Phytotherapy research **17**(9): 987-1000.

Altschul, S. F., et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Alves, C., et al. (2017). "Species identification in forensic samples using the SPInDel approach: A GHEP-ISFG inter-laboratory collaborative exercise." Forensic Science International: Genetics **28**: 219-224.

Apweiler, R., et al. (2001). "The InterPro database, an integrated documentation resource for protein families, domains and functional sites." Nucleic Acids Res **29**(1): 37-40.

Arenas, M., et al. (2017). "Forensic genetics and genomics: Much more than just a human affair." PLoS Genetics **13**(9): e1006960.

Arumuganathan, K. and E. D. Earle (1991). "Nuclear DNA content of some important plant species." Plant Molecular Biology Reporter **9**(3): 208-218.

Attwood, T. K. (2002). "The PRINTS database: a resource for identification of protein families." Brief Bioinform **3**(3): 252-263.

Baraket, G., et al. (2008). "Chloroplast DNA analysis in Tunisian fig cultivars (Ficus carica L.): Sequence variations of the trnL-trnF intergenic spacer." Biochemical Systematics and Ecology **36**(11): 828-835.

Barolo, M. I., et al. (2014). "Ficus carica L.(Moraceae): An ancient source of food and health." Food Chem **164**: 119-127.

Bell, K. L., et al. (2016). "Review and future prospects for DNA barcoding methods in forensic palynology." Forensic Science International: Genetics **21**: 110-116.

Bennetzen, J. L. (2000). "Comparative sequence analysis of plant nuclear genomes:m microcolinearity and its many exceptions." Plant Cell **12**(7): 1021-1029.

Bhargava, M. and A. Sharma (2013). "DNA barcoding in plants: evolution and applications of in silico approaches and resources." Mol Phylogenet Evol **67**(3): 631-641.

Bommarco, R., et al. (2013). "Ecological intensification: harnessing ecosystem services for food security." Trends in Ecology & Evolution **28**(4): 230-238.

Bruneau, A., et al. (2001). "Phylogenetic relationships in the Caesalpinioideae (Leguminosae) as inferred from chloroplast trnL intron sequences." Systematic Botany **26**(3): 487-514.

Bustin, S. A. (2005). "Real-time PCR." Encyclopedia of diagnostic genomics and proteomics **10**(1): 117-111.

Cafasso, D., et al. (2004). "Chloroplast DNA inheritance in the orchid Anacamptis palustris using single-seed polymerase chain reaction." Journal of Heredity **96**(1): 66-70.

Carneiro, J., et al. (2012). "SPInDel: a multifunctional workbench for species identification using insertion/deletion variants." Mol Ecol Resour **12**(6): 1190-1195.

Carneiro, J., et al. (2017). "The HIV oligonucleotide database (HIVoligoDB)." Database **2017**(1).

Cassidy, E. S., et al. (2013). "Redefining agricultural yields: from tonnes to people nourished per hectare." Environmental Research Letters **8**(3): 034015.

Chase, M. W., et al. (2003). "DNA data and Orchidaceae systematics: a new phylogenetic classification." Orchid conservation **69**: 89.

Chase, M. W., et al. (2007). "A proposal for a standardised protocol to barcode all land plants." Taxon **56**(2): 295-299.

Chase, M. W. and M. F. Fay (2009). "Barcoding of plants and fungi." Science **325**(5941): 682-683.

Chen, S., et al. (2010). "Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species." PLoS One **5**(1): e8613.

Cheng, Y.-J., et al. (2003). "An efficient protocol for genomic DNA extraction from Citrus species." Plant Molecular Biology Reporter **21**(2): 177a.

Coyle, H. M. (2004). Forensic botany: principles and applications to criminal casework, CRC Press.

Coyle, H. M., et al. (2005). "Forensic botany: using plant evidence to aid in forensic death investigation." Croatian medical journal **46**(4): 606.

Curci, P. L., et al. (2015). "Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other Asteraceae." PLoS One **10**(3): e0120589.

da Cruz Cabral, L., et al. (2013). "Application of plant derived compounds to control fungal spoilage and mycotoxin production in foods." International Journal of Food Microbiology **166**(1): 1-14.

Damas, J., et al. (2013). "MitoBreak: the mitochondrial DNA breakpoints database." Nucleic Acids Res **42**(D1): D1261-D1268.

