Identification of three species of genus *Allium* using DNA barcoding

Sepideh Anvarkhah¹, Mohammad Khajeh-Hosseini², Mohammad - Hassan Rashed - Mohassel³, Ali Davari- Edalat - Panah⁴ and Hoda Hashemi⁵

¹, ², ³ Department of Crop Science, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran.
⁴ Research and Development Center of Astan-e- Qouds, Mashhad, Iran
⁵ Department of Crop Science, Faculty of Agriculture, Islamic Azad University of Shirvan, Shirvan, Iran.

*Corresponding author email:* sa_sah_sa25@yahoo.com

ABSTRACT: DNA barcoding is a relatively new method developed to provide rapid, accurate and automatable species identification using standardized DNA sequences as tags. Species identification in this way is usually achieved by the repossession of a short DNA sequence, the “barcode” from a standard part of the genome of a given species. An experiment was conducted to identify three species of genus *Allium* using five barcode regions including *ndhJ*, *rpoC1*, *rpoB*, *YCF5* and *rbcL*. Results indicated that *A. ampeloprasum persicum* was the only identified species. Barcode region of *YCF5* had the highest SRSP with 93.33%. The highest interspecific diversity and intraspecific divergence with 0.8% and 0.4% were observed in region of *ndhJ* which was the most sufficient barcode region for species identification.

Keywords: *Allium cepa*, *Allium ampeloprasum persicum*, *Allium schoenoprasum* L., DNA barcoding, intra specific divergence, SNP

INTRODUCTION

DNA barcoding is a relatively new method which has been developed to provide rapid, accurate and automatable species identification using standardized DNA sequences as tags (Hebert et al., 2003; Taberlet et al., 2007). First studies started with the seminal work of Hebert et al. (2003) who reported that individual species of a collection of 200 similar species of lepidopterans could be identified with 100% accuracy using the mitochondrial gene cytochrome c oxidase subunit I (*COI*). As *COI* was not useful in plants, many loci have been proposed as plant barcodes, including *ITS* (Chase et al., 2009; Kress et al., 2005), *rbcL* (Kress and Erickson 2007; Newmaster et al., 2006), *psbA-trnH* (Chase et al., 2009; Kress and Erickson, 2007; Lahaye et al., 2008), and *matK* (Chase et al., 2009). Barcoding is now a reliable technique for species identification (Vijayan and Tsou, 2010). Because of advances in sequencing and computational technologies, DNA sequences have become the main source of new information for evolutionary and genetic relationships studies (Hajibabaei et al., 2007). DNA barcoding is based on the primers that are used for amplification of the short standardized sequences chosen for distinguishing the species (Hebert et al., 2003). Species identification through barcoding is usually achieved by the repossession of a short DNA sequence, the “barcode” from a standard part of the genome of the studied specimen. Barcoding has various applications and has been used for ecological studies (Dick and Kress, 2009), such as cryptic taxon identification (Lahaye et al., 2008) and verification of medicinal plant samples (Xue and Li, 2011). Several chloroplast gene regions are usually used as plant barcodes, with matK (matK) and ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit (*rbcL*) considered core barcodes (Hollingsworth et al., 2009). The spacer between tRNA-His and photosystem II protein D1 (*trnH-psbA* spacer) and the nuclear internal transcribed spacer 2 (*ITS2*) are also widely used (Chen et al., 2010; Fu et al., 2011; Gao et al., 2010). Barcoding works by matching sequence data from a studied sample (an unknown specimen) to a reference sequence (from a voucher specimen). The barcode sequence from each unknown specimen compared with a library of reference barcode sequences based on individuals of known character. A specimen is identified if its sequence closely
matches one in the barcode library. Otherwise, the new record can lead to a novel barcode sequence for a given species, or it can guide to identification of a new species (Hajibabaei et al., 2007).

Lahaye et al. (2008) reported that PCRs were generally successful with all studied barcodes, except ndhJ and YCF5, which did not amplify efficiently in orchids. Results showed that rbcL was not variable enough in the achlorophyllous Hydnora johannis (THUNB.) but worked in other parasitic plants. A matK exon segmentation was also amplified easily by using primers of 390F and 1326R in orchids. Studies of Schori and Showalter (2011) showed MatK, rbcL, and psbA-trnH sequences for Linum usitatissimum (L.) Griesb. 99-100% success but for L. usitatissimum vouchers and L. bienne (Mill.), none of these barcodes worked efficiently. Other species had no close matches for barcoding regions, which could happen due to poor sampling. Although Justicia adhatoda L. (Adulsa, Malabar Nut) had no similar matches for matK, other species of Justicia have no matK sequences in GenBank. Based on results of other studies, matK and rbcL were not always useful as barcodes for certain groups of plants (Fu et al., 2011; Roy et al., 2010).

