DOI: 10.32615/bp.2019.101 BIOLOGIA PLANTARUM 63: 565-574, 2019

*This is an open access article distributed under the terms of the Creative Commons BY-NC-ND Licence*

# Comparative analysis of complete chloroplast genomes sequences of *Arctium lappa* and *A. tomentosum*

Y.-P. XING[1], L. XU[1]*, S.-Y. CHEN[1], Y.-M. LIANG[1], J.-H. WANG[1], C.-S. LIU[2], T. LIU[3], and T.-G. KANG[1]*

*College of Traditional Chinese Medicine, Liaoning University of Traditional Chinese Medicine, Dalian 116600, P.R. China[1]*
*College of Traditional Chinese Medicine, Beijing University of Chinese Medicine, Beijing 100102, P.R. China[2]*
*School of Pharmacy, China Medical University, Shenyang 110122, P.R. China[3]*

## Abstract

*Arctium lappa* and *A. tomentosum* are known medicinal plants in China. The complete chloroplast genomes from *A. lappa* and *A. tomentosum* were sequenced using *Illumina* sequencing technology. The total genome sizes of the complete chloroplast genomes of *A. lappa* and *A. tomentosum* were 152 767 bp and 152 688 bp, respectively, and contained a pair of inverted repeats of the same length (15,181 bp). The small single-copies were 18 584 bp and 18 582 bp, and the large single-copies were 83 821 bp and 83 744 bp, respectively. We identified and annotated 134 and 126 genes from *A. lappa* and *A. tomentosum* including, respectively, 90 and 89 protein-coding genes, 36 and 29 tRNAs, and eight rRNAs. *A. lappa* was found to have 10 tRNAs different from those in *A. tomentosum*, and *A. tomentosum* had three tRNAs different from those in *A. lappa.* There were only two types of simple sequence repeats of two species, mononucleotide and dinucleotide, and the sequences were A and T rich. In addition, the two ways of phylogenetic analysis show that the position of *A. lappa* and *A. tomentosum* is consistent within *Asteraceae*.

*Additional key words*: medicinal plants, phylogenetic analysis, simple sequence repeats.

## Introduction

There are approximately 11 species of *Arctium* (*Asteraceae*) in the world, and they are widely distributed in temperate regions of Eurasia. *Arctium lappa* L. and *A. tomentosum* Mill. are distributed in China and used as medicinal plants. Recent studies of *Arctium* from China have focused on medicinal quality, internal transcribed spacer sequence identification, chemical composition, and pharmacological effects (Kang *et al*. 2009, Xu *et al*. 2011), but limited research has been focused on chloroplast (cp) genome identification, phylogenetic position, and species diversity.

In general, the cp genomes of vascular plants share a similar structure and are highly conserved (Jeffrey *et al.* 1982, Jansen *et al.* 2005, Choi *et al.* 2015). The highly conservative structure of the cp genome makes it suitable for comparative analysis of distant and closely related species (Raubeson and Jansen 2005). The cp genome size varies between 120 and 220 kb, with one inverted repeat (IR) region (20 - 28 kb), a large single copy (LSC) region (80 - 90 kb), and a small single copy (SSC) region (16 - 27 kb) (Wicke *et al*. 2017). The cp genome usually contains about 110 - 130 genes, including about 79 protein coding genes, about 30 transfer RNA (tRNA) encoding genes, and about 4 ribosomal RNA (rRNA) encoding genes. They are mainly involved in photosynthesis or gene expression (Jansen *et al.* 2005, Yang *et al.* 2013). With the emergence and development of high-throughput sequencing technology, more and more cp genomes of *Asteraceae* species have been dissected recently. Previous studies have demonstrated that the *Asteraceae* cp genomes had a relatively conservative number of genes and the size of intergenic regions contributed mainly to the variation of cp genome size (Wang *et al.* 2015). In the *Asteraceae*, the cp genomes share a 22.8 kb large inversion and a 3.3 kb

*Submitted* 1 March 2018, *last revision* 7 October 2018, *accepted* 11 December 2018.
*Abbreviations*: cp - chloroplast; IR - inverted repeat; LSC - large single copy; SNP - single nucleotide polymorphism; SSC - small single copy; SSR - simple sequence repeat; ML - maximum-likelihood.
*Acknowledgements*: This work was supported by the National Natural Science Foundation of China (General Program, Nos81773852 and 81874338) and the Liaoning Province Education Department (Liaoning Higher School Outstanding Young Scholar Growth Plan, No. LJQ2014101).
* Corresponding authors e-mails: 861364054@qq.com, kangtingguo@163.com

smaller inversion nested within it (Kim *et al*. 2005). Many genes from cp genomes, such as *psb*A-*trn*H, *trn*C-*ycf*6, *ycf*6-*psb*M, *trn*Y-*rpo*B, *rps*4-*trn*T, *trn*L-F, *rpl*16, *mat*K, and *rbc*L have been used for analysis of the phylogeny of *Chrysanthemum*, *A. lappa* and other plants (Liu *et al.* 2012, Tseng *et al*. 2012, Kim *et al.* 2016). Simple sequence repeat (SSR) is a class of tandemly repeated sequences that consists of 1 - 6 nucleotide repeat units. The SSRs are important in plant typing and widely developed as molecular genetic markers for species identification (Wu *et al*. 2018). The cp SSRs are also used to assess the molecular phylogeny within the family *Cucurbitaceae* and *Compositae* (Chung *et al*. 2003, Wills *et al*. 2005).

The aim of this study was to sequence whole cp genomes of *A. lappa* and *A. tomentosum* and characterize their structure and the differences between species. We then aimed to find the evolutionary relationships in the *Arctium* genus to identify appropriate regions or genes for use as markers.

## Materials and methods

**Plant materials:** Fresh *Arctium lappa* leaves were collected from Dalian, China (E 121° 52′, N 39° 03′) and fresh *A. tomentosum* leaves were collected from Urumqi, China (E 84° 33′, N 44° 07′). Both were identified by Dr. T. Kang from the Liaoning University of Traditional Chinese Medicine in Shenyang, China. Voucher specimens were deposited in the Liaoning University of Traditional Chinese Medicine Herbarium (*A. lappa* 20170417005LY, *A. tomentosum* 20170616001LY).

**Chloroplast DNA extraction and sequencing:** Approximately 5 g of fresh leaves was harvested for chloroplast DNA isolation using a method of Mcpherson *et al.* (2013). Then, 1 μg of purified DNA was fragmented and used to construct short-insert libraries (insert size 430 bp) according to the manufacturer's instructions (*Illumina,* USA), then sequenced on an *Illumina Hiseq 4000* (Borgstrom *et al.* 2011).