Daniell, H., et al. (2016). "Chloroplast genomes: diversity, evolution, and applications in genetic engineering." Genome Biol **17**(1): 134.

De Castro, O., et al. (2017). "What is in your cup of tea? DNA Verity Test to characterize black and green commercial teas." PLoS One **12**(5): e0178262.

Dellaporta, S. L., et al. (1983). "A plant DNA minipreparation: version II." Plant Molecular Biology Reporter **1**(4): 19-21.

Derocles, S. A., et al. (2015). "Determining plant–leaf miner–parasitoid interactions: a DNA barcoding approach." PLoS One **10**(2): e0117872.

Devantier, Y. A., et al. (1993). "Microprojectile-mediated DNA delivery to the Salicaceae family." Canadian journal of botany **71**(11): 1458-1466.

Díaz, S., et al. (2006). "Biodiversity loss threatens human well-being." PLoS biology **4**(8): e277.

Domenech, B. A.-L., C. B.; Baker, W. J.; and E. P. Alapetite, J.; and Nadot, S. (2014). "A phylogenetic analysis of palm subtribe Archontophoenicinae (Aracaceae) based on 14 DNA regions." Botanical Journal of the Linnean Society **175**: 469-481.

Dong, W., et al. (2014). "Discriminating plants using the DNA barcode rbcLb: an appraisal based on a large data set." Mol Ecol Resour **14**(2): 336-343.

Dong, W., et al. (2012). "Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding." PLoS One **7**(4): e35071.

Dong, W., et al. (2015). "ycf1, the most promising plastid DNA barcode of land plants." Sci Rep **5**: 8348.

Doyle, J. J. (1987). "A rapid DNA isolation procedure for small quantities of fresh leaf tissue." Phytochem. Bull. **19**: 11-15.

Doyle, J. J., et al. (1992). "Chloroplast DNA inversions and the origin of the grass family (Poaceae)." Proceedings of the National Academy of Sciences **89**(16): 7722-7726.

Drummond AJ, A. B., Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A (2009). "Geneious Pro 4.8.2.".

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Egan, A. N., et al. (2012). "Applications of next-generation sequencing in plant biology." Am J Bot **99**(2): 175-185.

Fazekas, A. J., et al. (2008). "Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well." PLoS One **3**(7): e2802.

Ferri, G., et al. (2015). "Forensic botany II, DNA barcode for land plants: Which markers after the international agreement?" Forensic Science International: Genetics **15**: 131-136.

Feuillet, C., et al. (2011). "Crop genome sequencing: lessons and rationales." Trends Plant Sci **16**(2): 77-88.

Ford, C. S., et al. (2009). "Selection of candidate coding DNA barcoding regions for use on land plants." Botanical Journal of the Linnean Society **159**(1): 1-11.

Franzke, A., et al. (2011). "Cabbage family affairs: the evolutionary history of Brassicaceae." Trends Plant Sci **16**(2): 108-116.

Frézal, L. and R. Leblois (2008). "Four years of DNA barcoding: current advances and prospects." Infection, Genetics and Evolution **8**(5): 727-736.

Gao, T., et al. (2010). "Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2." J Ethnopharmacol **130**(1): 116-121.

Gawel, N. and R. Jarret (1991). "A modified CTAB DNA extraction procedure for Musa and Ipomoea." Plant Molecular Biology Reporter **9**(3): 262-266.

Geller, J., et al. (2013). "Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys." Mol Ecol Resour **13**(5): 851-861.

Ghada, B., et al. (2010). "Molecular evolution of chloroplast DNA in fig (Ficus carica L.): Footprints of sweep selection and recent expansion." Biochemical Systematics and Ecology **38**(4): 563-575.

Ghahramanzadeh, R., et al. (2013). "Efficient distinction of invasive aquatic plant species from non-invasive related species using DNA barcoding." Mol Ecol Resour **13**(1): 21-31.

Golczyk, H., et al. (2014). "Chloroplast DNA in mature and senescing leaves: a reappraisal." Plant Cell **26**(3): 847-854.

Gonçalves, J., et al. (2015). "A multiplex PCR assay for identification of the red fox (Vulpes vulpes) using the mitochondrial ribosomal RNA genes." Conservation Genetics Resources **7**(1): 45-48.