As some species of Allium genus are very close to each other morphologically, their seeds cannot be distinguished even by machine vision (Anvarkhah, 2013). In present research we used DNA barcoding technique to identify some species of this genus.

MATERIALS AND METHODS

To screen the appropriate levels of sequence divergence in the plastid genome, 15 samples of three closely related species of Allium genus were selected for comparison (Table 1). The samples were collected from natural landscape or cultivated species in the Medicinal Plant Garden of Ferdowsi University of Mashhad, Iran, in 2011.

Table1. List of species used in the study

<table>
<thead>
<tr>
<th>No.</th>
<th>Scientific name</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Allium ampeloprasum persicum</td>
<td>Alliaceae</td>
</tr>
<tr>
<td>2</td>
<td>Allium schoenoprasum L.</td>
<td>Alliaceae</td>
</tr>
<tr>
<td>3</td>
<td>Allium cepa</td>
<td>Alliaceae</td>
</tr>
</tbody>
</table>

To obtain high quality DNA from fresh leave tissues DNA extraction kit (Fermentase) was used following the manufacturer’s instructions.

Universal primers for seven DNA loci (Table 2) were used according to Kress and Erickson (2007). Each PCR contained 2.5 µl of 10X PCR buffer, 2.5 µl MgCl2, 2 pM of each dNTP, 0.25 unit of Taq polymerase (CinnaGen Co., Iran), 4 pM of each primer and 20 ng of genomic DNA template.

Table2. PCR primers used for amplification of plastid DNA sequences

<table>
<thead>
<tr>
<th>region</th>
<th>Primer name</th>
<th>Sequence(5'----&gt; 3' order)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ndhJ</td>
<td>LP1-For</td>
<td>CATAGACCATTTTRGGTTTYGA</td>
</tr>
<tr>
<td>ndhJ</td>
<td>LP4-Rev</td>
<td>ACCAATCCAAATATCRGGGC</td>
</tr>
<tr>
<td>rpoC1</td>
<td>1-For</td>
<td>GTGGATACACCTTTGGATAATGG</td>
</tr>
<tr>
<td>rpoC1</td>
<td>3-Rev</td>
<td>TGGAGAAAACATAAGTAAACGGGC</td>
</tr>
<tr>
<td>rpoB</td>
<td>1-For</td>
<td>AAGTGCAATTTGGAACGTGG</td>
</tr>
<tr>
<td>rpoB</td>
<td>3-Rev</td>
<td>CGTATGAAAAGAGTGAA</td>
</tr>
<tr>
<td>YCF5</td>
<td>1-For</td>
<td>GATTATTAGTCACCTCGATG</td>
</tr>
<tr>
<td>YCF5</td>
<td>3-Rev</td>
<td>ACTTACGTCATCATTTACCA</td>
</tr>
<tr>
<td>rbcL</td>
<td>rbcLa-SI_For</td>
<td>ATGTCACCCACAAACAGAGACTAAAGC</td>
</tr>
<tr>
<td>rbcL</td>
<td>rbcL_SI_Rev</td>
<td>GCAAATGACGTCAGTGC</td>
</tr>
<tr>
<td>trnH-psbA</td>
<td>psbA_For</td>
<td>GTTATGCATGAAGTATGGTCT</td>
</tr>
<tr>
<td>trnH-psbA</td>
<td>trnH_Rev</td>
<td>CGCGCATGGTGATTTCACAA</td>
</tr>
<tr>
<td>matK</td>
<td>kew-matk-intR</td>
<td>AGAGACGCCCCTTTGATGAA</td>
</tr>
<tr>
<td>matK</td>
<td>kew-matk-intF</td>
<td>TTGATCCAAAGGCGGTCT</td>
</tr>
</tbody>
</table>

The PCR thermal cycle consisted of one cycle of 5 min at 94°C; 35cycles of 30s at 94°C, 30s at 55°C, 60s at 72°C, and a final cycle of 10 min at 72°C. PCR products were additionally purified before sequencing. Direct sequencing of (diluted) PCR was performed. Data was imported into MEGA 4.0 where sequences were aligned using ClustalW (Thompson et al., 1994). Then genetic distances were calculated according to the p-distance model. The average intra-specific distance was calculated to determine the intra-specific variation by the p-distance.
model (Meyer and Paulay, 2005). Moreover, the average and range of inter-specific distance were used to investigate inter-specific divergences (Chen et al., 2010; Meier et al., 2008; Meyer and Paulay, 2005). Dendrograms were constructed based on the sequence alignments by neighbor-joining using p-distance in the MEGA 4.0 software.