**Genomic assembly:** Prior to assembly, raw reads were filtered in order to remove the reads with adaptors, the reads showing a quality score below 20 ($Q < 20$), the reads containing a percentage of uncalled bases ("N" characters) equal or greater than 10 %, and the duplicated sequences. The cp genome was reconstructed using a combination of *de novo* and reference-guided assemblies, and the following three steps were used (Cronn *et al.* 2008, Koren *et al*. 2012): *1*) the filtered reads were assembled into contigs using *SOAP denovo 2.04* (Luo *et al.* 2012); *2*) contigs were aligned to the reference genome of two species using *BLAST*, and aligned contigs (≥ 80 % similarity and query coverage) were ordered according to the reference genome; and *3*) clean reads were mapped to the assembled draft cp genome to correct the wrong bases, and the majority of gaps were filled through local assembly.

**Genomic anotation and analysis:** The cp genes were annotated using an online *DOGMA* tool (Wyman *et al.* 2012) using default parameters to predict genes encoding proteins, tRNA, and rRNA. Whole chloroplast genomes *BLAST* search (Altschul *et al.* 1990) (E-value ≤ $1e^{-5}$, a minimal alignment length > 40 %) was performed against five databases. They were *KEGG* (Kyoto Encyclopedia of Genes and Genomes; Kanehisa *et al.* 1990, 1997, 2006), *COG* (Clusters of Orthologous Groups; Tatusov *et al.* 1997, 2003), *NR* (Non-Redundant Protein Database), *Swiss-Prot* (Magrane 2011), and *GO* (Gene Ontology; Ashburner *et al.* 2000). The SSR software *Micro Satellite* (*MISA*; http://pgrc.ipk-gatersleben.de/misa/) was used to identify the SSR sequences, and tandem repeats of 1 - 6 nucleotides were considered as microsatellites. The minimum numbers of repeats were set to 11, 6, 5, 5, 5, and 5 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides, respectively. The maximal number of bases interrupting 2 SSRs in a compound microsatellite was 100. We focused on perfect repeat sequences. The *mVISTA* was used to analyze similarities between four *Asteraceae* species (Frazer *et al.* 2004). Web-based *REPuter* (http://bibiserv.techfak.uni-bielefeld.de/reputer/) was used to analyze the long repeat sequences, which included forward, reverse, and tandem repeats with a minimum sequence length of 30 bp and edit distances of 3 bp.

**Chloroplast genome mapping:** The chloroplast genomes of *A. lappa* and *A. tomentosum* were exported in the GenBank format using the *Sepquin* software, and the cp genome was mapped using the annotation results (Wang *et al.* 2015) (http://ogdraw.mpimp-golm.mpg.de/index.shtml). Finally, the complete cp genomes were submitted to the *NCBI* GenBank database (accession number: MH375874).

**Phylogenetic analysis:** To identify the phylogenetic position of *Arctium* species within *Asteraceae* and their relationship to other families, phylogenetic trees were constructed by the cp genome sequences from 36 species, 34 of them were downloaded from GenBank. Among them, three species, *Lathyrus clymenum*, *Arabidopsis thaliana*, and *Nicotiana tabacum*, were set as outgroups. Thirty one *Asteraceae* species and two species from *Araliaceae* and *Campanulaceae* were also included in the phylogeny. The analysis was run using whole chloroplast genome single nucleotide polymorphisms (SNPs) and 25 genes were shared by all 36 species (Table 1 Suppl.). Maximum-likelihood (ML) methods were performed for the phylogenetic analyses using *PhyML 3.0*, and a model *GTR+I+G* was selected for ML analyses with 100 bootstrap replicates to calculate bootstrap values.

## Results and discussion

The total length of the chloroplast (cp) genomes of *Arctium lappa* and *A. tomentosum* were 152 767 bp and 152 688 bp (Fig. 1*A*,*B*), respectively. This is similar to the cp genomes of other genera in *Asteraceae*, which were downloaded from *GenBank* indicating that the *Asteraceae*

