



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

**A DISSERTATION FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY**

**Defining of plastid-mitochondrial
genome flux and establishment of plant
barcoding technology**

**엽록체-미토콘드리아 간 유전체 전이 규명 및 식물
바코딩 기술 개발**

By

HYUN-SEUNG PARK

FEBRUARY, 2019

**MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY
DEPARTMENT OF PLANT SCIENCE
COLLEGE OF AGRICULTURE AND LIFE SCIENCES
THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY**

Defining of plastid-mitochondrial genome flux and establishment of plant barcoding technology

UNDER THE DIRECTION OF DR. TAE-JIN YANG
SUBMITTED TO THE FACULTY OF THE GRADUATE
SCHOOL OF SEOUL NATIONAL UNIVERSITY

By

HYUN-SEUNG PARK

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY
DEPARTMENT OF PLANT SCIENCE

FEBRUARY, 2019

APPROVED AS QUALIFIED DISSERTATION OF HYUN-SEUNG
PARK FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

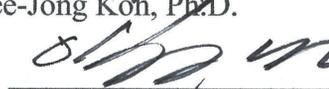
BY THE COMMITTEE MEMBERS

CHAIRMAN


Hee-Jong Koh, Ph.D.

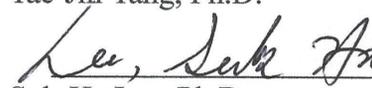


VICE-CHAIRMAN


Tae-Jin Yang, Ph.D.

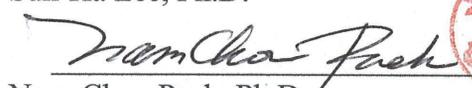


MEMBER


Suk-Ha Lee, Ph.D.

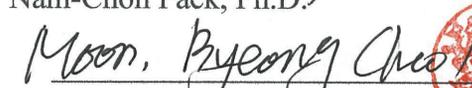


MEMBER


Nam-Chon Paek, Ph.D.



MEMBER


Byeong Cheol Moon, Ph.D.



Defining of plastid-mitochondrial genome flux and establishment of plant barcoding technology

HYUN-SEUNG PARK

**Department of Plant Science
The Graduate School of Seoul National University**

GENERAL ABSTRACT

DNA barcoding technology is used to classify and authenticate herbal plants and contributes to prevent economically motivated adulteration (EMA) of those products, which is vital for the stable growth of bio-industry. The herbal industry have spent 10 to 15 billion dollars per year since 2010, to manage this issue and provide the confident herbal materials. In this respect, DNA barcoding is regarded to be a promising one to discriminate target species and EMA counterfeit. However, underestimating the genomic complexity of the plant species can cause undiscovered shortcomings in DNA barcoding and molecular taxonomy. The herbal plant has a wide range of natural diversity as they have grown in wild without breeding approach. Complex nature of plant genome such as biparental inheritance or horizontal transfer of organelle genome is often ignored and underestimated. Without considering these distinct features of the plant genome, DNA barcoding can cause confusion or even disruption of entire industry like the cases happened in 2015 about the New York attorney general and GNC in the USA, and the adulteration of *Cynanchum wilfordii* with *C. auriculatum* in Korea.

In the first chapter, I assembled complete chloroplast and mitochondrial genome of two *Cynanchum* species. By comparative analysis, I found that 35% of the plastid genome of both species were transferred to their mitochondrial genome (mitochondrial plastid DNA, MTPT) from their common ancestor and the fragments were maintained in conserved form due to the slow mutation rate of plant mitochondrial genome. I identified diverse and lineage-specific MTPT transfer from 81 plants mitochondrial genomes and this flux contributed to the complexity of mitochondrial genome structure. Furthermore, I inspected possible DNA barcoding paradox from co-amplification of MTPT and developed recommended guidelines for regulation of economically motivated adulteration and protecting the herbal industry.

In the second chapter, I developed multiple plant barcoding primer pairs from 23 chloroplast genomes in four families including Araliaceae, Apiaceae, Papaveraceae, and Cannabaceae. The conserved regions of chloroplast genome across the family level were discovered by multiple alignments of their chloroplast genomes and the primer sequences that perfectly match to all of the 23 chloroplast genomes were picked to reduce PCR biases. From two universal barcoding region, *matK* and *rbcL* primer pairs with reduced amplicon size were also designed for further application to the NGS platform. PCR analysis was conducted to each pair from actual plant DNA and showed successful amplification results. *In silico* PCR experiment to registered chloroplast genomes from both of monocot and dicot plants in Genbank demonstrates the potential use of these pairs as universal plant barcoding.

Keywords: chloroplast, mitochondria, genome transfer, economically motivated adulteration, plant barcoding,

Student Number : 2010-21149

CONTENTS

GENERAL ABSTRACT	I
LIST OF TABLES	VI
LIST OF FIGURES	VII
LIST OF ABBREVIATIONS	IX
GENERAL INTRODUCTION	1
REFERENCES	4
CHAPTER 1	7
Horizontal plastid genome transfer creates mitochondrial genome complexity and DNA barcoding paradox	
ABSTRACT	8
INTRODUCTION	9
RESULTS	11
Plastid and mitochondrial genome sequences of <i>C. wilfordii</i> and <i>C. auriculatum</i>	11
MTPTs in the two <i>Cynanchum</i> species	14
Nucleotide substitution rates in plastid and mitochondrial genomes ..	17
Plastid genome flux into mitochondrial genome in angiosperms	18
DNA barcode markers based on inter- and intra-species plastid polymorphism	23
DISCUSSION	32
MATERIALS AND METHODS	39
Assembly of plastid and mitochondrial genomes	39
Identification of MTPTs in mitochondrial genome.....	39
Plastid marker design and PCR amplification	39

Ks value calculation for genes in plastid and mitochondrial genomes	.41
REFERENCES	42
CHAPTER II	46
Development of chloroplast universal primer for plant metabarcoding from unknown mixture	
ABSTRACT	47
INTRODUCTION	48
RESULTS	50
Selection of conserved regions in chloroplast genomes and primers design	50
Designing primers from conventional universal barcoding primer with reduced amplicon size	54
Phylogenetic relationship for evaluating sequence diversity	55
Overlapping of amplicon with mitochondrial plastid DNA (MTPT)	57
<i>In silico</i> PCR analysis for extended application of markers to overall plant lineage	57
PCR validation of primers and selection of final candidate	60
NGS application of PCR products from mixed DNA template	65
DISCUSSION	72
MATERIALS AND METHODS	75
Selection of barcoding candidate regions	75
Discovering conserved region in chloroplast genome	75
Sequence diversity of amplicon and phylogenetic analysis	75
Designing of primers from conventional barcoding region	76
Validation of primers by PCR analysis	76

NGS application of PCR product from mixed DNA	77
REFERENCES	79
ABSTRACT IN KOREAN	83

LIST OF TABLES

- Table 1-1.** Plastid genes commonly transferred to the mitochondrial genome in *Cynanchum* species
- Table 1-2.** Sequence similarity of *matK* genes between species and between organelle genomes of C
- Table 1-3.** Mode value of synonymous substitution (Ks) among four Apocynaceae species calculated from plastid and mitochondrial genes
- Table 1-4.** List of plant mitochondrial genomes used in this study
- Table 1-5.** Plastid DNA markers used in authentication of *Cynanchum* species
- Table 1-6.** The estimated proportion of NGS reads for plastid and mitochondrial genomes in Cw and Ca.
- Table 2-1.** List of chloroplast genomes used for primer design
- Table 2-2.** Information of designed primer pairs
- Table 2-3.** List of dicot order and representative species used for *in silico* PCR
- Table 2-4.** List of monocot order and representative species used for *in silico* PCR
- Table 2-5.** *Summary of in silico PCR success rate from order level of monocot and dicot and in vitro PCR result*
- Table 2-6.** List of plant DNA used for PCR Test
- Table 2-7.** List of samples and Primers used for 1st NGS application
- Table 2-8.** The 1st NGS result of two replication from each of markers
- Table 2-9.** List of samples and primer used for 2nd NGS application
- Table 2-10.** The 2nd NGS results of each amplicon from mixed DNA with 11 species.
- Table 2-11** Sequence of Primers used in this study