RESULTS

Two barcode loci of trnH-psbA and matK were not amplified well in all studied species. Therefore, they were not used for further analyses.

**Allium cepa** L.

Three loci including ndhJ, rpoC1 and YCF5 had high Species Rate of Amplification Percentage (SRAP) (80-100%) and sequencing (80-100%) for **A. cepa** (Table 3). rpoB locus had 50% amplification and 75% sequencing. rbcL locus showed 100% amplification but appropriate results were not achieved for sequencing (0%) (Table 3).

**Allium schoenoprasum** L.

SRAP ranged between 40-100% in **A. schoenoprasum** L. for ndhJ, rpoC1, rpoB, YCF5 and rbcL. The highest SRAP observed in YCF5 locus (100%) and the lowest one occurred for rpoB (40%) (Table 3). Their Species Rate of Sequencing Percentage (SRSP) also ranged between 0-100% which 100% of sequencing in locus ndhJ and 0% for rbcL (Table 3).

**Allium ampoloprasum persicum** L.

High percentages of amplification for species of **A. ampoloprasum persicum** shown in four barcode loci of ndhJ, rpoC1, YCF5 and rbcL with 100% amplification. The rpoB locus, however, showed 20% of amplification (Table 3). Sequencing results ranged between 0-100%. The highest SRSP (100%) was shown for rpoC1, rpoB and YCF5. 0% SRSP was shown for rbcL (Table 3).

**ndhJ**

Locus of ndhJ with 318 base pair length had 85.71% and 92.85% success in amplification and sequencing (Table 4). This barcode locus had 8 variable sites, and only one species, **A. ampoloprasum persicum**, could be distinguished by unique SNPs in 70th and 280th positions and had 1 deletion type indels, in 1st position. These SNPs contained G for **A. ampoloprasum persicum** as for two other species contained C. The intra specific divergence for **A. cepa**, **A. schoenoprasum** L. and **A. ampoloprasum persicum** were 0.2%, 0.2% and 0.8%. The distance between groups ranged between 0.2-1.1% with the average value of 0.8. Dendrogram of this locus (Figure 1) represents that using ndhJ locus had led to **A. ampoloprasum persicum** identification among other studied species.
Table 3. Results of PCR and sequencing of barcoding loci for *A. cepa, A. schoenoprasum* L. and *A. ampeloprasum persicum*

<table>
<thead>
<tr>
<th>Barcode region</th>
<th>Species</th>
<th>PCR success (%)</th>
<th>Sequencing success (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ndhJ</td>
<td><em>A. cepa</em></td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td><em>A. schoenoprasum</em> L.</td>
<td>66.66</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td><em>A. ampeloprasum persicum</em></td>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td><em>A. cepa</em></td>
<td>80</td>
<td>100</td>
</tr>
<tr>
<td>rpoC1</td>
<td><em>A. schoenoprasum</em> L.</td>
<td>60</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td><em>A. ampeloprasum persicum</em></td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td><em>A. cepa</em></td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td>rpoB</td>
<td><em>A. schoenoprasum</em> L.</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td><em>A. ampeloprasum persicum</em></td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td><em>A. cepa</em></td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>YCF5</td>
<td><em>A. schoenoprasum</em> L.</td>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td><em>A. ampeloprasum persicum</em></td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td><em>A. cepa</em></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>rbcL</td>
<td><em>A. schoenoprasum</em> L.</td>
<td>66.66</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td><em>A. ampeloprasum persicum</em></td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td><em>A. cepa</em></td>
<td>0</td>
<td>. . .</td>
</tr>
<tr>
<td>tmH-psbA</td>
<td><em>A. schoenoprasum</em> L.</td>
<td>0</td>
<td>. . .</td>
</tr>
<tr>
<td></td>
<td><em>A. ampeloprasum persicum</em></td>
<td>0</td>
<td>. . .</td>
</tr>
<tr>
<td></td>
<td><em>A. cepa</em></td>
<td>0</td>
<td>. . .</td>
</tr>
<tr>
<td>matk</td>
<td><em>A. schoenoprasum</em> L.</td>
<td>0</td>
<td>. . .</td>
</tr>
<tr>
<td></td>
<td><em>A. ampeloprasum persicum</em></td>
<td>0</td>
<td>. . .</td>
</tr>
</tbody>
</table>

