Plant Breed. Biotech. 2020 (December) 8(4):439~444
https://doi.org/10.9787/PBB.2020.8.4.439
Online ISSN: 2287-9366
Print ISSN: 2287-9358

RAPID COMMUNICATION

Complete Chloroplast Genome of a Milk Thistle (*Silybum marianum*) Acc. '912036'

Jeehyoung Shim[1,2], Jae-Hyuk Han[3], Na-Hyun Shin[3], Jae-Eun Lee[4], Jung-Sook Sung[5], Yeisoo Yu[6], Sanghyun Lee[2], Kwang Hoon Ahn[1]*, Joong Hyoun Chin[3]*

[1]EL&I Co., Ltd., Gimje 54318, Korea
[2]Department of Plant Science and Technology, Chung-Ang University, Anseong 17546, Korea
[3]Department of Integrative Biological Sciences and Industry, Sejong University, Seoul 05006, Korea
[4]National Agrobiodiversity Center, National Institute of Agricultural Sciences, Jeonju 54874, Korea
[5]Department of Southern Area Crop Science, National Institute of Crop Science, Miryang 50424, Korea
[6]DNACare Co., Ltd., Seoul 06730, Korea

ABSTRACT Milk thistle (*Silybum marianum* Gaertn.) is a well-known medicinal plant which has been used for more than 2,000 years around the world. It produces silymarin, which cures the liver from hepatitis and toxin damages. In this study, a selfed and purified breeding line of the milk thistle from the Korean environment was used as a source of chloroplast genome construction. It showed high concentration of silybin B (3.50 mg/g) in its dried seeds. The complete chloroplast genome of *S. marianum* acc. '912036' is 152,556 bp in length and G+C content is 37.69%. A total of 87 protein coding genes with 104 exons were annotated. Chloroplast genomes of five accessions from different countries were compared with that of '912036', and no sequence polymorphism among them was identified. Thus, the chloroplast genome from this study can be used to develop *S. marianum*-specific DNA markers when compared with other diverse *S. marianum* accessions and Asteraceae species.

Keywords Chloroplast, Genome, NGS, Milk thistle, *Silybum marianum*

INTRODUCTION

Silybum marianum Gaertn., commonly known as milk thistle, is one of the high-value plant resources to provide silymarin which has been well known for its medicinal effect in liver health (Polyak *et al.* 2010; Bhattacharya 2011; Toyang and Verpoorte 2013). Its seeds have been used as a medicine for more than 2,000 years (Corchete 2008). Silymarin has been used for alcoholic liver treatment and for acute viral hepatitis (Abenavoli *et al*. 2010). Moreover, milk thistle oil is highly beneficial with unsaturated fatty acids (Ghavami and Ramin 2008). The morphology of a milk thistle plant is similar to *Cirsium* spp. because both are included in the *Asteraceae* (or *Compositae*) family.

Although milk thistle can be distinguished from *Cirsium* spp. by the unique white patterns on their leaves, they are mistakenly used and managed mainly due to their similar common names and flower shape. On the chemical aspect, silybin B is known as a determinant chemical composition from the other species (Rodriguez *et al*. 2018). However, chemical analysis is an expensive and destructive method. For this reason, species-specific genomic comparison should be conducted. There is one reference sequence of milk thistle chloroplast genome that is publicly available online (NC_028027 derived from a plant of SMAR20150709);

Received September 10, 2020; Revised October 9, 2020; Accepted October 10, 2020; Published December 1, 2020
***Corresponding author** Joong Hyoun Chin, jhchin@sejong.ac.kr, Tel: +82-2-3408-3897, Fax: +82-2-3408-3897
***Corresponding author** Kwang Hoon Ahn, kevinan1122@gmail.com, Tel: +82-63-544-1237, Fax: +82-504-180-1159

however, there is no publication using multiple accessions of milk thistle. Thus, in this study, we have constructed one naturalized accession in Korea, showing the high concentration of silybin B. Then, chloroplast genomes of five accessions collected from different countries were mapped against our construct and compared.