LIST OF FIGURES

- Figure 1-2.** Plastid–mitochondrial genome structure, flux, and evolution in *Cynanchum* species
- Figure 1-3.** Multiple sequence alignment of two plastid targets with their counterpart MTPT targets in *C. wilfordii* (Cw) and *C. auriculatum* (Ca).
- Figure 1-4.** Schematic representation of MTPTs in 81 plant mitochondrial genomes.
- Figure 1-5.** Genome-wide comparison of mitochondrial genomes of three Solanaceae species: pepper, tomato, and potato.
- Figure 1-6.** Circos plot showing MTPT in *Capsicum annuum* and *Solanum lycopersicum*.
- Figure 1-7.** DNA marker paradox derived from MTPTs
- Figure 1-8.** Estimated proportion for plastid and mitochondrial genomes using WGS of *C. wilfordii* (Cw) and *C. auriculatum* (Ca).
- Figure 1-9.** Nine additional plastid markers without MTPT counterparts for authentication of *C. wilfordii* (Cw) and *C. auriculatum* (Ca).
- Figure 1-10.** Plastid map and intra- and inter-species polymorphism of *C. wilfordii*.
- Figure 1-11.** Genotyping of *C. wilfordii* and *C. auriculatum* collections for three plastid-derived markers
- Figure 1-12.** Mitochondrial genome sizes and MTPT amounts in the 81 plants used in this study
- Figure 1-13.** Intraspecies plastid diversity and DNA barcoding paradox.
- Figure 2-1.** Location of primer pairs on chloroplast map
- Figure 2-2.** Schematic representation of primer designing strategy from universal barcoding regions

- Figure 2-3.** Phylogenetic relationship of amplicon from each primer pairs
- Figure 2-4.** PCR amplification of chloroplast specific primers (M1 – M18)
- Figure 2-5.** PCR amplification of chloroplast specific primers (M19 – M22)
- Figure 2-6.** PCR amplification of nine universal primers (M23 – M31)
- Figure 2-7.** Relative NGS read depth of four plant species from each amplicon (1st NGS analysis)
- Figure 2-8** Relative NGS read depth of four plant species from each amplicon of multiplex PCR
- Figure 2-9** Relative read depth of 11 plant species from each amplicon (2nd NGS analysis)
- Figure 2-10** Diagram of metabarcoding primers for constructing NGS library based on two step PCR

LIST OF ABBREVIATIONS

COI	Cytochrome oxidase I
matK	Maturase K
rbcL	Ribulose bisphosphate carboxylase large chain
ITS2	Internal transcribed spacer 2
EMA	Economically motivated adulteration
Cw	<i>Cynancum wilfordii</i>
Ca	<i>Cynancum auriculatum</i>
MTPT	Mitochondrial plastid DNA
NUPT	Nuclear plastid DNA
WGS	Whole-genome sequencing
InDel	Insertion and deletion
SNP	Single nucleotide polymorphism
MYA	Million years ago
MFDS	Ministry of Food and Drug Safety
IGS	Intergenic spacer
tRNA	Transfer RNA
rRNA	Ribosomal RNA
PCR	Polymerase chain reactions
HRM	High-resolution melting curve analysis
KASP	Kompetitive allele-specific PCR
PT	Plastid
MT	Mitochondria
TCM	Traditional Chinese medicine

GENERAL INTRODUCTION

DNA barcoding is a term of a technique using sequence information of short DNA fragment for authentication of species. This technique widely used from microbiome¹ to animal or plant for species authentication²⁻⁴, molecular taxonomy⁵, and even bio-diversity in ecology⁶. The mitochondrial cytochrome oxidase I (*COI*) gene is a universal target region for DNA barcoding in animal genomes⁷ and the plastidial maturase K (*matK*), ribulose biphosphate carboxylase large chain (*rbcL*), and internal transcribed spacer 2 of 45s ribosomal DNA (ITS2) are widely used barcoding region in Plants^{8,9}. These regions have high copy number and structural stability, which enabling them to exist relatively intact form even in the old specimen or processed commercial herbal products^{10,11}.

Chloroplast and mitochondria are cellular organelles with independent genome apart from nuclear one. Mitochondrial genome in animal and chloroplast genome in a plant are highly conserved in their length and circular structure¹². However, plant mitochondrial genome is extremely large with a diverse size range from several hundred to several thousand kilobase¹³ and existed in diverse multichromosomal structure mediated by homologous recombination^{14,15}. Horizontal transfer of foreign genome such as nuclear, plastid, bacteria, and virus contribute to this enlarged genome size of plant mitochondria¹³, as well as introns, extended intergenic sequences, and high repeat contents¹⁶.

Owing to technical improvement, genome information of many medicinal plants have been accumulated mainly targeting chloroplast genome. The high quality draft genome of *Panax ginseng*, the king of herbal medicine, was reported in 2018¹⁷. Chloroplast genomes of *Panax* and related species were also published and their barcoding markers were developed from intra- and interspecies level polymorphic sites^{3,18,19}.

However, still there are many plants with no genomic information and natural diversity of their genomes have not yet been analyzed.

DNA markers provide a precise and convenient tool to identify plant species in a mixed sample. Trace amounts of DNA can be detected even after industrial processing which has a benefit on detecting materials with economically motivated adulteration (EMA). Identification and further quantification of each species from a mixture of plants are theoretically applicable from the most of functional foods. Although DNA barcoding has several benefits over conventional morphology or chemical-based discrimination²⁰, wrong application of DNA markers could cause severe social problems. In February 2015, the New York state attorney general's office accused major herbal product retailers of selling of adulterated product based on the application of DNA barcode²¹ but the products were restored after evaluation of the company. In April 2015, the sales volume of the best-selling herbal product derived from *Cynancum wilfordii* (Cw) for the relief of menopausal disorder had fallen steeply due to contamination issues with its related species *C. auriculatum* (Ca). The company was claimed for EMA by a civil organization but the claim was acquitted from the Korean Supreme Court after inspecting all of the Cw products in the market. Both cases are based on the application of one or two chloroplast target DNA markers

In the first chapter, I obtained the complete chloroplast and mitochondrial genome sequence from Cw and Ca using NGS platform and identified large scale chloroplast-mitochondrial genome flux in both species that make real confusion for DNA barcoding. I inspected a total of 81 mitochondrial genome sequence and identified dynamic mitochondrial plastid DNA (MTPT) with lineage-unique patterns. Collectively, here I show that the MTPTs can make a DNA barcoding paradox of mal-authentication of the target species. I also suggest to a reliable guideline for barcode marker application protocol for the EMA issue.

In the second chapter, I designed new primer pairs for plant metabarcoding from multiple alignment of complete chloroplast genome of 23 species and conventional barcoding regions. Amplifying ability of primer pairs were evaluated by actual PCR to various real plant DNA and *in silico* primer specificity investigation with order level of monocot and dicot chloroplast genomes in Genbank. Finally, NGS application using four selected plant primers from chloroplast and two nuclear primers was conducted with mixed DNA samples for validation of those primers.

REFERENCES

- 1 Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences* **109**, 6241-6246 (2012).
- 2 Nguyen, V. B. *et al.* Authentication markers for five major *Panax* species developed via comparative analysis of complete chloroplast genome sequences. *Journal of agricultural and food chemistry* **65**, 6298-6306 (2017).
- 3 Nguyen, V. B. *et al.* Comprehensive comparative analysis of chloroplast genomes from seven *Panax* species and development of an authentication system based on species-unique SNP markers. *Journal of Ginseng Research* (2018).
- 4 Joh, H. J. *et al.* Authentication of golden-berry *P. ginseng* cultivar ‘Gumpoong’ from a landrace ‘Hwangsook’ based on pooling method using chloroplast-derived markers. *Plant Breed Biotech* **5**, 16-24 (2017).
- 5 Purty, R. & Chatterjee, S. DNA Barcoding: An effective technique in molecular taxonomy. *Austin J Biotechnol Bioeng* **3**, 1059 (2016).
- 6 Kress, W. J., García-Robledo, C., Uriarte, M. & Erickson, D. L. DNA barcodes for ecology, evolution, and conservation. *Trends in ecology & evolution* **30**, 25-35 (2015).
- 7 Hebert, P. D., Ratnasingham, S. & de Waard, J. R. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences* **270**, S96-S99 (2003).
- 8 Group, C. P. W. *et al.* A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* **106**, 12794-12797 (2009).
- 9 Gao, T. *et al.* Identification of medicinal plants in the family

- Fabaceae using a potential DNA barcode ITS2. *Journal of ethnopharmacology* **130**, 116-121 (2010).
- 10 Wallace, L. J. *et al.* DNA barcodes for everyday life: Routine authentication of Natural Health Products. *Food Research International* **49**, 446-452 (2012).
 - 11 Shokralla, S. *et al.* Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. *Molecular ecology resources* **14**, 892-901 (2014).
 - 12 Parveen, I., Gafner, S., Techen, N., Murch, S. J. & Khan, I. A. DNA barcoding for the identification of botanicals in herbal medicine and dietary supplements: strengths and limitations. *Planta medica* **82**, 1225-1235 (2016).
 - 13 Alverson, A. J., Rice, D. W., Dickinson, S., Barry, K. & Palmer, J. D. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. *The Plant Cell*, tpc. 111.087189 (2011).
 - 14 Lonsdale, D., Brears, T., Hodge, T., Melville, S. E. & Rottmann, W. The plant mitochondrial genome: homologous recombination as a mechanism for generating heterogeneity. *Phil. Trans. R. Soc. Lond. B* **319**, 149-163 (1988).
 - 15 Sloan, D. B. *et al.* Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS biology* **10**, e1001241 (2012).
 - 16 Mower, J. P., Sloan, D. B. & Alverson, A. J. in *Plant Genome Diversity Volume 1* 123-144 (Springer, 2012).
 - 17 Kim, N. H. *et al.* Genome and evolution of the shade-requiring medicinal herb *Panax ginseng*. *Plant biotechnology journal* (2018).
 - 18 Kim, K. *et al.* Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species.