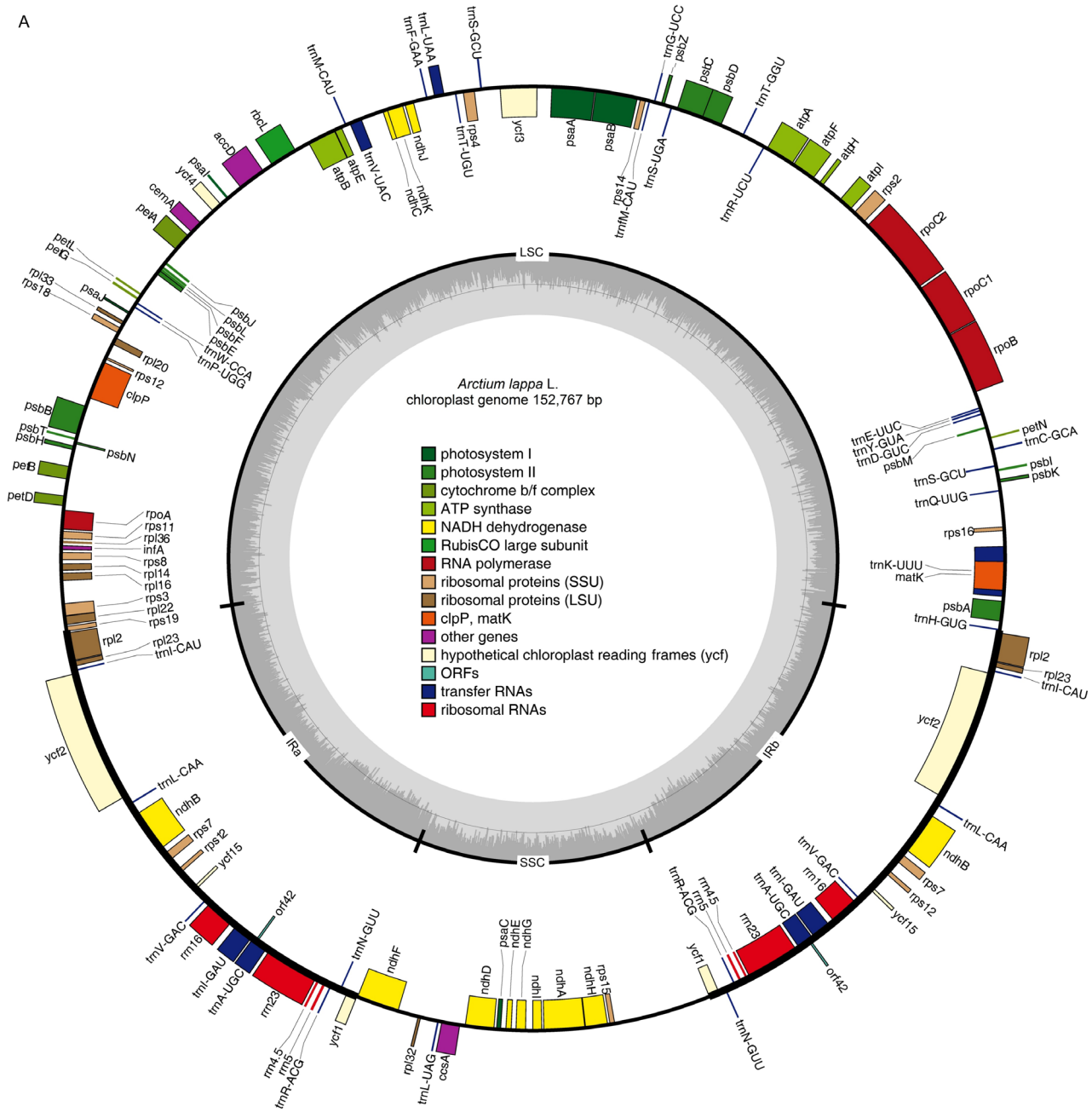


Fig. 1A. A chloroplast genome map illustrates *Arctium lappa.* The outer ring shows the gene, non-coding RNA and the positions of other genomic components corresponding to the gene name; the inner ring shows the genomic GC content.

cp genomes are very conservative (Li *et al.* 2013). The LSC lengths of *A. lappa* and *A. tomentosum* were 83 821 bp and 83 744 bp, the SSC lengths were 18 584 bp and 18 582 bp, the IR (IRa, IRb) region of them shared the same length of 25 181 bp, and the GC content were 38.02 and 37.69 %, respectively (Table 1). There were 134 and 126 genes annotated for *A. lappa* and *A. tomentosum*, respectively.

*Arctium lappa* contained 36 tRNA genes, and 9 rRNA and 90 protein coding genes (Table 2 Suppl.). Fourteen tRNA genes and all rRNAs were located in the IR region (Fig. 1*A*). *Arctium tomentosum* contained 89 protein coding genes, 29 tRNA genes, and 8 rRNA genes in the IR region, and 10 tRNAs in the IR region (Fig. 1*B*). Based on comparisons of published chloroplast genome sequences of *Panax* (NC_006290.1) (Kim *et al.* 2004), two inversions were present in the chloroplast genomes of *A. lappa* and *A. tomentosum.* One boarder of the larger inversion was located between the *trn*S-GCU and *trn*C-GCA genes, and the other boarder was located between the *trn*T-GCT and *trn*R-TCT genes. The smaller inversion had one boarder located between the *trn*C-GCA and *rpo*B

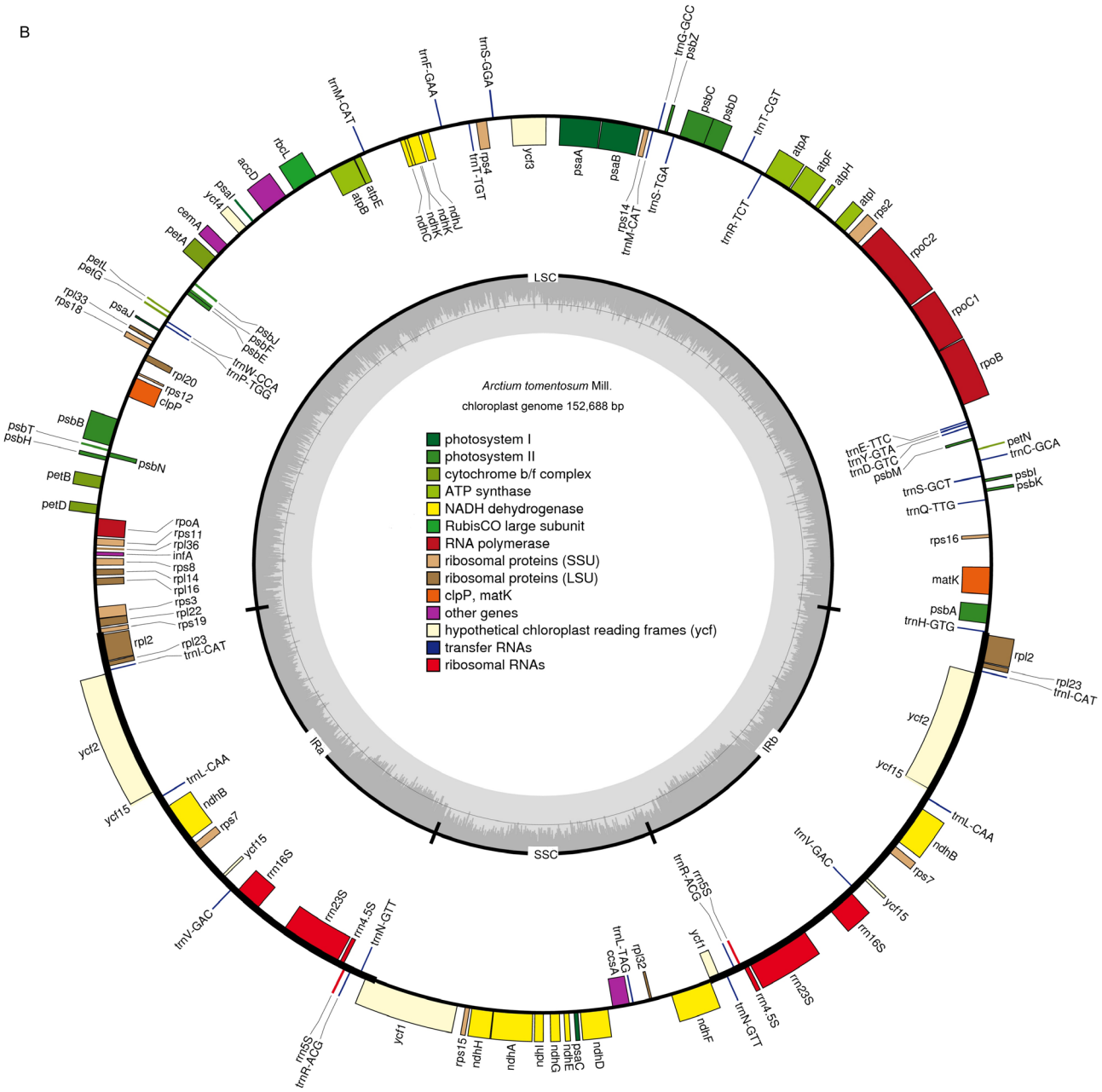


Fig. 1B. A chloroplast genome map illustrates *Arctium tomentosum.* The outer ring shows the gene, non-coding RNA and the positions of other genomic components corresponding to the gene name; the inner ring shows the genomic GC content.

genes and the other boarder between the *psb*M and *trn*D-GUC (Fig. 1 Suppl.).

