The perspectives for DNA barcoding of *Rhaponticum carthamoides* (Willd.) Iljin using *rbcL* gene sequence

M V Protopopova1, N A Shvetsova1,2 and V V Pavlichenko1

1 Siberian Institute of Plant Physiology and Biochemistry SB RAS, Irkutsk, 664033 Russia
2 Irkutsk State University, Irkutsk, 664003 Russia

E-mail: marina.v.protopopova@gmail.com

Abstract. The methods of biological species identification using nucleotide sequences of short genome regions (DNA barcoding) are actively developed. The universal DNA barcode for plants remains to be discovered, and one of the leading candidates is the plastid gene of the large subunit of ribulose-bisphosphate carboxylase gene (*rbcL*). In our study, we estimated the part of *rbcL* gene as a possible marker for molecular identification of *Rhaponticum carthamoides* (Willd.) Iljin. Due to its official properties, the species is susceptible to uncontrolled and illegal harvesting from natural populations. Today, the species needs to be protected and therefore is included into the Red Data Books of the Russian Federation and certain regions. The study was carried out using plants from the natural populations sampled from the Khamar-Daban Ridge (South Siberia) and considering now as *Rh. carthamoides* var. *chamarense* (Peschkova) O S Zhirova. It was shown that *rbcL* gene can be used to identify *Rh. carthamoides* at least from the populations of the Khamar-Daban Ridge using a fragment of the maximum length or its 3’ region. Apparently, the 5’ region of the gene (*rbcLa*) most often used as DNA barcode for plants may be of lesser importance for *Rh. carthamoides*. The *rbcL* gene sequences can be also used for the development of approaches for *Rh. carthamoides* identification in the medicinal preparations and products containing dried tissues to prevent their falsification and illegal harvesting of this species. The combination of *rbcL* gene with additional markers seems to be highly desirable to create effective DNA barcodes for *Rhaponticum* species.

1. Introduction
At present, the methods of molecular identification of species are being actively developed. The use of such approaches is especially important if the determination of biological samples by morphological criteria is difficult, including cases with damaged, incomplete, or even fragmented organisms. The method principle is to determine the unique nucleotide sequences of short genome regions (molecular markers), which will play a role of the genetic barcode of the species. This idea formed the basis of the genetic database purposed the identification for all living organisms – Barcode of Life Data System (BOLD), and the method was called DNA barcoding [1]. This approach may be very useful for monitoring the distribution of natural populations of vulnerable and endangered species; it also enables effective control of the origin of biological raw materials or finished products based on it in order to counteract falsification and illegal harvesting of rare species [2].

For DNA barcoding of animal species is based on the nucleotide sequence of the 5’ fragment encoding the first subunit of cytochrome C oxidase (CO1 or *cox1*) [1]. Attempts to use a universal marker for plants were less successful [1, 3, 4] and the development of taxon-specific methods is required [2].
The aim of this work was to develop approaches for the identification of *Rhaponticum carthamoides* (Willd.) Iljin (Asteraceae) using DNA barcoding.

*Rh. carthamoides* [=*Leuzea carthamoides* (Willd.) DC. =*Stemmacantha carthamoides* (Willd.) Dittrich =*Fornicium carthamoides* (Willd.) Kamelin] commonly known as a ‘maral root’ or Russian leuzea is a vulnerable medicinal species and is included into the Red Data Books of the Russian Federation (2008) and of most regions where it grows. The species range covers the mountains of South Siberia (the Altai, the Kuznetsky Alatau, the Western and Eastern Sayan mountains, the Republic of Tyva’ mountains, the Khamar-Daban Ridge), and is also found in Central and Eastern Asia. The extract and individual compounds isolated from different parts of the plant have specific biological effects as immunostimulation, eliminating free radicals to prevent oxidizing pathology, increasing protein biosynthesis and physical work capacity along with endurance and performance, enhancing cardiovascular functions, and mental work capacity [5]. The plant rhizomes and roots are mainly used as raw material for the production of balms, elixirs, extracts, tablets, capsules, and etc. *Rh. carthamoides*, like other medicinal plants, is subject to falsification. Therefore, the development of approaches to its identification is necessary to counter the illegal harvesting of this species and the falsification of goods and raw materials based on it. Taking into account the distribution of the species, DNA-barcoding can be effectively used to establish the region of origin of plant raw materials.