- PloS one* **10**, e0117159 (2015).
- 19 Kim, K. *et al.* Evolution of the Araliaceae family inferred from complete chloroplast genomes and 45S nrDNAs of 10 *Panax*-related species. *Scientific reports* **7**, 4917 (2017).
- 20 Chen, S. *et al.* A renaissance in herbal medicine identification: from morphology to DNA. *Biotechnology advances* **32**, 1237-1244 (2014).
- 21 Schneiderman, E. *AG Schneiderman asks major retailers to halt sales of certain herbal supplements as DNA tests fail to detect plant materials listed on majority of products tested*, <https://ag.ny.gov/press-release/ag-schneiderman-asks-major-retailers-halt-sales-certain-herbal-supplements-dna-tests> (2015).

CHAPTER 1

**Horizontal plastid genome transfer creates
mitochondrial genome complexity and DNA
barcoding paradox**

ABSTRACT

DNA barcoding technology is used to classify and authenticate herbal plants and contributes to prevent economically motivated adulteration (EMA) of those products, which is vital for the stable growth of bio-industry. However, underestimating the genomic complexity mediated by horizontal plastid genome transfer can cause undiscovered shortcomings in DNA barcoding and molecular taxonomy. I assembled plastid and mitochondrial genomes of *Cynanchum wilfordii*, an herb used as a functional food in the treatment of menopausal disorders, and *C. auriculatum*, which is found in EMA of *C. wilfordii*. In both species, the mitochondrial genome contained sequences related to ~35% of the plastid genome, termed mitochondrial sequences of plastid origin (MTPTs). I identified dynamic and lineage-specific horizontal plastid–mitochondrial genome transfer of up to 75 kb that contributed to diversifying mitochondrial genome structure complexity across 81 plant species. Additionally, co-amplification of MTPTs caused a DNA barcoding paradox in which herbal products could be mis-authenticated or mis-positioned taxonomically. I also characterized intraspecies diversity that could cause confusion in authentication. My results demonstrate frequent and lineage-unique MTPT distribution and show that it is conserved due to slow mutation rates in plant mitochondrial genomes. Co-amplification of MTPTs and intraspecies diversity create a DNA barcoding paradox, in which the very tool used for authentication misidentifies the product. I suggest guidelines for DNA barcoding in EMA regulation.

INTRODUCTION

Economically motivated adulteration (EMA) of herbal products, or “herb fraud,” has become a prevalent and serious threat to herbal industries as the global market for herbal products has increased^{1,2}. Such fraud costs the industry approximately \$10–\$15 billion globally every year owing to the need to authenticate products and eliminate EMA targets, and also damages the market for natural products more broadly by eroding consumer trust. DNA barcoding based on plastid (or chloroplast) and 45S nuclear ribosomal DNA (45S rDNA) is a credible and appropriate means to trace plant species in mixed samples or after industrial processing. However, the misapplication of such DNA markers can be highly problematic for the industry. In February 2015, the New York State attorney general’s office, on the basis of DNA barcode data, accused major herbal product retailers of selling adulterated products (<https://ag.ny.gov/press-release/ag-schneiderman-asks-major-retailers-halt-sales-certain-herbal-supplements-dna-tests>). The products were restored to market after an investigation determined that they had been produced in compliance with the guidelines of the US Food and Drug Administration and that the barcode testing was flawed. In April 2015, a similar case occurred in Korea related to a functional food based on an extract of the medicinal plant *Cynanchum wilfordii* (Cw), which is used in treatments of menopausal disorders³. Cw extract has been registered and approved as an ingredient for health functional food by the Ministry of Food and Drug Safety and was a best-selling health functional food during 2012–2015 in Korea. However, sales of the most popular Cw product plummeted after accusations that it was contaminated with the related species *Cynanchum auriculatum* (Ca). The manufacturer was sued for product adulteration by the Korea Consumer Agency, but later acquitted by the Korean Supreme Court. In both cases, one or two plastid DNA markers played key roles in the controversy.

Although plastids are well conserved in most plants, their genomes can potentially confound barcoding analysis and cause species mis-authentication. Plastid genomes have substantial intraspecies diversity although they are usually maternally inherited in plant species. Moreover, some species have more diverse plastid genomes because of biparental inherited heteroplasmy⁴. Additional unexpected results can be derived from horizontal plastid genome transfer into the mitochondrial and nuclear genomes, giving rise to what are known as mitochondrial sequences of plastid origin (MTPTs) and nuclear genome sequences of plastid origin (NUPTs), respectively⁵⁻¹⁰. Here, I assembled the complete plastid and mitochondrial genomes from individuals of two *Cynanchum* species and demonstrated large-scale plastid–mitochondrial genome flux. Inspection of 81 mitochondrial genomes revealed lineage-unique MTPT patterns and MTPTs capable of causing mis-authentication. Finally, I developed a set of recommended guidelines to address EMA and improve plant DNA barcoding system through a genomics-based approach.

RESULTS

Plastid and mitochondrial genome sequences of *C. wilfordii* and *C. auriculatum*

I assembled the complete plastid and mitochondrial genomes^{11,12} of Cw and Ca using low-coverage whole-genome sequencing (WGS) (**Figure 1-1**). For CW, I obtained three types of circular forms of the mitochondrial genome. Types 1 and 2 shared 50% conserved downstream sequence, while type 3 had no sequence homology with the other two. Total genome lengths were 379,601 bp, 352,774 bp and 111,332 bp for types 1, 2 and 3, respectively. The Ca mitochondrial genome was assembled into one linear chromosome 652,279 bp in length, along with two minor types of circular forms. Both mitochondrial genomes showed large-scale collinearity with some structural rearrangement (**Figure 1-2a**). The collinear regions showed high sequence similarity, and the sequences of all of the mitochondrial genes were identical except for the copy number of the *atp9* gene, which was two and one in Cw and Ca, respectively.