The annotated genes from *A. lappa* and *A. tomentosum* (Table 2 Suppl.) were generally similar, but some differences were observed. Specifically, the *mat*K gene in *A. lappa* was located in the intron of the *trn*K-UUU gene, whereas *A. tomentosum* had no *trn*K-UUU gene. There was specific trans-splicing in *rps*12 of *A. tomentosum*, the 5' end of the exon was located in the LSC region and the 3′ end in the IR region, which is consistent with *Magnolia grandiflora* (Li *et al.* 2013) and ginseng (Kim *et al*. 2004). The *rps*12 was located in the IR region of *A. lappa.*, which had more transfer RNA genes annotated: *trn*G-UCC, *trn*I-GAU, *trn*K-UUU, *trn*L-UAA, *trn*T-GGU, *trn*V-UAC, *trn*A-UGC, and *trn*I-GAU with the first six located in the LSC region and the last two in the IR area. The genes *trn*S-GGA, *trn*T-CGT, and *trn*G-GCC from the *A. tomentosum* were not annotaed in the *A. lappa* cp genome. In the subunits of photosystem II, the gene *psb*L was absent in the *A. tomentosum.* In the conserved hypothetical chloroplast reading genes, the *ycf*15 gene was located in the IR region with two copies in each IR region of *A. tomentosum,* but

Table 1. Comparison of general features of the chloroplast genomes in four *Asteraceae* species. LSC - large single-copies; SSC - small single copies; IR - inverted repeat.

| Species | *Arctium lappa* | *A. tomentosum* | *Chrysanthemum* × *morifolium* | *C. indicum* |
|---|---|---|---|---|
| Gene length [bp] | 152767 | 152688 | 151003 | 151129 |
| GC content [%] | 38.02 | 37.69 | 37.48 | 37.42 |
| LSC length [bp] | 83821 | 83744 | 82782 | 82810 |
| SSC length [bp] | 18584 | 18582 | 18354 | 18377 |
| IR length [bp] | 25181 | 25181 | 24953 | 24971 |
| Gene number | 134. | 126 | 129 | 129 |
| Gene number in IR regions | 42 | 36 | 17 | 17 |
| Protein-coding gene number | 90 | 89 | 85 | 85 |
| rRNA gene number | 8 | 8 | 8 | 8 |
| tRNA gene number | 36 | 29 | 36 | 36 |

it had one copy in each IR region of *A. lappa*. *A. lappa* had the putative genes *orf*42-D2 and *orf*42, which were not annotated for *A. tomentosum*. The genes *mat*K, *trn*K-UUU, *trn*S-GGA, *trn*T-CGT, *trn*G-GCC, and *psb*L could be used to develop molecular makers for the species.

Introns play an important role in the regulation of gene expressions. Some recent studies have found that many introns can improve the expression and timing of exogenous genes at specific locations, which may lead to the appearance of expected agronomic traits in genetically modified plants; therefore, introns can be useful tools for improving desired agronomic traits (Jiao *et al.* 2012). *A. lappa* chloroplast DNA had 10 coding genes with introns. Among them, the *clp*P gene contained three introns and *ycf*3 contained two introns. *A. tomentosum* chloroplast DNA had 11 genes containing introns. Among them, *ycf*1 contained three introns, and *ycf*3 contained two introns (Table 3 Suppl.). These introns may help to improve plant resilience and developing new cultivars.

Simple sequence repeats are effective molecular markers with a vast number of applications (Ravi *et al*. 2008, Kang *et al.* 2017). The SSRs are rich in quantity, co-dominant, and highly repetitive and also have a simple genomic structure, which is relatively conserved making them widely used for species identification and genetic analysis of individuals and groups (Khakhlova and Bock 2006, Curci 2015).

In the cp genomes of *A. lappa* and *A. tomentosum*, we found 16 and 24 SSRs, respectively (Table 2). The total lengths of SSRs in plastomes of *A. lappa* and *A. tomentosum* were 118 and 274 bp, respectively. There were only two types of SSRs (mononucleotides and dinucleotides). The dinucleotide SSRs of *A. lappa* were AT/AT and TA/TA repeats, and there were eight mononucleotide SSRs located in the LSC region. There was one mononucleotide SSR located in the SSC region and two located in the IR region. *Arctium tomentosum* had 19 SSRs located in the LSC region, 2 of which were dinucleotide SSRs, and 4 mononucleotide SSRs located in the IR region, and one mononucleotide SSR sequence located in the SSC region (Table 4 Suppl.) Both SSRs were A/T rich consistently with the *A. lappa* and *A. tomentosum* chloroplast genomes. We compared the abundance of the SSRs with two related species of *C.* (Fig. 2). When analyzed the SSR sequences of the four species of *Asteraceae,* all had more mononucleotide sequences and fewer trinucleotide sequences. Only *C. indicum* had a small amount of trinucleotide SSRs, and the SSRs of several plant species of *Asteraceae* were similar, which may provide information for finding SSR markers.

Large repeat sequences showed repeats with length ≥ 30 bp each. Forty-six and 36 pairs of large repeat sequences with a sequence identity > 90 % were found in the *A. lappa* (Table 5 Suppl.) and *A. tomentosum* cp genomes (Table 6 Suppl.). The repeats from *A. lappa* ranged from 30 to 116 bp in length and from *A. tomentosum* ranged from 30 to 46 bp in length. A total of 16 and 19 large repeat sequences were located in protein-coding genes in *A. lappa* and *A. tomentosum*, respectively. Thirty and 17 large repeat sequences were located in the intergenic regions

Table 2. Type and abundance of different single sequence repeats (SSR) in *Arctium lappa* and *A. tomentosum*.

| | *A. lappa* | | | *A. tomentosum* | | |
|---|---|---|---|---|---|---|
| | SSR repeats | SSR abundances | Percent abundance [%] | SSR repeats | SSR abundances | Percent abundance [%] |
| Mononucleotide | A/T | 12 | 75 | A/T | 21 | 87.5 |
| | G/C | 2 | 12.5 | G/C | 1 | 4.17 |
| Dinucleotide | AT/AT | 1 | 6.25 | AT/AT | 2 | 8.33 |
| | TA/TA | 1 | 6.25 | - | - | - |

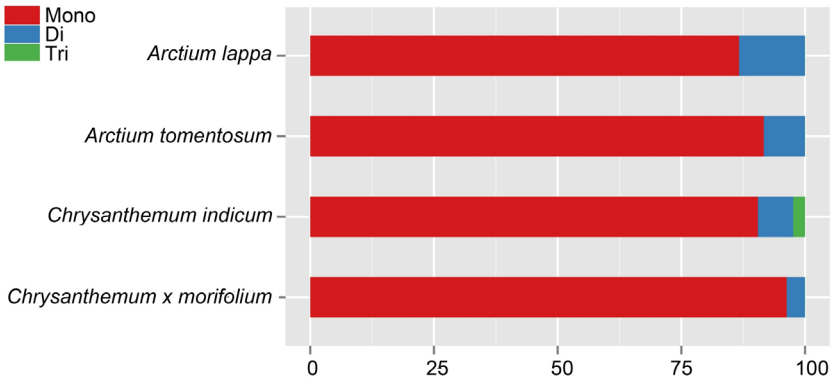


Fig. 2. Single sequence repeats in the chloroplast genomes of two species of *Arctium* and two species of *Chrysanthemum.* Mono represents mononucleotide repeats, Di represents dinucleotide repeats, and Tri represents trinucleotide repeats.*Arctium* spp. and *Chrysanthemum* spp.

in *A. lappa* and *A. tomentosum*, respectively. Numerous repeat sequences have been identified in the intergenic spacer regions in cp genomes of several angiosperm lineages (Yang *et al*. 2013).