Graham, S. W., et al. (2000). "Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference." International Journal of Plant Sciences **161**(S6): S83-S96.

Green, B. R. (2011). "Chloroplast genomes of photosynthetic eukaryotes." The plant journal **66**(1): 34-44.

Group, C. P. W., et al. (2009). "A DNA barcode for land plants." Proceedings of the National Academy of Sciences **106**(31): 12794-12797.

Gualberto, J. M. and K. J. Newton (2017). "Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation." Annu Rev Plant Biol **68**: 225-252.

Hajibabaei, M., et al. (2007). "DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics." TRENDS in Genetics **23**(4): 167-172.

Hamilton, M. B., et al. (2003). "Patterns and relative rates of nucleotide and insertion/deletion evolution at six chloroplast intergenic regions in New World species of the Lecythidaceae." Mol Biol Evol **20**(10): 1710-1721.

Hebert, P. D., et al. (2003). "Biological identifications through DNA barcodes." Proceedings of the Royal Society of London B: Biological Sciences **270**(1512): 313-321.

Hollingsworth, M. L., et al. (2009). "Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants." Mol Ecol Resour **9**(2): 439-457.

Hollingsworth, P. M., et al. (2011). "Choosing and using a plant DNA barcode." PLoS One **6**(5): e19254.

Hussain, S., et al. (2014). "Evaluation of in-vitro anti-mycobacterial activity and isolation of active constituents from Crocus sativus L.(Iridaceae)." Asian Journal of Medical and Pharmaceutical Researches **4**(2): 130-135.

Hwang, S.-G., et al. (2015). "Chloroplast markers for detecting rice grain-derived food ingredients in commercial mixed-flour products." Genes & Genomics **37**(12): 1027-1034.

Jansen, R. K., et al. (2007). "Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns." Proceedings of the National Academy of Sciences **104**(49): 19369-19374.

Jheng, C.-F., et al. (2012). "The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish Phalaenopsis orchids." Plant science **190**: 62-73.

Jin, J., et al. (2013). "PLncDB: plant long non-coding RNA database." Bioinformatics **29**(8): 1068-1071.

Jin, W.-T., et al. (2014). "Molecular systematics of subtribe Orchidinae and Asian taxa of Habenariinae (Orchideae, Orchidaceae) based on plastid matK, rbcL and nuclear ITS." Mol Phylogenet Evol **77**: 41-53.

Joly, A., et al. (2016). LifeCLEF 2016: multimedia life species identification challenges. International Conference of the Cross-Language Evaluation Forum for European Languages, Springer.

Jung, S., et al. (2007). "GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data." Nucleic Acids Res **36**(suppl_1): D1034-D1040.

Kajita, T., et al. (1998). "Molecular phylogeny of Dipetrocarpaceae in southeast Asia based on nucleotide sequences ofmatK, trnL Intron, andtrnL-trnF intergenic spacer region in chloroplast DNA." Mol Phylogenet Evol **10**(2): 202-209.

Kane, N., et al. (2012). "Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA." Am J Bot **99**(2): 320-329.

Karehed, J., et al. (2008). "The phylogenetic utility of chloroplast and nuclear DNA markers and the phylogeny of the Rubiaceae tribe Spermacoceae." Mol Phylogenet Evol **49**(3): 843-866.

Karp, A. and I. Shield (2008). "Bioenergy from plants and the sustainable yield challenge." New Phytologist **179**(1): 15-32.

Karrenberg, S., et al. (2002). "The life history of Salicaceae living in the active zone of floodplains." Freshwater Biology **47**(4): 733-748.

Kearse, M., et al. (2012). "Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data." Bioinformatics **28**(12): 1647-1649.

Kelchner, S. A. (2000). "The evolution of non-coding chloroplast DNA and its application in plant systematics." Annals of the Missouri Botanical Garden: 482-498.

Kellogg, E. A. and J. L. Bennetzen (2004). "The evolution of nuclear genome structure in seed plants." Am J Bot **91**(10): 1709-1725.

Khan, M. Q. and Z. K. Shinwari (2016). "The ethnomedicinal profile of family rosaceae; a study on pakistani plants." Pak. J. Bot **48**(2): 613-620.