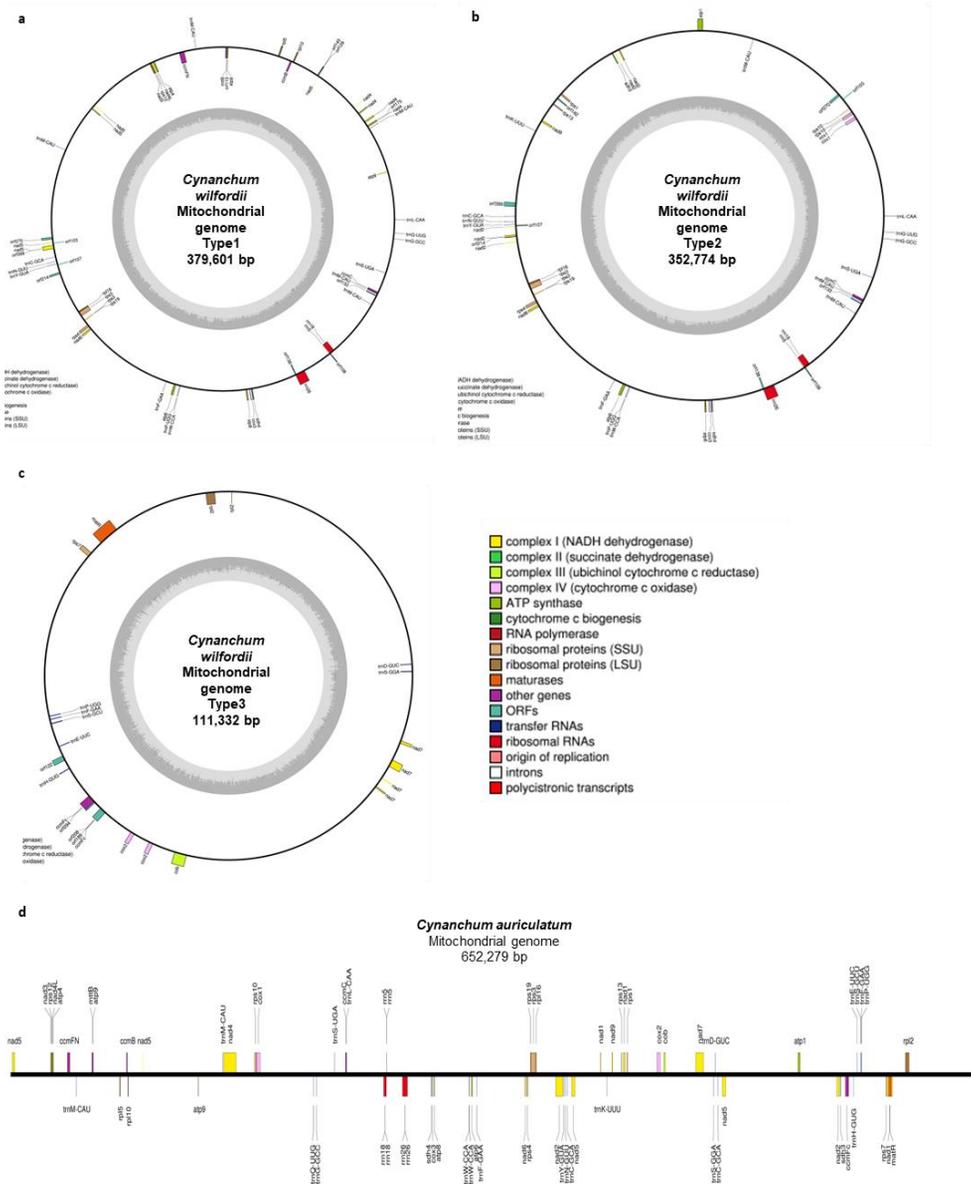


Figure 1-1. Mitochondrial genome maps of *C. wilfordii* and *C. auriculatum*. (a to c), Mitochondrial genome types 1, 2 and 3 for *C. wilfordii*. (d) Mitochondrial genome of *C. auriculatum*.

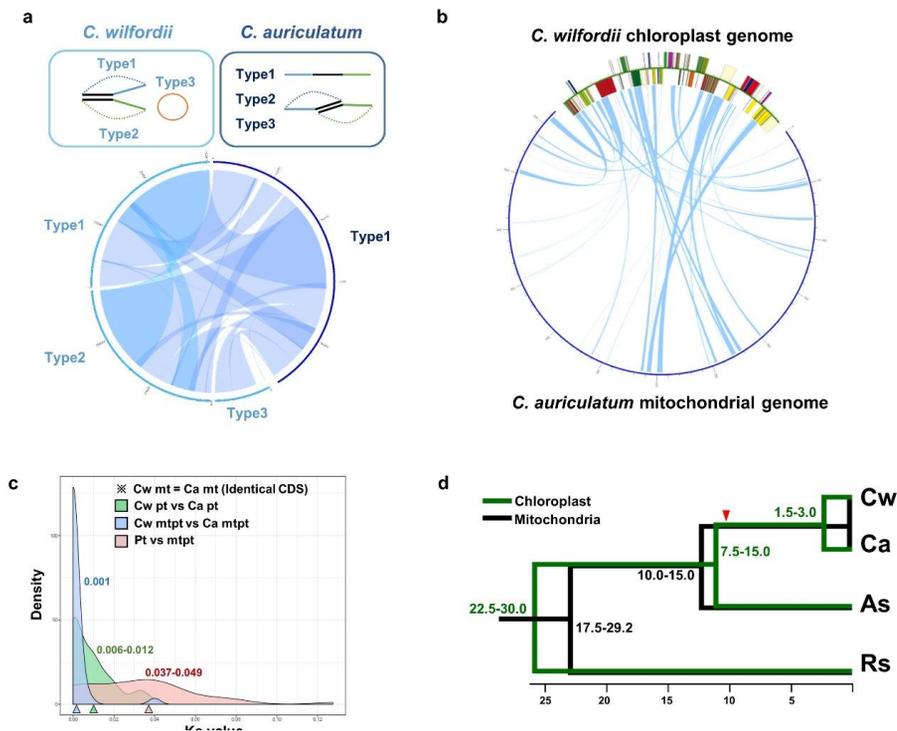


Figure 1-2. Plastid–mitochondrial genome structure, flux, and evolution in *Cynanchum* species. (a) Schematic representation of each type of mitochondrial genome and comparison in *C. wilfordii* (Cw) and *C. auriculatum* (Ca). Each line indicates a type of mitochondrial genome (of the three for each species), and genomic blocks with homology are connected at both the intra- and interspecies levels. The Circos plot shows synteny between the mitochondrial genomes of Cw and Ca. (b) Circos plot between plastid and mitochondrial genomes of the two *Cynanchum* species. (c) Density plot of the rates of nucleotide substitution between homologous genes (Ks values) in the plastid and mitochondrial genomes of the two *Cynanchum* species. The green line shows Ks values between plastid genes of Cw and Ca, and the gray line those between the MTPTs of the two species, whereas the red line shows Ks values between the MTPT and its plastid counterpart for each species. The mode value of Ks is marked with a triangle in each case (**Table 1-3**). (d) Estimation of divergence time among four Apocynaceae species based on Ks values of plastid (green) and mitochondrial genes (black). The red triangle indicates the estimated time of MTPT insertion in the common ancestor of the two *Cynanchum* species. The synonymous substitution rates per year per base were 2×10^{-9} for

plastid and 0.6×10^{-9} for mitochondria. As, *Asclepias syriaca*; Rs, *Rhazya stricta*.

MTPTs in the two *Cynanchum* species

In the plastid genomes, 34.3% and 37.7% of Cw and Ca sequences, respectively, showed homology with mitochondrial genome sequences of the same species (**Figure 1-2b**). Almost 50% of plastid protein-coding genes—36 of 75 genes—were identified in MTPTs of both *Cynanchum* species (**Table 1-1**). Two universal DNA barcoding genes, *matK* and *rpoB*, from the plastid genome were also identified in MTPTs of both species: while *rpoB* was identified in the mitochondrial genomes in a truncated form, the complete genic region of *matK* was found in the mitochondrial genomes. Although the MTPT and its plastid counterparts showed relatively high diversity of insertions and deletions (InDels), the sequence similarity between homologous sequences in the two *Cynanchum* species was high, 94.5%. The *matK* genes in the plastids of the two species showed 99.2% sequence similarity, and those in the MTPTs showed 99.6% sequence similarity (**Figure 1-3a, Table 1-2**).

Table 1-1. Plastid genes commonly transferred to the mitochondrial genome in *Cynanchum* species.

Gene product (transferred genes/total genes)	Gene names
ATP synthase (2/6)	<i>atpF, atpH</i>
Other proteins (3/6)	<i>matK*</i> , <i>cemA</i> , <i>infA</i>
NADH oxidoreductase (9/11)	<i>ndhK, ndhE, ndhG, ndhI,</i> <i>ndhA, ndhJ, ndhD,</i>
Cytochrome b6/f (5/6)	<i>petA, petD, petG, petL,</i> <i>petN,</i>
Photosystem I (4/7)	<i>psaA, psaB, psaC, psaI</i>
Photosystem II (7/16)	<i>psbA, psbC, psbD</i>
Large subunit ribosomal proteins (4/9)	<i>rpl22, rpl2, rpl36</i>
RNAP (3/4)	<i>rpoA, rpoB*</i> , <i>rpoC2</i>
Small subunit ribosomal proteins (5/11)	<i>rps3, rps11, rps14, rps16</i>
Proteins of unknown function (3/3)	<i>ycf2, ycf4, ycf15</i>

* Universal plant barcoding target regions in plastid

Table 1-2. Sequence similarity of *matK* genes between species and between organelle genomes of *C. wilfordii* and *C. auriculatum*.

InDels/SNPs	Similarity (%)			
	Cw-Pt	Ca-Pt	Cw-MTPT	Ca-MTPT
Cw-Pt		99.2	94.5	94.5
Ca-Pt	1/12		94.4	94.4
Cw-MTPT	11/79	12/82		99.6
Ca-MTPT	11/82	12/83	2/9	

Pt, plastid genome; MTPT, mitochondrial DNA of plastid origin

Numbers of SNPs/InDels and nucleotide similarity (%) are shown below and above the self-comparison diagonal, respectively.