Changes in genomic size are mainly due to differences in the length of LSC and IR regions (Chumley *et al.* 2006, Wu and Ge 2012). The length of the IR region in some species of *Magnoliaceae* have a positive relationship with the length of the complete cp genome sequence (Li *et al.* 2013). By comparing the boundary characteristics of *A. lappa, A. tomentosum, C.* × *morifolium*, and *C. indicum*, we found that the length of the IR region was conservative among the four species, ranging from 24 953 to 25 181 bp, but there was some minor variation present in the IR region among the species (Fig. 3). In angiosperms, the downstream sequences of IRb/SSC are mostly conserved, and the *ndh*F gene is adjacent to it (Raubeson *et al.* 2007), but in *A. lappa*, the *ndh*F gene was located downstream of the IRa/LSC border, which was different from the other species. The border between IRa and LSC regions in the four species (*A. lappa, A. tomentosum, C.* × *morifolium*, and *C. indicum*) was located inside the *rps*19 gene, which was different from *Lonicera japonica* (He *et al.* 2017), which border between IRa and LSC regions is located inside *rpl*23. The *rps*19 genes of the four species were 279 bp long, and their extension areas to IRa were similar in size (59 - 61 bp). The IRa region extends into the *ycf*1 gene, and the *ycf*1 and *ndh*F genes overlap in *A. lappa.* In *C.* × *morifolium* and *C. indicum*, *ycf*1 was a cross-border gene with a portion in both the IRa region and the SSC region. Because IRa and IRb had the same gene, they both contained a part of the *ycf*1 gene, however, the other part of *ycf*1 was present only in one end of the LSC region, which was close to the IRa/LSC border, but not the other, which was close to the LSC/IRb border. Therefore, the part of *ycf*1 in the IRb area could not constitute the complete gene and produce vacancies of 884 bp (*C. indicum*) and 886 bp (*C.* × *morifolium*) within the IRb. The *trn*H genes were located in the LSC region and the distance from the IRa/LSC border were different, the smallest was 0 bp in *C. indicum*. The IR region may be highly conserved within the genus, with a high homology within the family that we

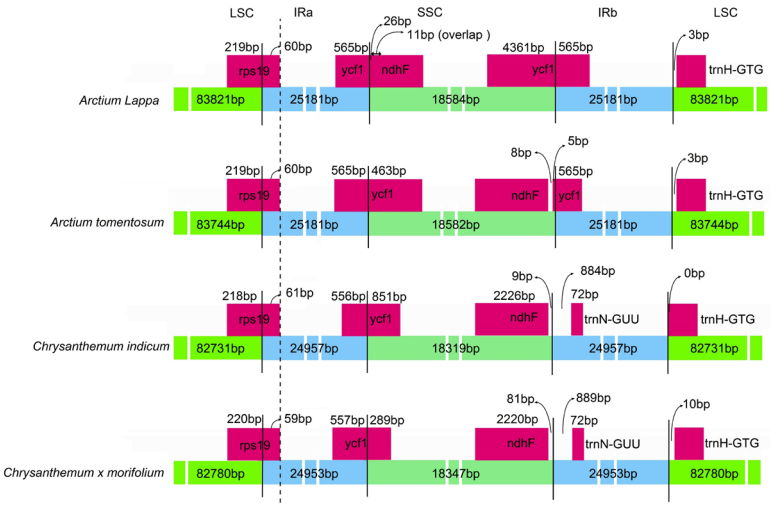


Fig. 3. Schematic representations of large single copies (LSC), small single copies (SSC), and inverted repeat (IRs) border regions in *Arctium* spp. and *Chrysanthemum* spp.