Khanuja, S. P., et al. (1999). "Rapid isolation of DNA from dry and fresh samples of plants producing large amounts of secondary metabolites and essential oils." Plant Molecular Biology Reporter **17**(1): 74-74.

Kikkawa, H. S., et al. (2016). "Real-Time PCR Quantification of Chloroplast DNA Supports DNA Barcoding of Plant Species." Mol Biotechnol **58**(3): 212-219.

Koch, M., et al. (2001). "Molecular systematics of the Brassicaceae: evidence from coding plastidic matK and nuclear Chs sequences." Am J Bot **88**(3): 534-544.

Kress, W. J. and D. L. Erickson (2007). "A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region." PLoS One **2**(6): e508.

Kress, W. J., et al. (2005). "Use of DNA barcodes to identify flowering plants." Proc Natl Acad Sci U S A **102**(23): 8369-8374.

Kumar, A., et al. (2016). "Higher efficiency of ISSR markers over plastid psbA-trnH region in resolving taxonomical status of genus Ocimum L." Ecology and Evolution **6**(21): 7671-7682.

Kurata, N. and Y. Yamazaki (2006). "Oryzabase. An integrated biological and genome information database for rice." Plant Physiol **140**(1): 12-17.

Lahaye, R. S., V.; Duthoit, S.; Maurin, O. and van der Bank, M. (2008). "A test of psbK-psbI and atpF-atpH as potential plant DNA barcodes using the flora of the Kruger National Park as a model system (South Africa)." Nature Precedings: 21.

Lai, K., et al. (2012). "WheatGenome.info: an integrated database and portal for wheat genome information." Plant Cell Physiol **53**(2): e2.

Lee, S. Y., et al. (2016). "Rapid species identification of highly degraded agarwood products from Aquilaria using real-time PCR." Conservation Genetics Resources **8**(4): 581-585.

Li, L., et al. (2015). "Thuniopsis: A New Orchid Genus and Phylogeny of the Tribe Arethuseae (Orchidaceae)." PLoS One **10**(8): e0132777.

Li, X., et al. (2015). "Plant DNA barcoding: from gene to genome." Biol Rev Camb Philos Soc **90**(1): 157-166.

Lin, J.-Y., et al. (2015). "Evaluation of chloroplast DNA markers for distinguishing Phalaenopsis species." Scientia Horticulturae **192**: 302-310.

Linacre, A. and S. S. Tobe (2011). "An overview to the investigative approach to species testing in wildlife forensic science." Investigative genetics **2**(1): 2.

Liu, Z., et al. (2012). "Identification of medicinal vines by ITS2 using complementary discrimination methods." J Ethnopharmacol **141**(1): 242-249.

Lohse, M., et al. (2014). "Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data." Plant Cell Environ **37**(5): 1250-1258.

Loreau, M., et al. (2001). "Biodiversity and ecosystem functioning: current knowledge and future challenges." Science **294**(5543): 804-808.

Löytynoja, A. (2014). "Phylogeny-aware alignment with PRANK." Multiple sequence alignment methods: 155-170.

Loytynoja, A., et al. (2012). "Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm." Bioinformatics **28**(13): 1684-1691.

Lynch, M., et al. (2006). "Mutation pressure and the evolution of organelle genomic architecture." Science **311**(5768): 1727-1730.

Ma, X. Y., et al. (2010). "Species identification of medicinal pteridophytes by a DNA barcode marker, the chloroplast psbA-trnH intergenic region." Biol Pharm Bull **33**(11): 1919-1924.

Mahadani, P. and S. K. Ghosh (2014). "Utility of indels for species-level identification of a biologically complex plant group: a study with intergenic spacer in Citrus." Mol Biol Rep **41**(11): 7217-7222.

Marsh, A. J., et al. (2014). "Sequence-based analysis of the bacterial and fungal compositions of multiple kombucha (tea fungus) samples." Food microbiology **38**: 171-178.

Meyer, S., et al. (2005). "PoMaMo--a comprehensive database for potato genome data." Nucleic Acids Res **33**(Database issue): D666-670.

Mishra, P., et al. (2016). "DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market." Plant Biotechnol J **14**(1): 8-21.

Moon, J. C., et al. (2016). "Development of multiplex PCR for species-specific identification of the Poaceae family based on chloroplast gene, rpoC2." Applied Biological Chemistry **59**(2): 201-207.