**rpoC1**

Subunit locus *rpoC1* with 412 base pair had 92.85% and 92.85% success in amplification and sequencing (Table 4). It had 4 variable sites in which the species of *A. ampeloprasum persicum* had unique SNPs in 228th and 318th positions. These SNPs contained C for *A. ampeloprasum persicum* but contained T for two other species. Species of *A. schoenoprasum* L. had 2 insertion type indels in 1st and 2nd positions and *A. ampeloprasum persicum* had 6 insertion type indels in 1st, 408th, 409th, 410th, 411th, and 412th positions. The intra specific divergence for *A. cepa, A. schoenoprasum* L. and *A. ampeloprasum persicum* were 0.1%. The distances between groups varied between 0.1-0.6 and average of 0.4. Dendrogram of this locus (Figure 2) clarifies that the only species which was discriminated and identified accurately is *A. ampeloprasum persicum*.

Table 4. Total results of PCR and sequencing of barcode regions

<table>
<thead>
<tr>
<th>Barcode region</th>
<th>PCR success (%)</th>
<th>Sequencing success (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ndhJ</td>
<td>85.71</td>
<td>92.85</td>
</tr>
<tr>
<td>rpoC1</td>
<td>92.85</td>
<td>92.85</td>
</tr>
<tr>
<td>rpoB</td>
<td>35.71</td>
<td>76.92</td>
</tr>
<tr>
<td>YCF5</td>
<td>85.71</td>
<td>93.33</td>
</tr>
<tr>
<td>rbcL</td>
<td>92.85</td>
<td>0</td>
</tr>
</tbody>
</table>

**rpoB**

*rpoB* locus with 381 base pair length had 35.71% and 76.92 % success in amplification and sequencing (Table 4). There were 3 variable sites and the unique SNPs in 196th and 293th positions, belonged to *A. ampeloprasum persicum*. Position of 196 contained C for *A. ampeloprasum persicum* and T for two other species and position of 293 contained T for *A. ampeloprasum persicum* and C for two other species. Species of *A. cepa* had 2 insertion type idels in 380th and 381th positions, *A. schoenoprasum* L. had 3 insertion type indels in 379th, 380th and 381th positions and *A. ampeloprasum persicum* had 2 deletion type indels in 380th and 381th positions. The intra specific divergence for *A. cepa, A. schoenoprasum* L. and *A. ampeloprasum persicum* were 0%. The distance between groups ranged between 0-0.5 with the average of 0.3. In this locus *A. ampeloprasum persicum* could have distinguished out of other two species. It means all samples of *A. ampeloprasum persicum* were located in a separated cluster Dendrogeram of this primer (Figure 3).
YCF5

For YCF5 locus with 370 base pair length 85.71% and 93.33 % success were observed in amplification and sequencing (Table 4). It had 3 variable sites. There were unique SNPs for only species of A. ampeloprasum persicum in 146th, 164th and 261th positions, which contained T, T, and A for A. ampeloprasum persicum while hold C, C, and G for two other species. No indel was detected using YCF5. The intra specific divergence for A. cepa, A. schoenoprasum L. and A. ampeloprasum persicum were 0%, 0.4% and 0%. The distances between groups were 0.2-0.8% and the average was 0.5. Dendrogram of this region indicates that A. ampeloprasum persicum was distinguished. However some confusions were raised from cluster dendrogram analysis and species of A. schoenoprasum L. could not be distinguished out of A. ampeloprasum persicum in some samples (Figure 4).

rbcL

rbcL locus with 484 nucleotide pairs had 92.85% amplification but no success in sequencing (data not shown) (Table 4).

Figure 1. Neighbor-joining tree based on the ndhJ region using the p-distance -parameter model. Bootstrap values (>50%) are shown above the relevant branches. Voucher numbers follow the species names.
Figure 2. Neighbor-joining tree based on the rpoC1 region using the p-distance -parameter model. Bootstrap values (>50%) are shown above the relevant branches. Voucher numbers follow the species names.

Figure 3. Neighbor-joining tree based on the rpoB region using the p-distance -parameter model. Bootstrap values (>50%) are shown above the relevant branches. Voucher numbers follow the species names.
Figure 4. Neighbor-joining tree based on the YCF5 region using the p-distance -parameter model. Bootstrap values (>50%) are shown above the relevant branches. Voucher numbers follow the species names.