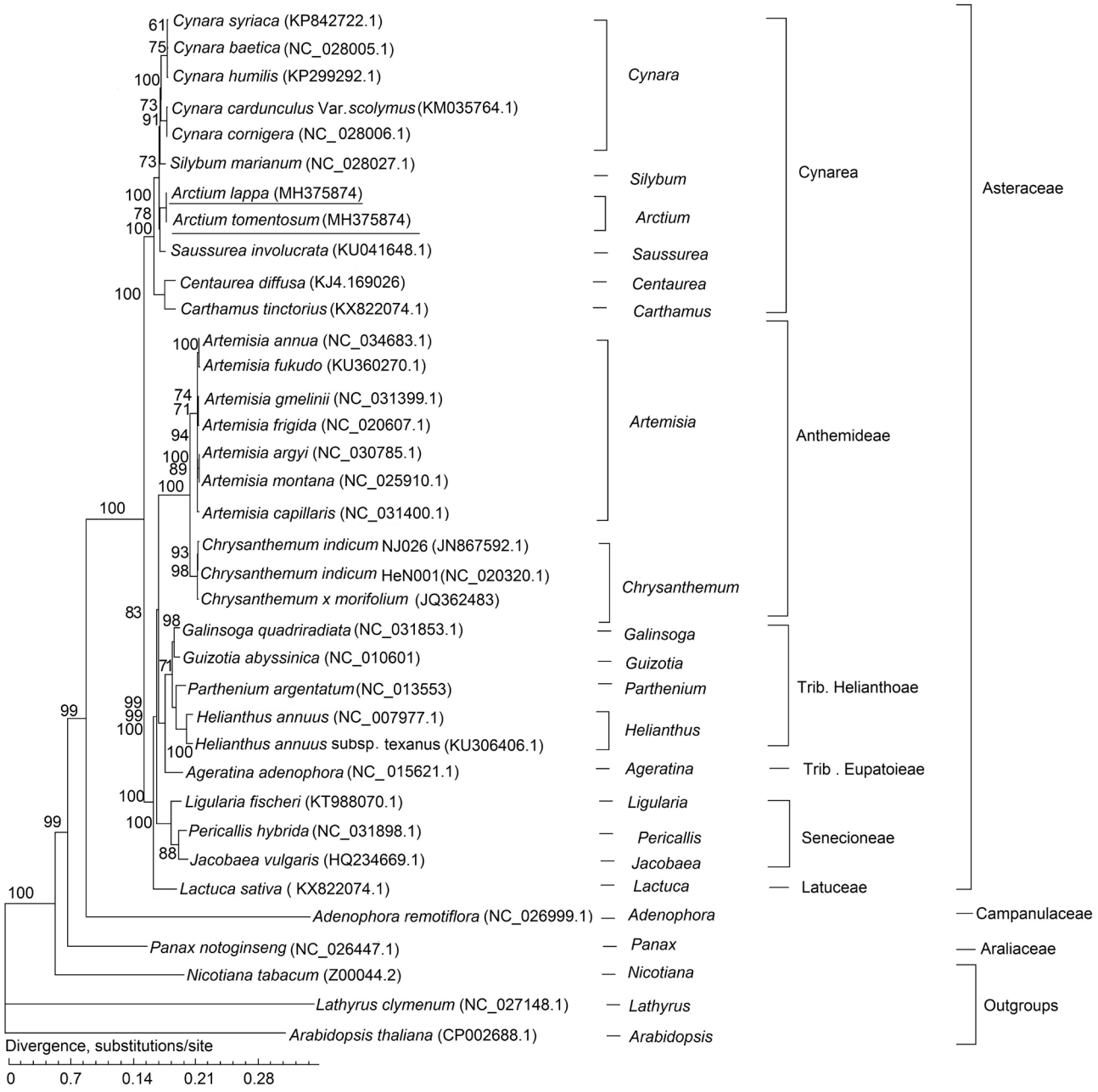


Fig. 4. A molecular phylogenetic tree of 36 species of whole genome single nucleotide polymorphisms (SNPs). The tree was constructed using a maximum likelihood analysis using PhyML 3.0. The model GTR+I+G was selected for maximum-likelihood analyses with 100 bootstrap replicates to calculate the bootstrap values.

study, and the IR region was a relatively conserved region due to the ability to correct replication after mutation (Yao *et al.* 2015).

The overall sequences of the chloroplast genomes of the four *Asteraceae* species, *A. lappa*, *A. tomentosum, C. indicum*, and *C.× morifolium,* were presented using *mVISTA* with *Silybum marianum* (NC_028027.1) as a reference (Fig. 2 Suppl.). The cp genomes in *Arctium* and *Chrysanthemum* were similar, but some differences between the two genera were obvious. In *Arctium*, the divergent regions were between the genes *atp*H and *atp*F and the genes *psb*Z and *rps*14, which were non-coding regions. These different regions could be developed as molecular makers for the identification and phylogenetic analysis.

Because cp genes are conserved within species but vary greatly across species, they are often used for phylogenetic analyses. A phylogenetic tree constructed from genome-wide population SNPs (Fig. 4) shows that there was a total of 33 nodes with 13 nodes 100 % supported and

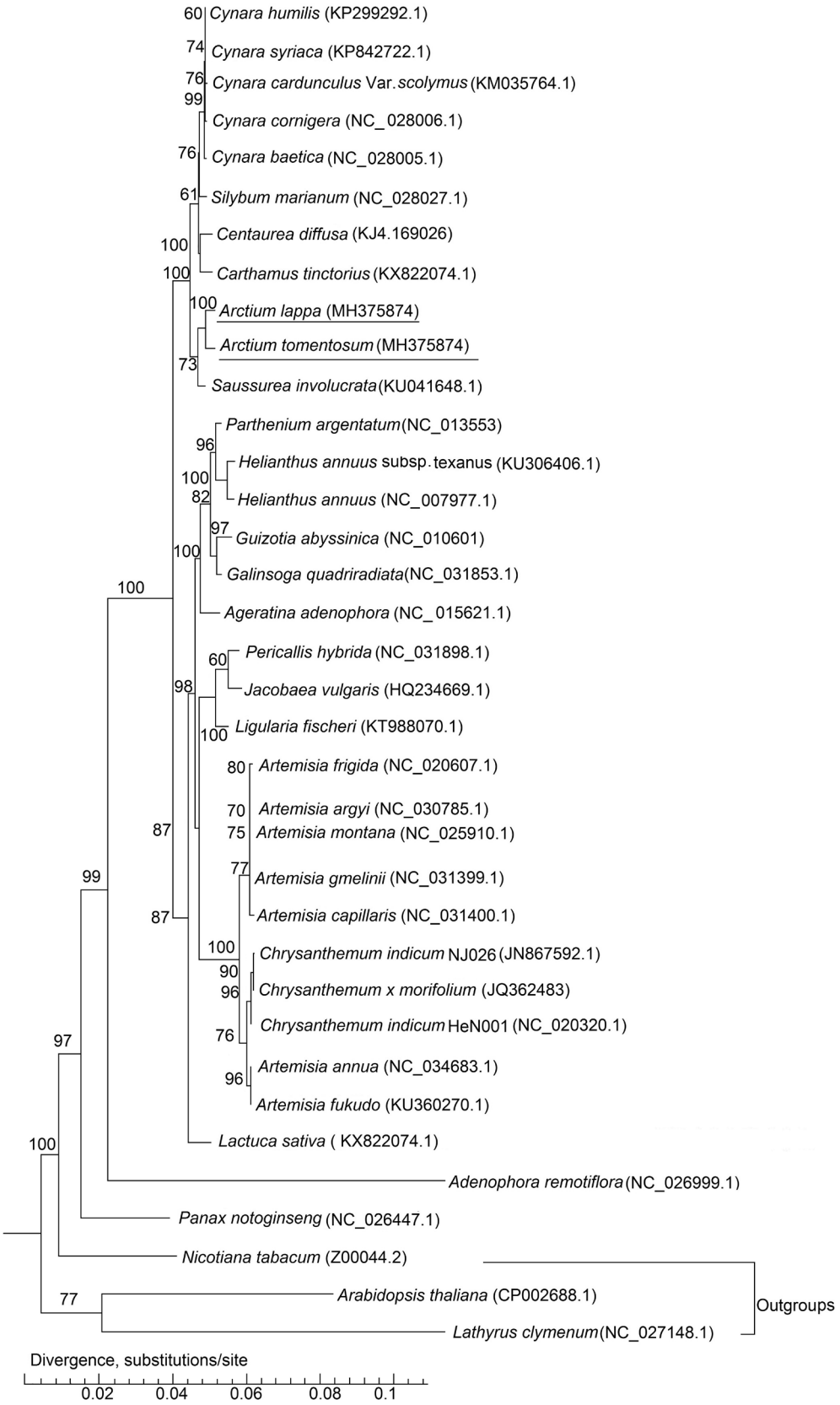


Fig. 5. A molecular phylogenetic tree of 36 species based on 25 shared protein-coding genes in the chloroplast genome. The tree was constructed using a maximum likelihood analysis using PhyML 3.0. The model GTR+I+G was selected for ML analyses with 100 bootstrap replicates to calculate the bootstrap values.