MATERIALS AND METHODS

Plant materials and DNA preparation

A total of six accessions of milk thistle have been used in this study. Four of them were from National Agrobiodiversity Center, National Institute of Agricultural Sciences of Rural Development of Administration (K001033 from Canada, K044886 from Germany, K153821 from North Korea, K227004 from Moldova) and the other two were bought from a local market (unknown genetic sources, '912036' and '912171' from EL&I, Co., Ltd.) in Gyeonggi-do, Korea. The collected seeds have been grown and observed in pots to develop homogeneous plants. The selfed seeds of the six milk thistles were separately sown in May. DNA from a single plant of each accession was extracted by the Cetyltrimethylammonium Bromide (CTAB) method (Murray and Thomson 1980). Each DNA was quantified by NanoDrop 2000 (Thermo Fisher Scientific, USA) and only the high-quality DNA samples were used for genome sequencing.

Sequencing and chloroplast genome construction

Illumina paired-end (PE) library with a 400 bp insert size was constructed according to the manufacture's recommendation, and the library was sequenced on Illumina Novaseq with 2 × 150 bp. The low quality sequences (Phred score ≤ 20) and Illumina adapter sequences were removed in raw fastq files using Trimmomatic v.0.39 (http://www.usadellab.org/cms/?page=trimmomatic) and the chloroplast sequences were collected by mapping the trimmed fastq files to the chloroplast sequence of milk thistle (Genbank acc# KT267161) using BWA (v0.7.17). *De novo* assembly using the selected chloroplast reads was conducted using Newbler v.2.9 (https://www.roche.com/) and the assembly was cross-validated with that of SPAdes (http://cab.spbu.ru/software/spades/). Gene prediction and manual editing were conducted using DOGMA (https://dogma.ccbb.utexas.edu/) and Artemis v.17.0.1 (https://www.sanger.ac.uk/science/tools/artemis), and the final chloroplast genome was visualized using OGDraw v.1.3.1 (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html). Phylogenetic analysis of chloroplast genomes among *S. marianum* and 10 relative plants was performed using MAFFT v.7.407 (https://mafft.cbrc.jp/alignment/software/) and MEGA v.10.0.5 (https://www.megasoftware.net/; Kumar *et al*. 2018), and the tree was generated using Neighbor-joining method with 1000 bootstrapping.

Whole genome sequences from the other five additional accessions were aligned to the '912036' chloroplast sequence using BWA-mem (v.0.7.17-r1188) and variants were called using a genome analysis toolkit (GATK v.3.8). Variants were filtered using vcftools (v.0.1.15) with the following conditions: minimum read coverage < 5; genotype quality < 20; genotype missing > 20%.

RESULTS AND DISCUSSION

Based on our preliminary chemical analysis and agronomic traits of the six plants which were used for sequencing, '912036' was selected for the chloroplast genome construction. '912036' produced the highest level of silybin B (around 3.50 mg/g) from the dried seeds and showed the most typical shape of flower sets with vigorous thorns.

After trimming, 127.7 million reads covering 18.9 Gb

Table 1. Pre-processing statistics of the sequencing products of the chloroplast of a *Silybum marianum* accession, '912036'.

	Reads	Length		Q30 (%)	Q20 (%)	GC (%)
Raw Data	149,012,860	22,500,941,860	-	88.61	95.32	36.23
Trimmed Data	127,679,160	18,932,927,244	84.14%	92.36	97.83	35.66
CP Data	10,036,686	1,495,223,450	7.90%	92.56	97.92	37.72

were retained from a total of 149 million raw reads (about 22.5 Gb). About 7.9% of total reads (~10 million reads) were identified as chloroplast reads in chloroplast mapping and they were used for assembly (Table 1). Chloroplast genome sequence was assembled *de novo* with Newbler and SPAdes assembler followed by manual correction and gap-filling.

Table 2. The complete chloroplast genome structure of a *Silybum marianum* accession, '912036'.