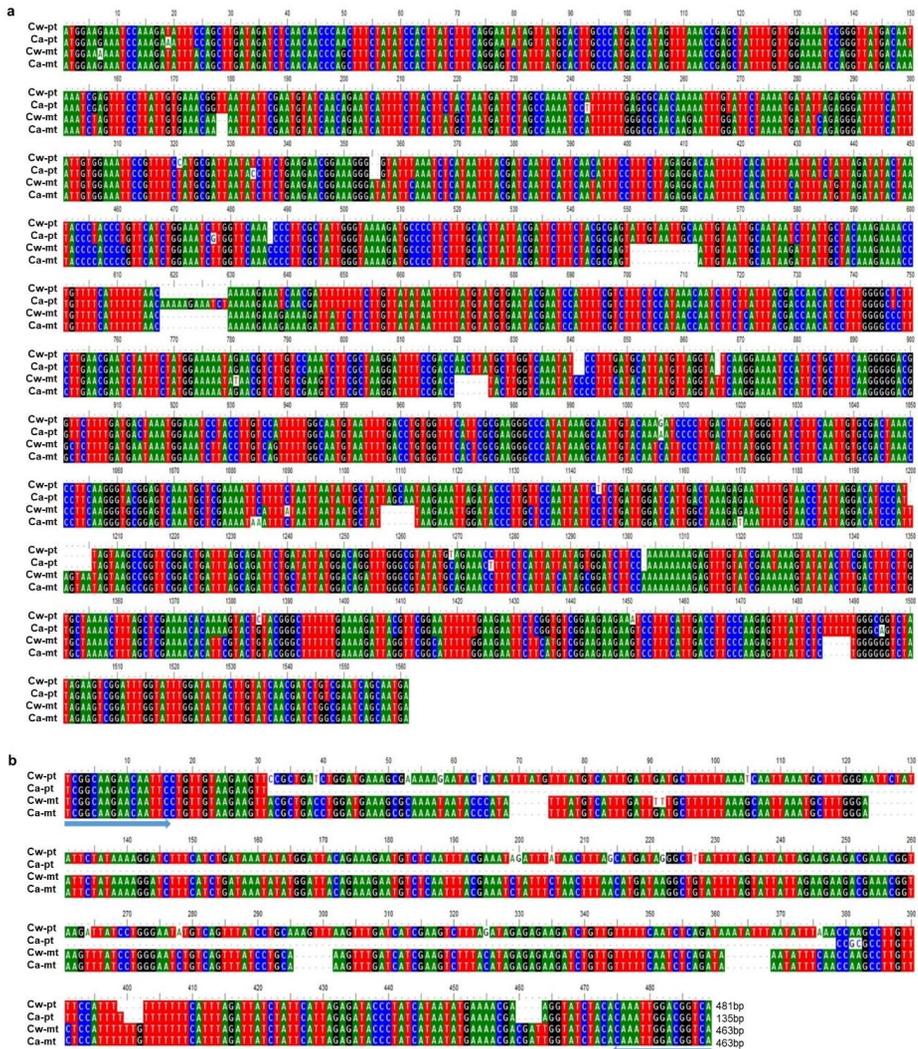


Figure 1-3. Multiple sequence alignment of two plastid targets with their counterpart MTPT targets in *C. wilfordii* (Cw) and *C. auriculatum* (Ca). (a) Multiple sequence alignment of the full sequence of *matK* gene and counterpart MTPTs in *C. wilfordii* and *C. auriculatum* (b) Multiple sequence alignment of the amplicon of Cw_i_1 representing plastid and counterpart MTPT targets in Cw and Ca. Blue arrows indicate primers in the intergenic region of *rps2* and *rpoC2*.

Nucleotide substitution rates in plastid and mitochondrial genomes

I next calculated the nucleotide substitution rates of homologous genes (synonymous substitutions per synonymous site, K_s) between species in the plastid, mitochondrion and MTPT regions (**Figure 1-2c, Table 1-3**). When I performed interspecies comparisons of each organelle between the two *Cynanchum* species, the mode of the K_s values of the 75 plastid genes was approximately 0.006–0.012 and that of the MTPT regions was around 0.001. The mitochondrial genes of both species were identical, and thus the K_s values were 0. However, when I performed intraspecies comparisons of plastid genes and their MTPT counterparts, the K_s values were much higher and broadly distributed (0.000–0.127) in both species.

I calculated the divergence times of the two *Cynanchum* species and members of two related genera, *Asclepias syriaca* and *Rhyza stricta*, based on the K_s values (**Figure 1-2d**) of their plastid and mitochondrial genes. The divergence rate calculated from the mitochondrial genome was much slower than that calculated from the plastid genome. *R. stricta* first diverged around 17.5–30.0 million years ago (MYA), and *A. syriaca* and the two *Cynanchum* species separated next, around 7.5–15.0 MYA. The two *Cynanchum* species were estimated to have diverged around 1.5–3.0 MYA based on the plastid genome sequence divergence, although their mitochondrial genes are identical. However, the MTPT sequence similarity between *A. syriaca* and *Cynanchum* species was much lower than that of the counterpart plastid genes. MTPTs showed 90% sequence similarity between *A. syriaca* and *Cynanchum* species and 99% sequence similarity between Cw and Ca. I found that a recent episode of elevated plastid–mitochondrial genome flux occurred at around 10.7 MYA in the common ancestor of Cw and Ca, based on K_s values between plastid and MTPT counterparts.

Table 1-3. Mode value of synonymous substitution (Ks) among four Apocynaceae species calculated from plastid and mitochondrial genes

	Pt	Cw	Ca	As	Rs
Mt					
Cw			0.006 - 0.012	0.03-0.06	0.09-0.12
Ca		0		0.03-0.06	0.09-0.12
As		0.012-0.018	0.012-0.018		0.09-0.12
Rs		0.021-0.035	0.021-0.035	0.021-0.035	

Cw, *C. wilfordii*; Ca, *C. auriculatum*; As, *Asclepias syriaca*; Rs, *Rhazya stricta*; Pt, Plastid genome; MTPT, Mitochondrial DNA of plastid origin

Mode value of Ks between mitochondrial genes and between plastid genes are shown below and above the self-comparison diagonal, respectively. Ks values between MTPT genes were calculated only between Cw and Ca with average value of 0.0001

Plastid genome flux into mitochondrial genome in angiosperms

I investigated MTPTs in the mitochondrial genome sequences of 81 flowering plants (**Table 1-4**). All 78 protein-coding genes in the plastid genome of *Arabidopsis thaliana* were identified at least once as MTPTs among the 81 plant mitochondrial genomes (**Figure 1-4**). *rbcL* was the most frequently detected MTPT, followed by *atpB*, *psaA*, *psaB*, *psbC*, *psbD*, *rpl2*, *rpl23*, *rpoB*, *rps7*, *rps12* and *ycf2*, which were each identified in more than 20 plant species. The international recommendations for barcoding candidate regions for land plants include seven plastid targets, four genic and three intergenic regions¹³. Among the four genic regions, I found that *rbcL* and *rpoB* belonged to the most frequent MTPT group; *rpoC1* was grouped in the moderately frequent group, found in more than 10 species; and *matK* was rarely detected as a MTPT. The three intergenic regions (*atpF-atpH*, *psbK-psbI* and *trnH-psbA*) and their flanking genes were seldom identified as MTPTs.

Table 1-4. List of plant mitochondrial genomes used in this study.