Moreira, F., et al. (2017). "A proposal for standardization of transgenic reference sequences used in food forensics." Forensic Science International: Genetics **29**: e26-e28.

Moritz, C. and C. Cicero (2004). "DNA barcoding: promise and pitfalls." PLoS biology **2**(10): e354.

Murray, M. G. and W. F. Thompson (1980). "Rapid isolation of high molecular weight plant DNA." Nucleic Acids Res **8**(19): 4321-4326.

Nam, M., et al. (2015). "Development of multiplex RT-PCR for simultaneous detection of garlic viruses and the incidence of garlic viral disease in garlic genetic resources." Plant Pathol J **31**(1): 90.

Narayan, L., et al. (2015). "A genotyping protocol for multiple tissue types from the polyploid tree species Sequoia sempervirens (Cupressaceae)." Appl Plant Sci **3**(3): 1400110.

Neubig, K. M., et al. (2009). "Phylogenetic utility of ycf1 in orchids: a plastid gene more variable than matK." Plant Systematics and Evolution **277**(1-2): 75-84.

Nock, C. J., et al. (2011). "Chloroplast genome sequences from total DNA for plant identification." Plant Biotechnol J **9**(3): 328-333.

Notsu, Y., et al. (2002). "The complete sequence of the rice (Oryza sativa L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants." Molecular Genetics and Genomics **268**(4): 434-445.

Numa, H. and T. Itoh (2014). "MEGANTE: a web-based system for integrated plant genome annotation." Plant Cell Physiol **55**(1): e2.

Ogden, R. and A. Linacre (2015). "Wildlife forensic science: a review of genetic geographic origin assignment." Forensic Science International: Genetics **18**: 152-159.

Olmstead, R. G. and J. D. Palmer (1994). "Chloroplast DNA systematics: a review of methods and data analysis." Am J Bot: 1205-1224.

Palmer, J. D., et al. (1988). "Chloroplast DNA variation and plant phylogeny." Annals of the Missouri Botanical Garden: 1180-1206.

Panero, J. L. and B. S. Crozier (2003). "Primers for PCR Amplification of Asteraceae Chloroplast DNA." Lundellia.

Pang, X., et al. (2012). "Assessing the potential of candidate DNA barcodes for identifying non-flowering seed plants." Plant Biol (Stuttg) **14**(5): 839-844.

Parker, J. and A. J. Helmstetter (2017). "Field-based species identification of closely-related plants using real-time nanopore sequencing." **7**(1): 8345.

Parson, W., et al. (2000). "Species identification by means of the cytochrome b gene." International journal of legal medicine **114**(1): 23-28.

Pennisi, E. (2007). "Wanted: a barcode for plants." Science **318**(5848): 190-191.

Pereira, F., et al. (2008). "Identification of species with DNA-based technology: current progress and challenges." Recent patents on DNA & gene sequences **2**(3): 187-200.

Pereira, F., et al. (2010). "Identification of species by multiplex analysis of variable-length sequences." Nucleic Acids Res **38**(22): e203-e203.

Pereira, F., et al. (2010). "A guide for mitochondrial DNA analysis in non-human forensic investigations." Open Forensic Science Journal **3**: 33-44.

Pérez-Escobar, O. A., et al. (2015). "Rumbling orchids: how to assess divergent evolution between chloroplast endosymbionts and the nuclear host." Systematic biology **65**(1): 51-65.

Petit, R. J., et al. (2005). "Invited review: comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations." Molecular Ecology **14**(3): 689-701.

Peyachoknagul, S., et al. (2014). "Identification of native Dendrobium species in Thailand by PCR-RFLP of rDNA-ITS and chloroplast DNA." ScienceAsia **40**(2): 113-120.

Qian, J., et al. (2013). "The complete chloroplast genome sequence of the medicinal plant Salvia miltiorrhiza." PLoS One **8**(2): e57607.

Quandt, D. and M. Stech (2005). "Molecular evolution of the trnL UAA intron in bryophytes." Mol Phylogenet Evol **36**(3): 429-443.

Ran, J. H., et al. (2010). "A test of seven candidate barcode regions from the plastome in Picea (Pinaceae)." Journal of integrative plant biology **52**(12): 1109-1126.