Structure	Length	GC (%)	Start	End
LSC	83,535	35.81	1	83535
IR	25,195	43.1	83536	108730
SSC	18,631	31.45	108731	127361
IR	25,195	43.1	152556	127362
Total	152,556	37.69		

The complete chloroplast genome of *S. marianum* is 152,556 bp in length and G+C content is 37.69%. It showed a typical quadripartite structure, including a pair of IR (25,195 bp) separated by the large single copy (LSC; 83,535 bp) and small single copy (SSC; 18,631 bp) regions (Table 2 and Fig. 1). GC content of IR regions was 43.1%, which is higher than those of LSC and SSC regions, which was commonly reported previously (Shen *et al*. 2017).

A total of 87 protein coding genes with 104 exons were annotated (Fig. 1 and Table 3). The average size of the protein coding sequences is 854 bp, whose G+C content is 38.51%. Besides, 37 tRNAs and eight rRNAs were annotated in the chloroplast DNA. Most photosynthesis related genes were located within the LSC region.

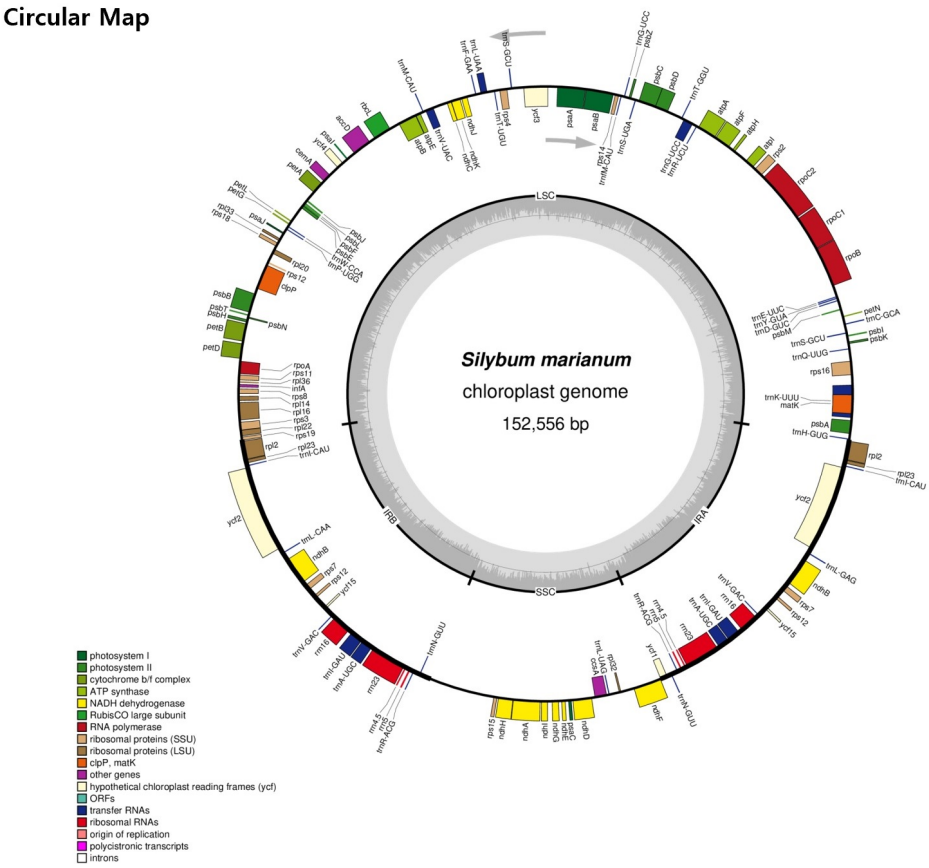

Fig. 1. Circular map of chloroplast genome of *Silybum marianum* acc. '912036'. The genes drawn outside and inside of the circle are transcribed in clockwise and counterclockwise directions, respectively. Genes were colored based on their functional groups. The inner circle shows the quadripartite structure of the chloroplast: small single copy (SSC), large single copy (LSC) and a pair of inverted repeats (IRa and IRb). The gray ring marks the GC content with the inner circle marking a 50% threshold. Asterisks mark genes that have introns.

The evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei 1987). The optimal tree with the sum of branch length = 0.07854853 is shown (Fig. 2). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein 1985). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura *et al*. 2004) and are in the units of the number of base substitutions per site. This analysis involved 11 nucleotide sequences. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were 158,398 positions in the final dataset. The complete chloroplast genome sequence is available at NCBI-SRA (accession no. MW00167).

Table 3. Annotation result of *Silybum marianum* chloroplast DNA.

Annotation Info	
Genome Size (bp)	152,556
G+C content (%)	37.69
Protein No	87
exons	104
Protein Coding (%) (excluding introns)	48.7
Average Size (bp)	854
Average exon Size (bp)	715.1
G+C content (%)	38.51
tRNAs	37
G+C content (%)	52.78
rRNA	8
G+C content (%)	55.21

The chloroplast genome assembled in this study was very close to *S. marianum* (KT267161.1 or SMAR20150709) (Fig. 2). The closest species were *Cirsium rhinoceros* and *Cirsium shansiense*. The chloroplast genome of *C. rhinoceros* was reported by Nam *et al*. (2019). It is Korean endemic species distributed in Jeju island, Republic of Korea, which has been utilized as traditional medicine, containing polyacetylene, three flavonoids, and noriso-

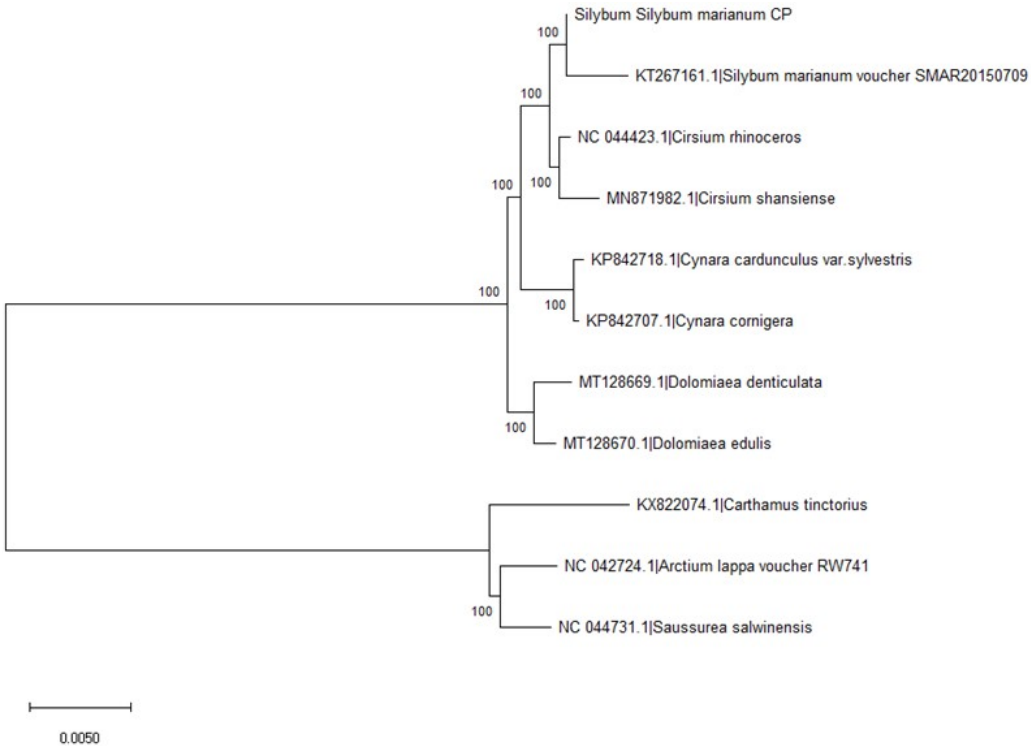

Fig. 2. Phylogenic tree of chloroplast genome sequences of '912036', one reference *Silybum marianum*, and nine related species. The evolutionary history was inferred using the Neighbor-Joining method. The evolutionary distances were computed using the Maximum Composite Likelihood method. This analysis involved 11 nucleotide sequences. Evolutionary analyses were conducted in MEGA X.

prenoids. The complete chloroplast genome of *C. shansiense*, commonly found in China, was recently reported by Xu *et al*. (2020). It is also consumed for medicinal purposes, which can be used for dealing with bleeding and hypertension (Ming *et al*. 2012). One of the related species, *Saussurea salwinensis* Anth., also can be found in China (https://plants.jstor.org/stable/10.5555/al.ap.specimen.e00394623). Interestingly, these are all included in the family Asteraceae (or *Compositae*), which is covering more than 32,000 species in plants.