#	Name	NCBI accession	#	Name	NCBI accession
1	<i>Butomus umbellatus</i>	KC208619.1	46	<i>Cannabis sativa</i>	KU310670.1
2	<i>Spirodela polyrhiza</i>	NC_017840.1	47	<i>Glycine max</i>	JX463295.1
3	<i>Phoenix dactylifera</i>	JN375330.1	48	<i>Vigna radiata</i>	HM367685.1
4	<i>Cocos nucifera</i>	KX028885.1	49	<i>Milletia pinnata</i>	JN872550.1
5	<i>Allium cepa</i>	KU318712.1	50	<i>Lotus japonicus</i>	NC_016743.2
6	<i>Oryza minuta</i>	KU176938.1	51	<i>Medicago truncatula</i>	KT971339.1
7	<i>Oryza rufipogon</i>	NC_013816.1	52	<i>Vicia faba</i>	KC189947.1
8	<i>Oryza sativa</i> subsp. Japonica	NC_011033.1	53	<i>Vitis vinifera</i>	NC_012119.1
9	<i>Bambusa oldhamii</i>	EU365401.1	54	<i>Geranium maderense</i>	KP940515.1
10	<i>Sorghum bicolor</i>	DQ984518.1	55	<i>Beta macrocarpa</i>	FQ378026.1
11	<i>Zea luxurians</i>	DQ645537.1	56	<i>Beta vulgaris</i> subsp. maritima	FP885834.1
12	<i>Zea mays</i> subsp. mays	NC_007982.1	57	<i>Silene latifolia</i>	HM562727.1
13	<i>Zea perennis</i>	DQ645538.1	58	<i>Silene vulgaris</i>	JF750427.1
14	<i>Tripsacum dactyloides</i>	NC_008362.1			JF750428.1
15	<i>Lolium perenne</i>	JX999996.1			JF750429.1
16	<i>Triticum aestivum</i>	AP008982.1			JF750430.1
17	<i>Brassica carinata</i>	JF920287.1	59	<i>Daucus carota</i> subsp. sativus	JQ248574.1
18	<i>Brassica nigra</i>	KP030753.1	60	<i>Helianthus annuus</i>	KF815390.1
19	<i>Sinapis arvensis</i>	KM851044.1	61	<i>Vaccinium macrocarpon</i>	KF386162.1
20	<i>Eruca vesicaria</i> subsp. sativa	KF442616.1	62	<i>Viscum album</i>	KJ129610.1
21	<i>Raphanus sativus</i>	JQ083668.1	63	<i>Rhazya stricta</i>	KJ485850.1
22	<i>Brassica juncea</i>	NC_016123.1	64	<i>Asclepias syriaca</i>	KF541337.1
23	<i>Brassica rapa</i> subsp. campestris	JF920285.1	65	<i>Cynanchum wilfordii</i>	MH931257
24	<i>Brassica oleracea</i>	KJ820683.1			MH931258
25	<i>Brassica napus</i>	NC_008285.1			MH931259
26	<i>Schrenkiella parvula</i>	KT988071.2	66	<i>Cynanchum auriculatum</i>	MH931260
27	<i>Arabidopsis thaliana</i>	NC_037304.1	67	<i>Hesperalaea palmeri</i>	KX545367.1
28	<i>Batis maritima</i>	KJ820684.1	68	<i>Boea hygrometrica</i>	JN107812.1
29	<i>Carica papaya</i>	EU431224.1	69	<i>Mimulus guttatus</i>	JN098455.1
30	<i>Corchorus capsularis</i>	KT894204.1	70	<i>Castilleja paramensis</i>	KT959112.1
31	<i>Corchorus olitorius</i>	KT894205.1	71	<i>Salvia miltiorrhiza</i>	KF177345.1
32	<i>Gossypium barbadense</i>	KP898249.1	72	<i>Ajuga reptans</i>	KF709392.1
33	<i>Gossypium harknessii</i>	JX944506.1	73	<i>Capsicum annuum</i>	KJ865409.1
34	<i>Gossypium hirsutum</i>	JX944505.1	74	<i>Solanum lycopersicum</i>	NC_035963.1
35	<i>Gossypium raimondii</i>	KU317325.1	75	<i>Solanum tuberosum</i>	MF989953.1
36	<i>Ricinus communis</i>	HQ874649.1			MF989954.1
37	<i>Populus tremula</i>	KT337313.1			MF989955.1
38	<i>Salix purpurea</i>	KU198635.1			MF989956.1
39	<i>Salix suchowensis</i>	KU056812.1			MF989957.1
40	<i>Citrullus lanatus</i>	GQ856147.1	76	<i>Solanum commersonii</i>	MF989960.1
41	<i>Cucumis sativus</i>	NC_016004.1			MF989961.1
		NC_016005.1	77	<i>Hyoscyamus niger</i>	KM207685.1
		NC_016006.1	78	<i>Nicotiana tabacum</i>	NC_006581.1
42	<i>Cucurbita pepo</i>	GQ856148.1	79	<i>Nicotiana sylvestris</i>	KT997964.1
43	<i>Malus hupehensis</i>	KR534606.1	80	<i>Liriodendron tulipifera</i>	KC821969.1
44	<i>Malus x domestica</i>	NC_018554.1	81	<i>Nelumbo nucifera</i>	KR610474.1

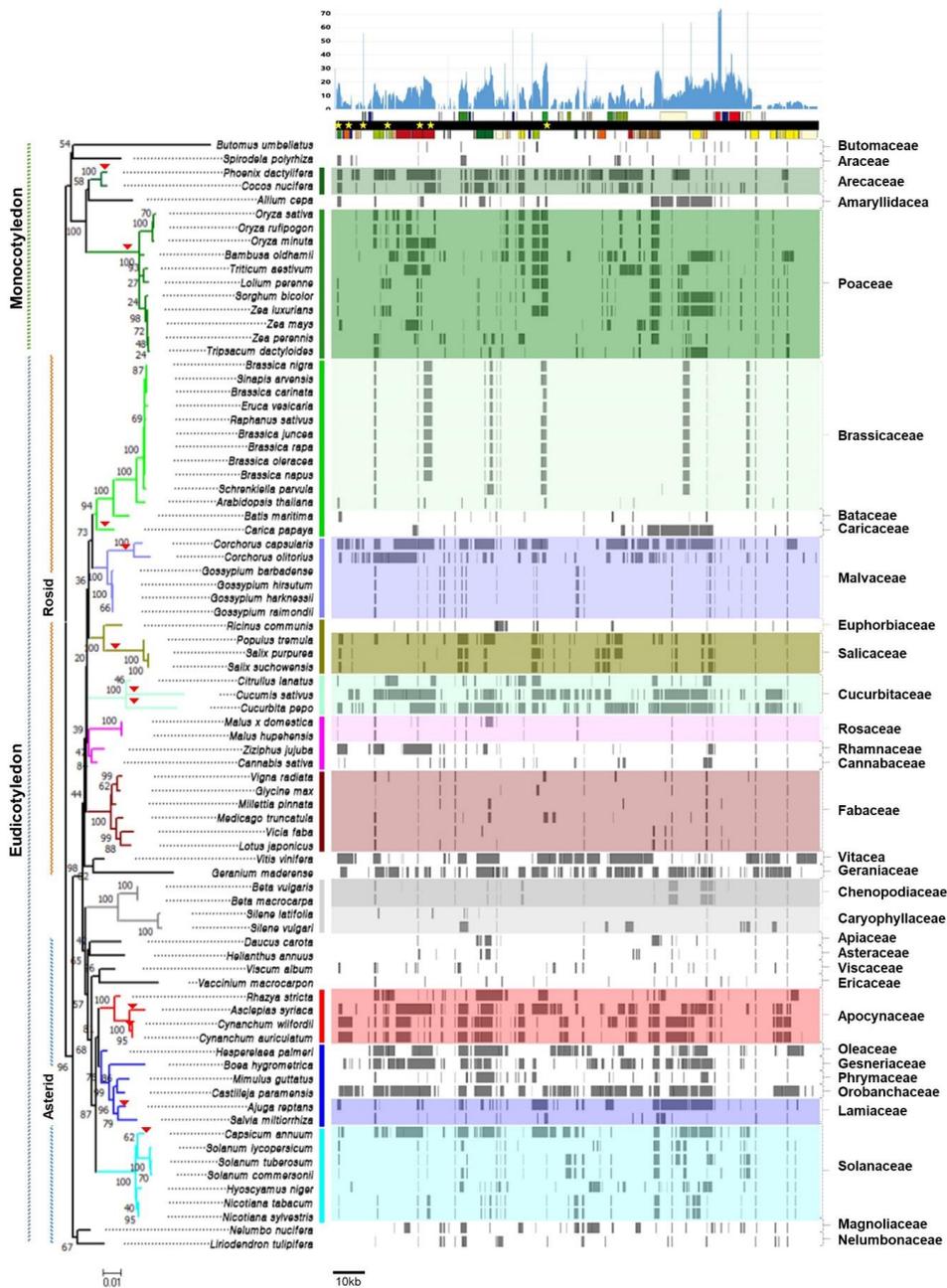


Figure 1-4. Schematic representation of MTPTs in 81 plant mitochondrial genomes. The plastid genome sequence of *Arabidopsis thaliana* was used as a backbone for the MTPTs in the mitochondrial genome of each species. The linear plastid genome map was generated by Organelle Genome DRAW, and depths of MTPTs are represented on the plastid genome map. Seven universal land plant barcoding regions¹⁴ are

marked on the map with yellow stars. The MTPT fragments in each mitochondrial genome were represented as gray blocks. tRNA and rRNA regions are not represented. The phylogenetic relationship was reconstructed using the *matR* sequences of 81 mitochondrial genomes. Areas of recent putative plastid–mitochondrial genome flux are marked with red triangles on the tree.

The MTPT distribution was coincident with the taxonomical groupings based on mitochondrial genes for the 81 plant species. In addition, the same pattern of MTPT distribution was identified at the family or genus level, with some exceptions. Certain species, such as two *Corchorus* species and *Capsicum annuum*, showed unique, highly abundant MTPT patterns that were extremely different from those of closely related species (denoted as arrowheads on the phylogenetic tree of **Figure 1-4**. MTPTs corresponded to 5.7% and 33.1% of the plastid genome in mitochondrial genomes of *S. lycopersicum* and *C. annuum*, respectively, indicating that there has been recent additional plastid–mitochondrial genome flux in *C. annuum* (**Figure 1-5 and 6**).