In the present study, we used the partial sequence of the plastid gene of the large subunit of the ribulose-bisphosphate carboxylase gene (*rbcL*) as a molecular marker for DNA based identification of *Rh. carthamoides*. *RbcL* gene is one of the leading candidates to the universal DNA barcode for plants and was recommended for molecular identification of plant species by Consortium for the Barcode of Life’s Plant Working Group [4]. For phylogenetic studies, the full-length *rbcL* gene is often used. For DNA barcoding the use of an approximately 650-bp fragment at the 5’ end of the *rbcL* gene (is also termed as *rbcLa*) is proposed [4], however, it was shown that 3’ fragment or region partially overlapping of these sequences (is also termed as *rbcLb*) might be also sufficiently variable to identify the species [6]. In our study, we compared the efficiency of using the fragment of *rbcL* gene of the maximum length and three types of shorter regions.

2. Materials and methods

2.1. Sampling

Development of approaches for DNA barcoding of *Rh. carthamoides* was carried out using plants collected from the natural populations on the Khamar-Daban Ridge (the upper courses of the Khara-Murin and the Snezhnaya river basins, the Lake Baikal basin, and South Siberia). The plants from the Khamar-Daban Ridge are now considered to belong to a separated variety of the species – *Rh. carthamoides* var. *chamarense* (Peschkova) O S Zhirova. At least six separate individuals were collected from each population and were kept in an individual filter paper bag prior to DNA isolation. Fresh plant material was dried and stored in silica gel.

2.2. DNA isolation, PCR and sequencing

Total DNA was isolated from silica gel-dried leaf tissue following the cetyltrimethylammonium bromide (CTAB) method [7], with modifications [2] and RNase A treatment for RNA removal.

Amplification of *rbcL* was performed using GoTaq Flexi DNA polymerase (Promega) and aF and cR primers for *Nicotiana tabacum* [8] in a reaction mixture of 20 µL containing 1x Green GoTaq Flexi Buffer, 1 unit of GoTaq Flexi DNA polymerase (Promega) and final concentrations of 2.5 mM of MgCl2, 250 µM of each dNTP, and 250 nM of each primer. The amplification efficiency was assessed using electrophoresis in 1% agarose gel by the intensity of PCR products staining by ethidium bromide. Obtained amplicons were gel-purified using the GeneJET Gel Extraction Kit (Thermo Fisher Scientific) and directly sequenced by Sanger method using BigDye Terminator Cycle Sequencing kit v. 3.1 (Applied Biosystems) on 3500 Genetic Analyzer (Applied Biosystems).
2.3. Sequence alignment and phylogenetic analysis

The raw sequencing data were primarily edited using SnapGene Viewer software v.2.6.2 (GSL Biotech), primers and 5’ and 3’ low-quality regions of reads were clipped and consensus to each sequence was generated based on data obtained from the forward and the reverse direction. GenBank accession numbers of sequences obtained in our study and the reference sequences are presented in Table 1. As references for phylogenetic reconstruction several *Rhaponticum* species were used. As out-group the clade combining *Syreitschikovia spinulosa* and *Saussurea salwinensis* was used.

| Species name [9]                                                                 | GenBank accession number¹   |
|---------------------------------------------------------------------------------|-----------------------------|
| *Rh. carthamoides* var. *chamarense* (Peschkova) O S Zhirova (the Snezhnaya River isolate) | MZ346034                    |
| *Rh. carthamoides* var. *chamarense* (the Khara-Murin River isolate)             | MZ346035                    |
| *Rh. uniflorum* (L.) DC.                                                         | MW683229.1                  |
| *Rh. cf. carthamoides* (Willd.) Iljin                                             | MN919077.1                  |
| *Rh. cf. scariosum* Lam.                                                         | MN919078.1                  |
| *R. australe* (Gaudich.) Soskov                                                | KC589888.1                  |
| *Rh. coniferum* (L.) Greuter                                                    | MN919079.1                  |
| *Rh. acaule* (L.) DC.                                                           | KC589887.1                  |
| *Rh. repens* (L.) Hidalgo                                                       | KC589889.1                  |
| *Syreitschikovia spinulosa* (Franch.) Pavlov                                    | KC589902.1                  |
| *Saussurea salwinensis* J. Anthony                                               | NC_044731.1                 |