8 nodes ≥ 90 % supported. Four genera of Cynareae, *Arctium, Saussurea, Silybum*, and *Cynara*, all clustered into one clade. These four genera are closely related to *Certaurea* and *Carthamus.* In addition, *Arctium* is most closely related to *Saussurea.* Other *Asteraceae* species were distributed in *Lactuceae*, *Chrysantheminae*, trib. *Heliantheae*, and tribs. *Eupatorieae* and *Senecioneae*. Among them, the trib. *Heliantheae* and trib. *Eupatorieae* were close. The other two families, *Campanulaceae* and *Araliaceae*, gathered in one clade. All the 31 species in the *Asteraceae* family clustered in one clade, which supports a monophyly of the *Asteraceae*and is similar to previous studies (Wang *et al.* 2015). Using the SNPs from the cp genome to analyze the phylogenetic relationships among a selection of *Asteraceae* spp. shows the relative conservatism of *Asteraceae* cp genes.

Another phylogenetic tree (Fig. 5) was constructed from a total of 25 shared protein-coding sequences for the analysis of 36 species, requiring protein similarity threshold values > 40 %. The number of tree nodes was 34, and there were 18 nodes with ≥ 90 % support and 9 nodes with 100 % support. In this tree, *Arctium* was a sister to *Saussurea* and *Silybum* as well as five *Cynara* species. Seven *Artemisia* species composed one clade, which was a sister to *Chrysanthemum*. This is different from the tree based upon SNPs. The position analysis of *A. lappa* and *A. tomentosum* by two ways are consistent within *Asteraceae* and with the result according to Chinese traditional morphological classification.

## Conclusions

We obtained the complete cp genome sequences from *A. lappa* and *A. tomentosum*, and the chloroplasts had similar genome sizes. The cp genomes from the *Arctium* spp. were compared with two *Chrysanthemum* spp., and the genome size was also found to be similar. There were some differences in the type and location of genes annotated in *A. lappa* and *A. tomentosum*, which may be suitable for developing markers and conducting phylogenetic analysis. In addition, there were only mononuleotide and dinucleotide SSRs, and those had a similar abundance in the two *Chrysanthemum* spp. In *A. lappa* and *A. tomentosum*, 46 and 36 pairs of large repeat sequences were found. These repeat motifs can be used to develop markers and analyze a phylogenetic tree. *A. lappa* and *A. tomentosum* had the same IR lengths and similar with the *Chrysanthemum* spp. Based on the SNPs from the chloroplast genome and shared coding-proteins, a phylogenetic analysis was conducted to illustrate the position of *A. lappa* and *A. tomentosum* in *Asteraceae*.

## References


Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. - J. mol. Biol. **215**: 403-410, 1990.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. - Nat. Genet. **25**: 25-29, 2000.

Borgstrom, E., Lundin, S., Lundeberg, J.: Large scale library generation for high throughput sequencing. - PLoS ONE **6**: e19119, 2011.

Chan, Y.S., Cheng, L.N., Wu, J.H., Chan, E., Kwan, Y.W., Lee, S.M., Leung, G.P., Yu, P.H., Chan, S.W.: A review of the pharmacological effects of *A. lappa* (burdock). - Inflammopharmacology **19**: 245-254, 2011.

Choi, K.S., Park, S.: The complete chloroplast genome sequence of *Aster spathulifolius* (*Asteraceae*); genomic features and relationship with *Asteraceae*. - Gene **572**: 214-221, 2015.

Chumley, T.W., Palmer, J.D., Mower, J.P., Fourcade, H.M., Calie, P.J., Boore, J.L., Jansen, R.K.: The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. - Mol. Biol. Evol. **23**: 2175-2190, 2006.

Chung, S.M., Deena S.D.W., Jack, E.S.: Genetic relationships within the *Cucurbitaceae* as assessed by consensus chloroplast simple sequence repeats (ccSSR) marker and sequence analyses. - Can. J. Bot. **81**: 814-832, 2003.

Conea, S., Mogoan, C., Vostinaru, O., Toma, C.C., Hepcal, I.C., Cazacu, I., Pop, C., Vlase., L.: Polyphenolic profile, anti-inflammatory and antinociceptive activity of an extract from *A. lappa* L. roots. - Not. Bot. Hort. Agrobot. **45**: 59-64, 2017.

Cronn, R., Liston, A., Parks, M., Gernandt, D.S., Shen, R.K.: Multiplex sequencing of plant chloroplast genomes using *Solexa* sequencing-by- synthesis technology. - Nucl. Acids Res. **36**: e122, 2008.

Curci, P.L., De, P.D., Danzi, D., Vendramin, G.G., Sonnante, G.: Complete chloroplast genome of the multifunctional crop globe artichoke and comparison with other *Asteraceae*. - PLoS ONE **10**: e0120589, 2015.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., Dubchak, A.I.: VISTA: computational tools for comparative genomics. - Nucl. Acids Res. **32**: W273-W279, 2004.

He, L., Qian, J., Li, X.W., Sun, Z.Y., Xu, X.L., Chen, S.L.: Complete chloroplast genome of medicinal plant *Lonicera japonica*: genome rearrangement, intron gain and loss, and implications for phylogenetic studies. - Molecules **22**: 249, 2017.

Jansen, R.K., Raubeson, L.A., Boore, J.L., De Pamphilis, C.W., Chumley, T.W., Haberle, R.C., Wyman, S.K., Alverson, A.J., Peery, R., Herman, S.J., Fourcade, H.M., Kuehl, J.V., McNeal, J.R., Leebens M.J., Cui, L.: Methods for obtaining and analyzing whole chloroplast genome sequences. - Methods Enzymol. **395**: 348-384, 2005.

Jiao, Y., Jia, H.M., Li, X.W., Chai, M.L., Jia, H.J., Chen, Z., Wang, G.Y., Chai, C.Y., Van de Weg, E., Gao, Z.S.: Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (*Myrica rubra*). - BMC Genomics **13**: 201, 2012.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M.: The KEGG resource for deciphering the genome. - Nucl. Acids Res. **32**: D277-D280, 2004.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG. - Nucl. Acids Res. **34**: D354-D357, 2006.

Kanehisa, M.: A database for post-genome analysis. - Trends Genet. **13**: 375, 1997.

Kang, J.S., Lee, B.Y., Kik, M.: The complete chloroplast genome

sequences of *Lychnwas wilfordii* and *Silene capitata* and comparative analyses with other *Caryophyllaceae* genomes. - PLoS ONE **12**: e0172924, 2017.