**DISCUSSION**

Based on the results, *A. ampeloprasum persicum* was discriminated from *A. cepa* and *A. schoenoprasum* L. in loci of *ndhJ*, *rpoC1* and *rpoB*, while two species of *A. cepa* and *A. schoenoprasum* L. could not be separated based on these barcodes. It seems that *YCF5* had little identification ability for the three species as some confusions observed in sequencing results. It seems it has been shown little discriminatory region in *Allium*. *rbcl* did not work for sequencing. The *ndhJ* proved the most effective barcoding locus for current species, tested in present study. From three studied species, *Allium* genus the lowest intraspecific divergence (average mean divergence) was seen in barcode *rpoB* and the highest one was shown in *ndhJ* barcoding locus.

Studies of Jarret (2008) on five cultivated *Capsicum* species showed that the *trnC-rpoB* could not separate *C. chacoense* (Hunz.) from *C. annuum* L., while it was effective in separating the other taxa from each other. Investigations on species of genus *Schizolobium* exhibited *YCF5* could not separate different species despite wide diversity between them (Turchetto-Zolet, 2009). Studies on different species of *Myrtacea* showed that *rpoB*, 407 base pair in length and 49 variable sites, could discriminate the species of two genera of *Hexachlamys* and *Eugenia*. The phylogenetic dendrogram of the barcoding locus however added confusion due to existence of the inter specific similarities (Da-Cruz, 2012). Lahaye et al. (2008) reported in orchids, the barcodes that signify the lowest intraspecific divergence (average mean divergence), were *rpoC1 accD/matK* and *matK* that calculations of intraspecific divergence did not somehow show a clear result. In the Kruger National Park, the lowest intraspecific divergence was provided by *ndhJ* with different metrics. *trnH-psbA* has been shown to be the least discriminatory region in *Aspalathus* (Fabaceae) when compared with the *ITS* and *trnT-trnL* intergenic spacers (Edwards et al., 2008). Several potential challenges can prevent DNA barcoding from being a reliable method for plant identification. For example, medicinal plants contain secondary compounds, such as tannins, alkaloids, and polysaccharides, which could be problematic by inhibiting the DNA extraction and amplification processes. On the other hand the primers might show high-quality results during the amplification phase (PCR), but may not be specific enough to work in sequencing experiments. Cycle sequencing may fail despite successful PCR, if the primer sequence is not an exact match to the primer region (Schori and Showalter, 2011). Studies of Ghahramanzadeh et al. (2013) showed that the *rbcl* locus was the easiest to sequence and align, but comes short for showing enough variation to enable identifying the tested species.

Despite various studies which reported very low divergence for *rbcl* (Kress et al., 2005; Newmaster et al., 2008), especially in very similar species, Liu et al. (2010) represented that *rbcl* could be suitable in Bryophyta barcoding. Studies of Cowan et al. (2006) on 96 species of *Sinningia* s.l. (Gesneriaceae) revealed 95% probability of correct identification using *trnS-trnG*, *trnT-trnL*, *rpl16*, *trnL-trnF*, *atpB-rbcL* and *ncpGS*. Kress et al. (2005)
reported 119 to >1,000 bp in length of trnH-psbA spacer among the studied angiosperm taxa. This length of variation, can lead to difficulties in alignment, which could cause problems for DNA barcoding due to the high number of insertion/deletion in trnH-psbA.

94% match between two trnH-psbA sequences for Trigonella foenum-graecum L. may cause intraspecific variation or misidentification of one voucher. Multiple voucher sequences should be used for each different barcode, especially for barcoding regions that are known very similar (Schori and Showalter, 2011).

CONCLUSION

Based on results of this study A. ampeloprasum persicum was the only identified species. Barcode locus of YCF5 had the highest sequencing percentage with 93.33%. The highest interspecific diversity and intraspecific divergence with 0.8% and 0.4% were observed in locus of ndhJ which was the most sufficient barcoding locus for species identification.

It seems that two species of A. cepa and A. schoenoprasum L. cannot be distinguished from each other, based on the used barcoding loci. For more accurate identification of species of Allium genus supplementary primers combinations are needed for each barcoding locus in future studies and also using more samples is recommended.

REFERENCES

Anvarkhah S. 2013. Seed identification of medicinal plants using advanced technologies. PhD thesis. Department of Crop Science, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran.