¹ The GenBank accession numbers of the original sequences obtained in this study are highlighted in bold typeface.

The multiple alignment of nucleotide sequences by the MUSCLE application with gap opening penalty equalled 100 and extension penalty equalled 1.01 was conducted in MEGA software v. 7.0.16 [10] followed by manual editing. The part that contained primers and missing data on 5’ and 3’ ends were completely removed from the analysis. To assess the possibility of identifying sequences by clustering the phylogenetic analysis was performed by both the neighbour-joining method (NJ) in MEGA and the Bayesian inference method implemented in MrBayes v. 3.2.5 [11]. Phylogenetic analyses were created based on fragments of *rbcL*α, *rbcL*β, 3’ region of *rbcL* separately and on the fragment of *rbcL* of the maximum length. For NJ-phylogeny the uncorrected paired p-distances with no among-site rate variation was used. A bootstrap of 1,000 replicates was used as a test of the phylogeny. Bayesian inference was performed using F81-like models [12] with the equal stationary state frequencies in MrBayes. Markov chain Monte Carlo (MCMC) analyses were run for 1,000,000 generations, with four simultaneous chains with sampling every 100 generations and diagnostic calculation every 1,000 generations. Finally, a phylogram was constructed from the posterior distribution of trees and edited in FigTree v.1.4.3 [13]. The ‘best close match’ test was used to determine the closest correspondence of a target sequence with all of the others in a set of aligned data [14]. The marker efficiencies in species identification were estimated by a portion of sequences that were classified as “correct” when the genetic distances between the target sequence and other sequences of the same species fit within 95% of the calculated limit [15].
3. Results and discussion

The length of obtained consensus sequences of \textit{rbcL} for \textit{Rh. carthamoides} after the editing and clipping was 1,311 bp. The size of the multiple alignment including the sequences of \textit{rbcL} gene for \textit{Rh. carthamoides} and related species after removal of the missing data on 5’ and 3’ ends was 1,288 bp, of which 18 variable sites (1.4%) and only 6 parsimony-informative sites were found (0.47%) (figure 1).

\textbf{Figure 1.} The parts of multiple alignment of \textit{rbcL} nucleotide sequences of several \textit{Rhaponticum} species.

Four types of nucleotides are marked with dots coloured in different shades of grey range. The dots indicate the same base with the reference sequence (\textit{Rh. carthamoides} var. \textit{chamarense} from the Snezhnaya River). Colour dots and letters indicate the polymorphic sites containing the mismatches comparatively to reference one. The stars indicated the parsimony-informative sites. Numbers indicate the positions in the multiple alignment.

Analysis of the \textit{rbcL} gene region in eight \textit{Rhaponticum} species revealed a rather low level of overall interspecies polymorphism (p<0.01) within the genus. Sequence analysis of two populations of \textit{Rh. carthamoides} from the Khamar-Daban Ridge did not reveal the presence of intraspecific polymorphism and species-specific sites (figure 1).