The NGS sequences of the other five milk thistle accessions were mapped against the reference of '912036', but we could not find sequence polymorphism among them although they were from different European and Asian countries. Therefore, the chloroplast genome from this study can be used to develop *S. marianum*-specific DNA marker when it is compared with the other diverse *S. marianum* accessions and Asteraceae species, although there are too many species within the family. However, several SNP and InDels were identified from the comparison between Genbank accessions (KT267161 and NC_028027 derived from the same voucher plant of SMAR20150709) and '912036'. Therefore, it is possible that different *S. marianum* might have some sequence variations in the chloroplast genome.

Recently, the importance of the identification of the useful herbal and medicinal plants is globally increasing. Using the genome sequence information, the uniformed and certified seed production and the proper identification can be achieved. At this point, utilizing chromosomal DNA for species identification will be useful.

ACKNOWLEDGEMENTS

This work was supported by the Cooperative Research Program for Agriculture Science & Technology Development (Project No. PJ01418503) of the Rural Development Administration, Republic of Korea.

REFERENCES

Abenavoli L, Capasso R, Milic N, Capasso F. 2010. Milk thistle in liver diseases: past, present, future. Phytother. Res. 24: 1423-1432.

Bhattacharya S. 2011. Phytotherapeutic properties of milk thistle seeds: An overview. J. Adv. Pharm. Educ. Res. 1: 69-79.

Corchete P. 2008. *Silybum marianum* (L.) Gaertn: the source of silymarin, p. 123-148. In: KG. Ramawat, JM. Merillon (Eds.). Bioactive Molecules and Medicinal Plants. Springer, Berlin, Heidelberg, Germany.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783-791.

Ghavami N, Ramin AA. 2008. Grain yield and active substances of milk thistle as affected by soil salinity. Commun. Soil Sci. Plant Anal. 39: 2608-2618.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35: 1547-1549.

Murray MG, Thompson WF. 1980. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 8: 4321-4326.

Nam SJ, Kim JM, Kim Y, Ku JJ, Jung SY, Lee YM, *et al*. 2019. The complete chloroplast genome of Korean endemic species, *Cirsium rhinoceros* (H. Lév. & vaniot) Nakai (Asteraceae). Mitochondrial DNA B Resour. 4: 2351-2352.

Polyak SJ, Morishima C, Lohmann V, Pal S, Lee DYW, Liu Y, *et al*. 2010. Identification of hepatoprotective flavonolignans from silymarin. Proc. Natl. Acad. Sci. U.S.A. 107: 5995-5999.

Rodriquez JP, Quilantang NG, Lee JS, Lee JM, Kim HY, Shim JS, *et al*. 2018. Determination of silybin B in the different parts of *Silybum marianum* using HPLC-UV. Nat. Prod. Sci. 24: 82-87.

Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406-425.

Shen X, Wu M, Liao B, Liu Z, Bai R, Xiao S, *et al*. 2017. Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant *Artemisia annua*. Molecules 22: 1330.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining

method. Proc. Natl. Acad. Sci. U.S.A. 101: 11030-11035.

Ming T, Qian L, Yan H. 2012. Infrared and thermal analysis identification of *Cirsium shansiense* Petrak and *Potentilla discolor* Bunge. Med. Plant 10: 49-50.

Toyang NJ, Verpoorte R. 2013. A review of the medicinal potentials of plants of the genus *Vernonia* (Asteraceae). J. Ethnopharmacol. 146: 681-723.