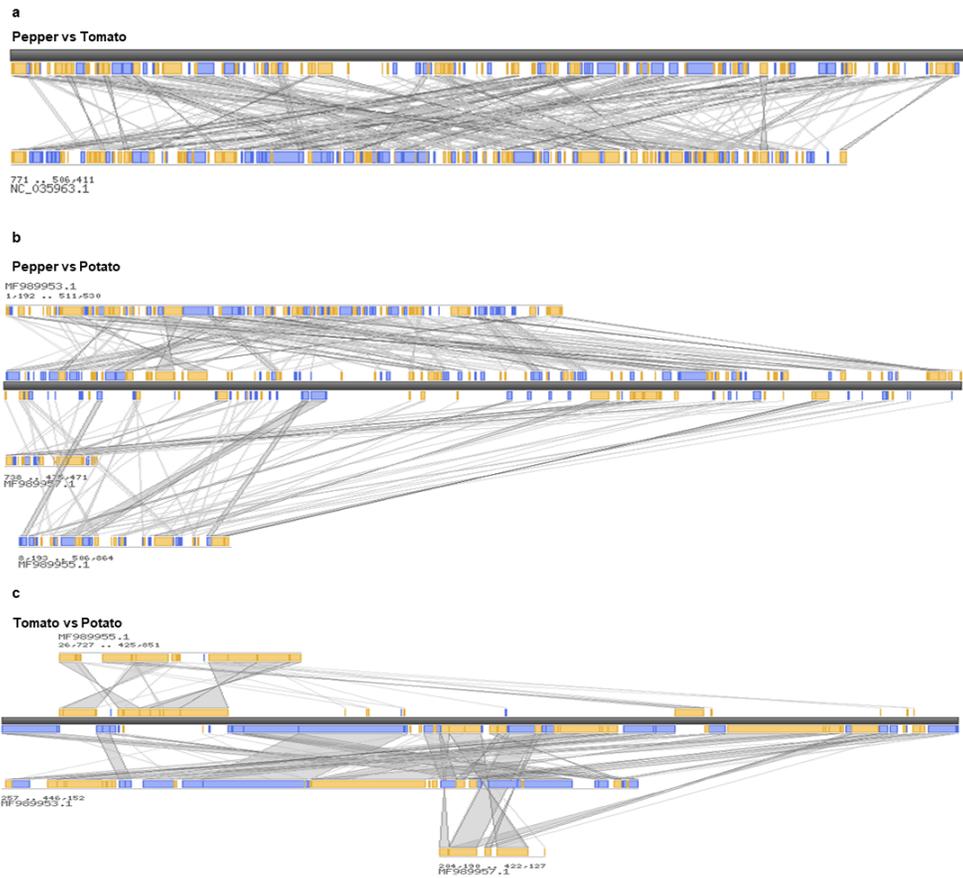


Figure 1-5. Genome-wide comparison of mitochondrial genomes of three Solanaceae species: pepper, tomato, and potato.

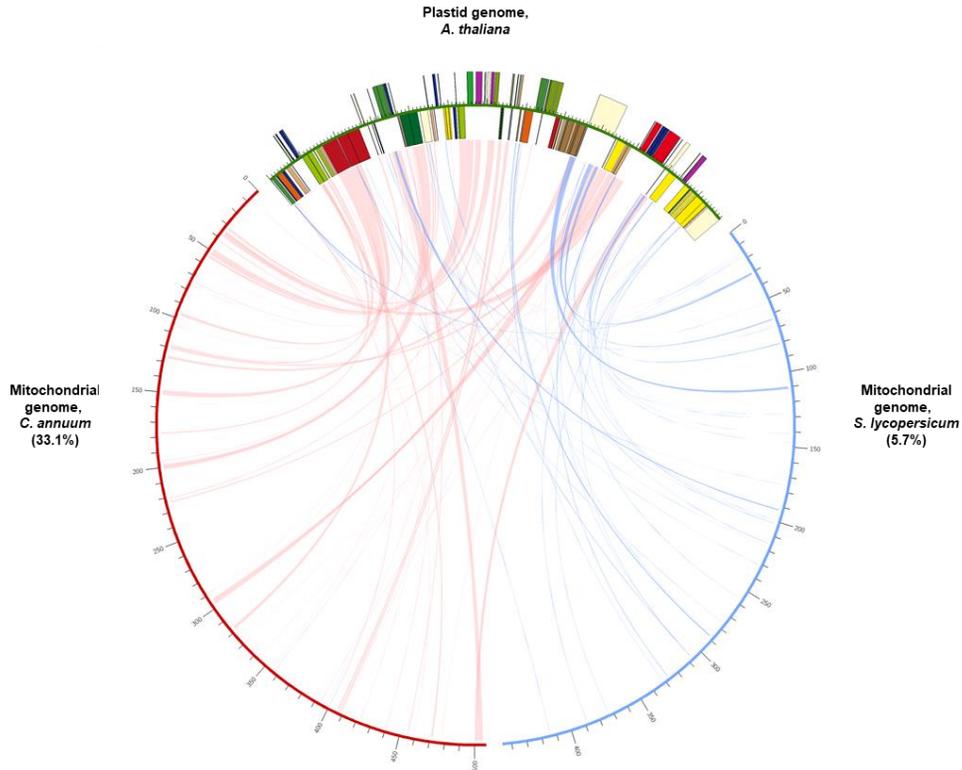


Figure 1-6. Circos plot showing MTPT in *Capsicum annuum* and *Solanum lycopersicum*. The plastid genome sequence was compared with the mitochondrial genomes *C. annuum* and *S. lycopersicum*. The plastid genome (top) and its MTPT counterparts are denoted by blue and red brackets for the mitochondrial genomes of *C. annuum* and *S. lycopersicum*, respectively.

DNA barcode markers based on inter- and intra-species plastid polymorphism

I identified polymorphic sites from the two *Cynanchum* species by pairwise alignment of plastid sequence. From this I developed a total of 12 DNA markers, including seven SNPs and five InDels, for authentication of each species (**Table 1-5**). I inspected three of these markers under different PCR conditions and found that two of them target the polymorphic plastid regions that are homologous to MTPT counterparts in the mitochondrial genomes of both species, while the third targets a

polymorphic plastid region that has no MTPT counterpart (**Figure 1-7**). The first marker is a Ca-specific marker derived from a SNP associated with *matK* and is currently used to detect Ca contamination in Cw products under regulation by the Ministry of Food and Drug Safety (MFDS) of Korea^{14,15} (**Figure 1-7a**). The second is based on codominant primers targeting a 348-bp InDel polymorphism located in the intergenic spacer (IGS) region between *rps2* and *rpoC2*, which produce a 481-bp band for Cw and a 135-bp band for Ca (**Figure 1-3b**). The third marker is also based on codominant primers, in this case targeting the IGS region between *rpoB* and *trnC-GCA* and producing a 347-bp and a 428-bp band for Cw and Ca, respectively. All three markers showed the expected genotype for the plastid genomes of both species through a moderate number of PCR amplification cycles (less than 25). However, for the first and second markers, unexpected bands were also detected upon increasing the number of PCR cycles or template DNA amounts (**Figure 1-7b,c**). I found that these unexpected bands were derived from MTPT targets. The MTPTs of both species are almost identical both to each other and to the *matK* gene of the Ca plastid (**Figure 1-7a and 1-3a**). Further, read mapping of WGS indicated that 88% and 12% of NGS reads were derived from the plastid and mitochondrial genomes, respectively (**Figure 1-8**, and **Table 1-6**). These numbers might represent copy numbers of the plastid and mitochondrial genomes within a cell and explain the different amounts of PCR products, i.e., the appearance of weaker bands derived from MTPTs compared to intense bands for the Ca_s_1 marker and Cw_i_1 marker in Cw and Ca, respectively (**Figure 1-7b,c**). I also designed nine additional DNA markers: three InDel and six SNP targets from plastid without MTPT homologs. All nine showed clear authentication of Cw and Ca without any noise, even after larger numbers of PCR cycles, such as the third marker in **Figure 1-7b** (**Figure 1-9**).

I assembled four more plastid genomes from different Cw collections and identified 11 targets showing intraspecies diversity—6 SNPs

and 5 InDel regions—among the five plastid genomes (**Figure 1-10**). I designed three additional markers from the five InDels and inspected Cw and Ca collections (**Figure 1-11, Table 1-5**). The three markers showed 2–3 haplotypes based on copy numbers of tandem repeats. When I assessed the genotypes of 27 Cw collections using these markers, 12 showed the same genotype as Ca for at least one of the three markers.