Khakhlova, O., Bock, R.: Elimination of deleterious mutations in plastid genomes by gene conversion. - Plant J. **46**: 85-94, 2006.

Kim, K.J., Lee, H.L.: Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. - DNA Res. **11**: 247-261, 2004.

Kim, K.J., Choi, K.S., Jansen, R.K.: Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (*Asteraceae*). - Mol. Biol. Evol. **22**: 1783-1792, 2005.

Kim, W.J., Moon, B.C., Yang, S., Han, K.S., Choi, G., Lee, A.Y.: Rapid authentication of the herbal medicine plant species *Aralia continentalis* Kitag., and *Angelica bwaserrata* C.Q. Yuan and R.H. Shan using ITS2 sequences and multiple x-SCAR markers. - Molecules **21**: 270, 2016.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W. R., Jarvis, E.D., Phillippy, A.M.: Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. - Nat. Biotechnol. **30**: 693-700, 2012.

Li, X.W., Gao, H.H., Wang, Y.T., Song, J.Y., Robert, H., Wu, H.Z., Hu, Z.G., Yao, H., Luo, H.M., Luo, K., Pan, H.L., Chen, S.L.: Complete chloroplast genome sequence of *Magnolia grandiflora* and comparative analysis with related species. - Sci. China Life Sci. **56**: 189-198, 2013.

Liu, L.P., Wan, Q., Guo, Y.P., Yang, J., Rao, G.Y.: Phylogeny of the genus *Chrysanthemum* L.: evidence from single-copy nuclear gene and chloroplast DNA sequences. - PLoS ONE **7**: e48970, 2012.

Lohse, M., Drechsel O., Kahlau, S., Bock, R.: Organellar Genome DRAW - a suite of tools for generating physical maps of plastid and mitochondrial genomes visualizing expression data sets. - Nucl. Acids Res. **41**: W575-W581, 2013.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung D.W., Yiu, S., Peng, S., Zhu, X., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam T.W., Wang, J.: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. - Gigascience **1**: 18, 2012.

Masalih, R.H., Majesky, L., Schwarzacher, T., Gornall, R., Heslop-Harrison, P.: Complete chloroplast genomes from apomictic *Taraxacum* (Asteraceae): Identity and variation between three microspecies. - PLoS ONE **12**: e0168008, 2017.

Magrane, M., Consortium, U.: UniProt knowledge base: a hub of integrated protein data database. - Databases (Oxford) **2011**: bar009, 2011.

McPherson, H., Van der Merwe, M., Delaney, S.K., Edwards, M.A., Henry, H.R., McIntosh, E., Robert, J.H., McIntosh, E., Rymer, P.D., Milner, M.L., Siow, J., Rossetto, M.: Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. - BMC Ecol. **13**: 8, 2013.

Ohyama, K., Fukuzai, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki,Y., Takeuchi, M., Chang, Z., Aota, S.I., Inokuchi, H., Ozeki, H.: Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. - Nature **322**: 572-574, 1986.

Palmer, J.D., Thompson, W.F.: Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence was lost. - Cell **29**: 537-550, 1982.

Raubeson, L.A., Jansen, R.K.: Chloroplast genomes of plants, plant diversity and evolution: genotypic and phenotypic variation in higher plants. - Mol. Phylogenet. Evol. 3: 45-68, 2005.

Raubeson, L.A., Peery, R., Chumley, T.W., Dziubek, C., Fourcade, H.M., Boore, J.L., Jansen, R.K.: Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. - BMC Genomics 8: 174, 2007.

Ravi, V., Khurana, J.P., Tyagi, A.K., Khurana, P.: An update on chloroplast genomes. - Plant Syst. Evol. **271**: 101-122, 2008.

Tatusov, R.L., Koonin, E.V., Lipman, D. J.: A genomic perspective on protein families. - Science **278**: 631-637, 1997.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Bachoti,S., Smirnov,S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I.,Yin, J.J., Natale, D.A., Yin, J.J., Natale, D.A.: The COG database: an updated version includes eukaryotes. - BMC Bioinformatics **4**: 41, 2003.

Tseng, M.C., Wong, S.L., Hsiung, D.S., Hwang, J.H., Lee, S.C., Chen, F.H.: Genetic diversity of the chloroplast *trn*L-*trn*F intergenic spacer and nuclear internal transcribed spacer of great burdock (*Arctium Lappa* L.) in Taiwan. - J. med. Plants Res. **6**: 5086-5093, 2012.

Wang, M., Cui, L., Feng, K., Deng, P., Du, X., Wan, F., Song, W., Nie, X.: Comparative analysis of *Asteraceae* chloroplast genomes: structural organization, RNA editing and evolution. - Plant mol. Biol. Rep. **33**: 1526-1538, 2015.

Wicke, S., Schneeweiss, G.M., De Pamphilis, C.W., Muller, K.F., Quandt, D.: The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. - Plant mol. Biol. **76**: 273-297, 2017.

Wills, D.M., Hester, M.L., Liu, A., Burke, J. M.: Chloroplast SSR polymorphisms in the *Compositae* and the mode of organellar inheritance in *Helianthus annuus*. - Theor. appl. Genet. **110**: 941-947, 2005.

Wu, M.L., Li, Q., Xu, J., Li, X.W.: Complete chloroplast genome of the medicinal plant *Amomum compactum*: gene organization, comparative analysis, and phylogenetic relationships within *Zingiberales*. - Chin. Med. **13**: 10, 2018.

Wu, Z.Q., Ge, S.: The phylogeny of the BEP clade in grasses revisited: evidence from the whole genome sequences of chloroplasts. - Mol. Phylogenet. Evol. **62**: 573-578, 2012.

Wyman, S.K., Jansen, R.K., Boore, J.L.: Automatic annotation of organellar genomes with DOGMA. - Bioinformatics **20**: 3252-3255, 2004.

Xu, L., Dou, D.Q., Wan, B., Yang, Y.Y., Kang, T.G., Liu, Y.: [Identification of traditional Medicine Fructus Arctii by nuclear ribosomal DNA ITS sequences.] - China J. Chin. Mater. Med. **36**: 45, 2011. [In Chin.]

Yang, J.B., Tang, M., Li, H.T., Zhang, Z.R., Li, D.Z.: Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. - BMC Evol. Biol. **13**: 84, 2013.

Yang, J.B., Yang, S.X., Li, H.T., Yang, J., Li, D.Z.: Comparative chloroplast genomes of *Camellia* species. - PLoS ONE **8**: e73053, 2013.

Yao, X., Tang, P., Li, Z., Li, D., Liu, Y., Huang, H.: The first complete chloroplast genome sequences in *Actinidiaceae*: genome structure and comparative analysis. - PLoS ONE **10**: e0129347, 2015.