Ratnasingham, S. and P. D. Hebert (2007). "bold: The Barcode of Life Data System (http://www.barcodinglife.org)." Mol Ecol Notes **7**(3): 355-364.

Rogers, S. O. and A. J. Bendich (1985). "Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues." Plant Mol Biol **5**(2): 69-76.

Ronning, S. B., et al. (2005). "Differentiation of important and closely related cereal plant species (Poaceae) in food by hybridization to an oligonucleotide array." J Agric Food Chem **53**(23): 8874-8880.

Rudall, P. (1994). "Anatomy and systematics of Iridaceae." Botanical Journal of the Linnean Society **114**(1): 1-21.

Saddhe, A. A., et al. (2017). "Evaluation of multilocus marker efficacy for delineating mangrove species of West Coast India." PLoS One **12**(8): e0183245.

Sakai, H., et al. (2013). "Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics." Plant Cell Physiol **54**(2): e6.

Santos, C. and F. Pereira (2017). "Design and evaluation of PCR primers for amplification of four chloroplast DNA regions in plants." Conservation Genetics Resources **9**(1): 9-12.

Sarwat, M. and M. M. Yamdagni (2016). "DNA barcoding, microarrays and next generation sequencing: recent tools for genetic diversity estimation and authentication of medicinal plants." Critical reviews in biotechnology **36**(2): 191-203.

Sass, C., et al. (2007). "DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads." PLoS One **2**(11): e1154.

Scriver, M., et al. (2015). "Development of species-specific environmental DNA (eDNA) markers for invasive aquatic plants." Aquatic Botany **122**: 27-31.

Seberg, O. and G. Petersen (2009). "How many loci does it take to DNA barcode a crocus?" PLoS One **4**(2): e4598.

Semerikova, S. A. and V. L. Semerikov (2014). "[Molecular phylogenetic analysis of the genus Abies (Pinaceae) based on the nucleotide sequence of chloroplast DNA]." Genetika **50**(1): 12-25.

Shaw, J., et al. (2005). "The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis." Am J Bot **92**(1): 142-166.

Shen, D., et al. (2012). "RadishBase: a database for genomics and genetics of radish." Plant and Cell Physiology **54**(2): e3-e3.

Sigurgeirsson, A. and A. E. Szmidt (1993). "Phylogenetic and biogeographic implications of chloroplast DNA variation in Picea." Nordic Journal of Botany **13**(3): 233-246.

Spaniolas, S., et al. (2010). "The potential of plastid trnL (UAA) intron polymorphisms for the identification of the botanical origin of plant oils." Food Chem **122**(3): 850-856.

Staats, M., et al. (2016). "Advances in DNA metabarcoding for food and wildlife forensic species identification." Anal Bioanal Chem **408**(17): 4615-4630.

Su, C.-l., et al. (2013). "Orchidstra: an integrated orchid functional genomics database." Plant and Cell Physiology **54**(2): e11-e11.

Taberlet, P., et al. (2007). "Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding." Nucleic Acids Res **35**(3): e14.

Taberlet, P., et al. (1991). "Universal primers for amplification of three non-coding regions of chloroplast DNA." Plant Mol Biol **17**(5): 1105-1109.

Tang, Y., et al. (2015). "Phylogeny and classification of the East Asian Amitostigma alliance (Orchidaceae: Orchideae) based on six DNA markers." BMC Evol Biol **15**(1): 96.

Techen, N., et al. (2014). "DNA barcoding of medicinal plant material for identification." Current Opinion in Biotechnology **25**: 103-110.

Teletchea, F., et al. (2005). "Food and forensic molecular identification: update and challenges." Trends in biotechnology **23**(7): 359-366.

Thomsen, P. F. and E. Willerslev (2015). "Environmental DNA–an emerging tool in conservation for monitoring past and present biodiversity." Biological Conservation **183**: 4-18.

Timme, R. E., et al. (2007). "A comparative analysis of the Lactuca and Helianthus (Asteraceae) plastid genomes: identification of divergent regions and categorization of shared repeats." Am J Bot **94**(3): 302-312.

Tsai, C.-C., et al. (2012). "Plastid trnL intron polymorphisms among Phalaenopsis species used for identifying the plastid genome type of Phalaenopsis hybrids." Scientia Horticulturae **142**: 84-91.