The most structured cladogram was obtained when analysing the fragment of \textit{rbcL} gene of the maximum length, 1,288 bp (figure 2). The populations of \textit{Rh. carthamoides} from the Khamar-Daban Ridge were combined into a supported clade both on the NJ (bootstrap, 74) and the Bayesian tree (posterior probability, 1.00). The \textit{rbcL} sequence listed in GenBank as FZ-2020 isolate of \textit{Rh. cf. carthamoides} (GenBank No: MN919077.1) significantly (for 4 out of 6 informative sites) differed from those obtained by us for this species (figure 1) and was combined with \textit{Rh. scariosum}, \textit{Rh. australie} and \textit{Rh. coniferum} on the Bayesian tree (posterior probability, 1.00). Using phylogeny based on ITS1-ITS2 (the data are not presented here) of the same isolate (FZ-2020) of \textit{Rh. cf. carthamoides} (GenBank No: MN918981.1), it was found that that ambiguous \textit{rbcL} sequence most likely belongs to \textit{Rh. cynaroides} (L.) O Bolòs (sequences identity, 100%). Clade combining our original sequences of \textit{Rh. carthamoides} from the Khamar-Daban Ridge forms a common group with \textit{Rh. uniflorum} which well correlates with the study dedicated to the phylogeny of \textit{Rhaponticum} [16]. However, there are mismatched patterns on the trees, for instance, in our study Australian \textit{Rh. australie}
nested in the clade combining European *Rh. cf. scariosum*, *Rh. coniferum*, and presumably *Rh. cynaroides* (corresponding to *Rh. cf. carthamoides* in the figure) but not in the clade combining central and east Asian species including *Rh. carthamoides* and *Rh. uniflorum* as was shown in the study of D Hidalgo with the colleagues [16].

![Cladogram of rbcL gene fragment of the maximum length based on both NJ and Bayesian 50% majority-rule consensus trees.](image)

Bootstrap values are marked above branches, posterior probabilities are below branches. The star indicates the species presumably incorrectly determined as *Rhaponticum carthamoides* (GenBank: MN919077.1).

The cladograms built based on the fragments of *rbcLa* (the length, 545 bp), *rbcLb* (the length, 861), and 3’ region (the length, 743) of *rbcL* separately showed a low efficiency of the sequences clusterization on both NJ and Bayesian trees (the data are not presented here). In none of the trees obtained, the sequences of *Rh. carthamoides* from the Khamar-Daban Ridge were combined into a separated clade, and all node supports were lower than in the tree built based on the *rbcL* fragment of the maximum length. Analysis based on *rbcLa* fragment showed that the *rbcL* sequences of *Rh. carthamoides* from the Khamar-Daban Ridge were nested in the common clade with the sequence of the species which was presumably incorrectly determined as *Rhaponticum carthamoides*. In the tree built based on the *rbcLb* region most of the branches, including the populations of *Rh. carthamoides* from the Khamar-Daban Ridge, turned out to be unresolved. On the tree built based on 3’ region of *rbcL* gene, the sequences of the populations of *Rh. carthamoides* from the Khamar-Daban Ridge were combined with *Rh. uniflorum*, and that tree was generally in line with expectations and had a very similar structure as the tree built at the fragment of the maximum length.

The best close match criterion showed that correct identification possible for at least 40% of sequences belonging to *Rhaponticum* using the fragment of *rbcL* of the maximum length; 17% if using 3’ region, and 0% if using *rbcLa* and *rbcLb* regions. The successful match between sequences of *Rh. carthamoides* from the Khamar-Daban Ridge appeared only in the case of using the fragment of *rbcL* of the maximum length and its 3’ region. In the case of using *rbcLa* and *rbcLb* regions, the sequences of *Rh. carthamoides* from the Khamar-Daban Ridge matched to the multiple species, and unambiguous identification was impossible.
4. Conclusion

RbcL gene can be used to identify *Rhaponticum carthamoides* using a fragment of the maximum length or its 3’ region. Apparently, the 5’ region of the gene (rbcLa) most often used as DNA barcode for plants may be of lesser importance for *Rh. carthamoides*. The rbcL gene sequences can be also used for the development of approaches for *Rh. carthamoides* identification in the medicinal preparations and products containing dried tissues to prevent their falsification and illegal harvesting of this species. The combination of rbcL gene with additional markers seems to be highly desirable to create effective DNA barcodes for *Rhaponticum* species.