Valentini, A., et al. (2016). "Next‐generation monitoring of aquatic biodiversity using environmental DNA metabarcoding." Molecular Ecology **25**(4): 929-942.

Vassou, S. L., et al. (2015). "DNA barcoding for species identification from dried and powdered plant parts: A case study with authentication of the raw drug market samples of Sida cordifolia." Gene **559**(1): 86-93.

Veidenberg, A., et al. (2016). "Wasabi: An Integrated Platform for Evolutionary Sequence Analysis and Data Visualization." Mol Biol Evol **33**(4): 1126-1130.

Wallinger, C., et al. (2012). "Rapid plant identification using species-and group-specific primers targeting chloroplast DNA." PLoS One **7**(1): e29473.

Wang, M., et al. (2015). "Comparative analysis of Asteraceae chloroplast genomes: Structural organization, RNA editing and evolution." Plant Molecular Biology Reporter **33**(5): 1526-1538.

Wang, W., et al. (2010). "DNA barcoding of the Lemnaceae, a family of aquatic monocots." BMC Plant Biol **10**(1): 205.

Wang, X., et al. (2014). "Identification of clinically relevant fungi and prototheca species by rRNA gene sequencing and multilocus PCR coupled with electrospray ionization mass spectrometry." PLoS One **9**(5): e98110.

Wolfe, K. H., et al. (1987). "Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs." Proceedings of the National Academy of Sciences **84**(24): 9054-9058.

Woolfe, M. and S. Primrose (2004). "Food forensics: using DNA technology to combat misdescription and fraud." Trends in biotechnology **22**(5): 222-226.

Xiong, A.-S., et al. (2009). "Gene duplication, transfer, and evolution in the chloroplast genome." Biotechnology advances **27**(4): 340-347.

Xu, J.-H., et al. (2015). "Dynamics of chloroplast genomes in green plants." Genomics **106**(4): 221-231.

Xu, Q., et al. (2013). "The draft genome of sweet orange (Citrus sinensis)." Nat Genet **45**(1): 59-66.

Yamane, K., et al. (2006). "Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice." DNA research **13**(5): 197-204.

Yang, F.-S. and X.-Q. Wang (2007). "Extensive length variation in the cpDNA trn T-trn F region of hemiparasitic Pedicularis and its phylogenetic implications." Plant Systematics and Evolution **264**(3): 251-264.

Yang, J. B., et al. (2014). "Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs." Mol Ecol Resour **14**(5): 1024-1031.

Yang, J. B., et al. (2013). "Complete chloroplast genome of the genus Cymbidium: lights into the species identification, phylogenetic implications and population genetic analyses." BMC Evol Biol **13**: 84.

Yang, Y. C., et al. (2015). "Development of primer pairs from diverse chloroplast genomes for use in plant phylogenetic research." Genet Mol Res **14**(4): 14857-14870.

Yao, H., et al. (2009). "Identification of Dendrobium species by a candidate DNA barcode sequence: the chloroplast psbA-trnH intergenic region." Planta Med **75**(06): 667-669.

Yao, H., et al. (2010). "Use of ITS2 region as the universal DNA barcode for plants and animals." PLoS One **5**(10).

Yu, J., et al. (2017). "PMDBase: a database for studying microsatellite DNA and marker development in plants." Nucleic Acids Res **45**(D1): D1046-D1053.

Yu, J., et al. (2011). "New universal matK primers for DNA barcoding angiosperms." Journal of Systematics and Evolution **49**(3): 176-181.

Yuan, J. S., et al. (2008). "Plants to power: bioenergy to fuel the future." Trends Plant Sci **13**(8): 421-429.

Zaiko, A., et al. (2015). "Metabarcoding approach for the ballast water surveillance--an advantageous solution or an awkward challenge?" Mar Pollut Bull **92**(1-2): 25-34.

Zaya, D. N. and M. V. Ashley (2012). "Plant genetics for forensic applications." Plant DNA Fingerprinting and Barcoding: Methods and Protocols: 35-52.

Zeng, S. Y., et al. (2017). "The Complete Chloroplast Genome Sequences of Six Rehmannia Species." Genes **8**(3).

Zhang, S., et al. (2013). "Apple gene function and gene family database: an integrated bioinformatics database for apple research." Plant growth regulation **70**(2): 199-206.