Acknowledgements

The research was done using the equipment of the Core Facilities Centre ‘Bioanalitika’ and collections of the Core Facilities Centre ‘Bioresource Centre’ and within the framework of the state assignments of Siberian Institute of Plant Physiology and Biochemistry, Siberian Branch, Russian Academy of Sciences (No. 121031300009-4). The research was partially supported by Irkutsk State University, project No. 091-21-319 ‘Development of DNA barcoding approaches for species identification of official plants with *Rhaponticum carthamoides* as an example’. The authors also thank V V Chepinoga for the help with sampling and species identification.

References

[1] Hebert P D N, Cywinska A, Ball S L and Dewaard J R 2003 Biological identifications through DNA barcodes *Proceedings of the Royal Society of London B* **270**(1512) 313–21

[2] Protopopova M, Pavlichenko V, Gnutikov A and Chepinoga V 2019 DNA Barcoding of *Waldsteinia* Wild. (Rosaceae) species based on ITS and *trnH-psbA* nucleotide sequences *Information Technologies in the Research of Biodiversity. Springer Proc. in Earth and Environmental Sciences* ed I Bychkov and V Voronin (Cham: Springer) pp 107–15

[3] Chase M W, Salamin N, Wilkinson M, Dunwell J M, Kesanakurthi R P, Haidar N and Savolainen V 2005 Land plants and DNA barcodes: short-term and long-term goals *Philosophical Transactions of the Royal Society of London Biological Sciences* **360**(1462) 1889–95

[4] CBOL Plant Working Group 2009 A DNA barcode for land plants *Proc. of the National Academy of Sciences* (USA) **106**(31) 12794–7

[5] Kokoska L and Janovska D 2009 Chemistry and pharmacology of *Rhaponticum carthamoides*: a review *Phytochemistry* **70** 842–55

[6] Dong W, Cheng T, Li C, Xu C, Long P, Chen C and Zhou S 2014 Discriminating plants using the DNA barcode rbcLb: an appraisal based on a large data set *Molecular Ecology Resources* **14**(2) 336–43

[7] Doyle J J and Doyle J L 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue *Phytochemical Bulletin* **19** 11–5

[8] Hasebe M, Omori T, Nakazawa M, Sano T, Kato M and Iwatsuki T 1994 rbcL gene sequences provide evidence for the evolutionary lineages of leptosporangiate ferns *Proc. of the National Academy of Sciences* (USA) **91** 5730–4

[9] Hassler M 2021 World Plants: Synonymic checklists of the vascular plants of the World (version 2021-03-17) *Species 2000 & ITIS Catalogue of Life* (Leiden: Naturalis)

[10] Kumar S, Stecher G and Tamura K 2016 MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets *Molecular Biology and Evolution* **33**(7) 1870–4

[11] Ronquist F, Teslenko M, van der Mark P, Ayres D L, Darling A, Hohna S, Larget B, Liu L, Suchard M A and Huelsenbeck J P 2012 MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space *Systematic Biology* **61**(3) 539–42

[12] Felsenstein J 1981 Evolutionary trees from DNA sequences: A maximum likelihood approach *Journal of Molecular Evolution* **17**(6) 368–76
[13] Rambaut A 2006-2016 FigTree: Tree figure drawing tool, version 1.4.3 (Edinburgh: Institute of Evolutionary Biology, University of Edinburgh) http://tree.bio.ed.ac.uk (accessed 30.09.2021)

[14] Meier R, Shiyang K, Vaidya G and Ng P K L 2006 DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success Systematic Biology 55(5) 715–28

[15] Bolson M, de Camargo Smidt E, Brotto M L and Silva-Pereira V 2015 ITS and trnH-psbA as efficient DNA barcodes to identify threatened commercial woody Angiosperms from Southern Brazilian Atlantic rainforests PLOS ONE 10(12) e0143049

[16] Hidalgo O, Garcia-Jacas N, Garnatje T and Susanna A 2006 Phylogeny of Rhaponticum (Asteraceae, Cardueae-Centaureinae) and related genera inferred from nuclear and chloroplast DNA sequence data: taxonomic and biogeographic implications Annals of Botany 97(5) 705–14