Table 1-5. Plastid DNA markers used in authentication of *Cynanchum* species.

Primer	Sequence (5' to 3')	Gene name	Product size (bp)	
			<i>C. wilfordii</i>	<i>C. auriculatum</i>
Cw_i_1	F: TCGGCAAGAACAATTCCTGT R: TGACCGTCCAATTTGTGTAGA	<i>rps2 – rpoC2</i>	481	135
Cw_i_2	F: AGATGATCTAGCAACGATGGGA R: CGGGTATTCAAGCGGATTGG	<i>rpoB – trnC-GCA</i>	347	428
Cw_i_3	F: TACACAAGCAGCAGGATGCC R: CGGTTTCGAGTCCGTATAGCC	<i>ndhC – trnV-UAC</i>	468	407
Cw_i_4	F: ACTCGGCCAATCTTTTCCT R: TGTGGATTCAAGACAACAAT	<i>rbcL – accD</i>	230	310
Cw_i_5	F: GTCTGAGACGGCCAGAAAG R: CCCGAAAGAACCGGACATGA	<i>petD</i> intron	269	301
Ca_s_1	F: CTGTTCCTCAATTATTC R: AATGAGAAAAGTTTCTG	<i>matK</i>	151	151
Cw_s_2	F: GCCGAATCCTTCTAGAGCCC R: CGGACGTTCCAGTGGACATT	<i>rpoB</i>	115	115
Cw_s_3	F: AGCGATCTTTTCGTAGACGTT R: TTCCCTTGTTCACTAATAAATCGAC	<i>rps18</i>	100	100
Cw_s_4	F: CCAAGACGAACTAATGCAGGG R: TTGCGACACCCATCAAAGGA	<i>psbH</i>	100	100
Cw_s_5	F: CATTCCCGCAGGAGATCCG R: ACTCCAGGGATGAATCGAAAAAGA	<i>ycf2</i>	120	120
Cw_s_6	F: AACCAATAGCGATTCATACAAGC R: TGGATTTGGATAAAGAGAACCATCT	<i>ycf1</i>	118	118
Cw_s_7	F: GCAAATTGATAAAACCCGGCG R: AGAGTTGAAGCCCCAAAGGG	<i>ndhH</i>	103	103

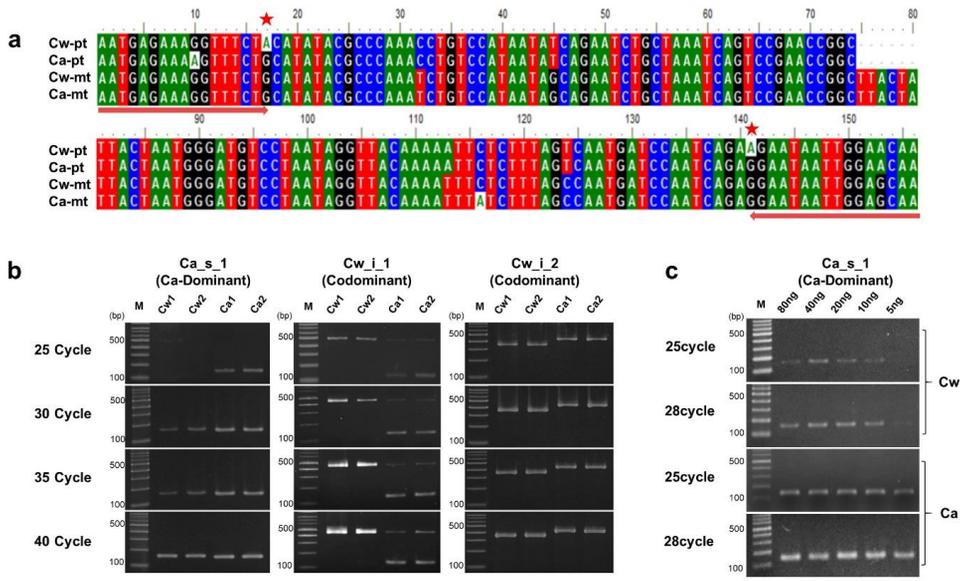


Figure 1-7. DNA marker paradox derived from MTPTs. (a) Amplicon of Ca-specific marker (Ca_s_1) designed from the *matK* region and its counterpart MTPT in *C. wilfordii* (Cw) and *C. auriculata* (Ca). Pt, plastid genome; mt, MTPT segment in mitochondrial genome. Primer regions and target SNPs for authentication of Cw and Ca are marked with red arrows and red stars, respectively. **(b)** Electrophoresis of PCR products from three authentication markers for *Cynanchum* species after different numbers of PCR cycles. Ca_s_1 is a Ca dominant marker, and Cw_i_1 and Cw_i_2 are codominant markers. Two accessions each of Cw and Ca were used. **(c)** Results from electrophoresis of PCR products obtained from Cw and Ca samples using the Ca-specific marker Ca_s_1 with different amounts of template DNA and PCR cycles

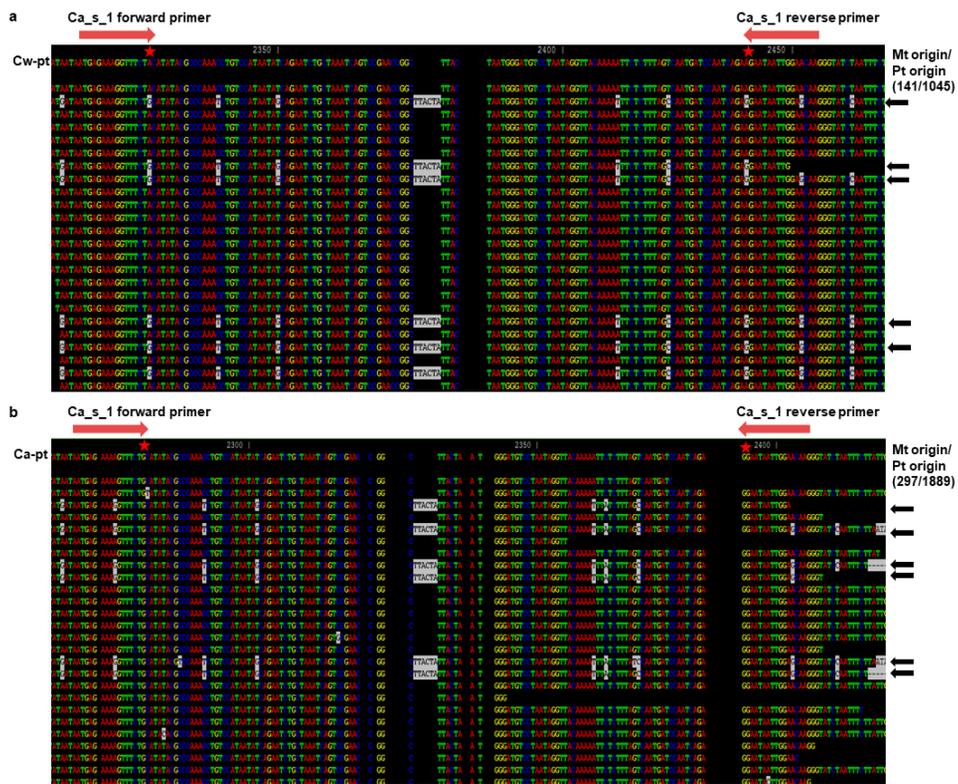


Figure 1-8. Estimated proportion for plastid and mitochondrial genomes using WGS of *C. wilfordii* (Cw) and *C. auriculatum* (Ca). (a and b), A partial image capture for NGS read mapping on the *matK* target using WGS of Cw and Ca, respectively. Red arrows represent primer regions and red stars indicate target SNP for discriminating Cw and Ca. The NGS reads putatively derived from mitochondrial genome are marked with black arrows.

Table 1-6. The estimated proportion of NGS reads for plastid and mitochondrial genomes in Cw and Ca.

Plants	Total reads (No.)	Total NGS data (Gbp)	Read depth at SNP sites*			
			Forward primer		Reverse primer	
			Pt origin	Mt origin	Pt origin	Mt origin
<i>C. wilfordii</i>	3,064,122	0.7	1,045 (a)	141 (G)	1,148 (a)	145 (G)
			88%	12%	88%	12%
<i>C. auriculatum</i>	3,350,936	1.0	1889 (G)	297 (G)	1833 (G)	246 (G)
			86%	14%	88%	12%

*Pt, Plastid; Mt, Mitochondria;

*The proportion was calculated based on read depth for the polymorphic sites between the plastid and counterpart MTPT (**Figure 1-8**). The haploid genome size of Cw was estimated as 430 Mbp, and plastid and mitochondrial genome copy numbers are estimated as 1,241 and 170 copies, respectively, in a somatic cell (2n).

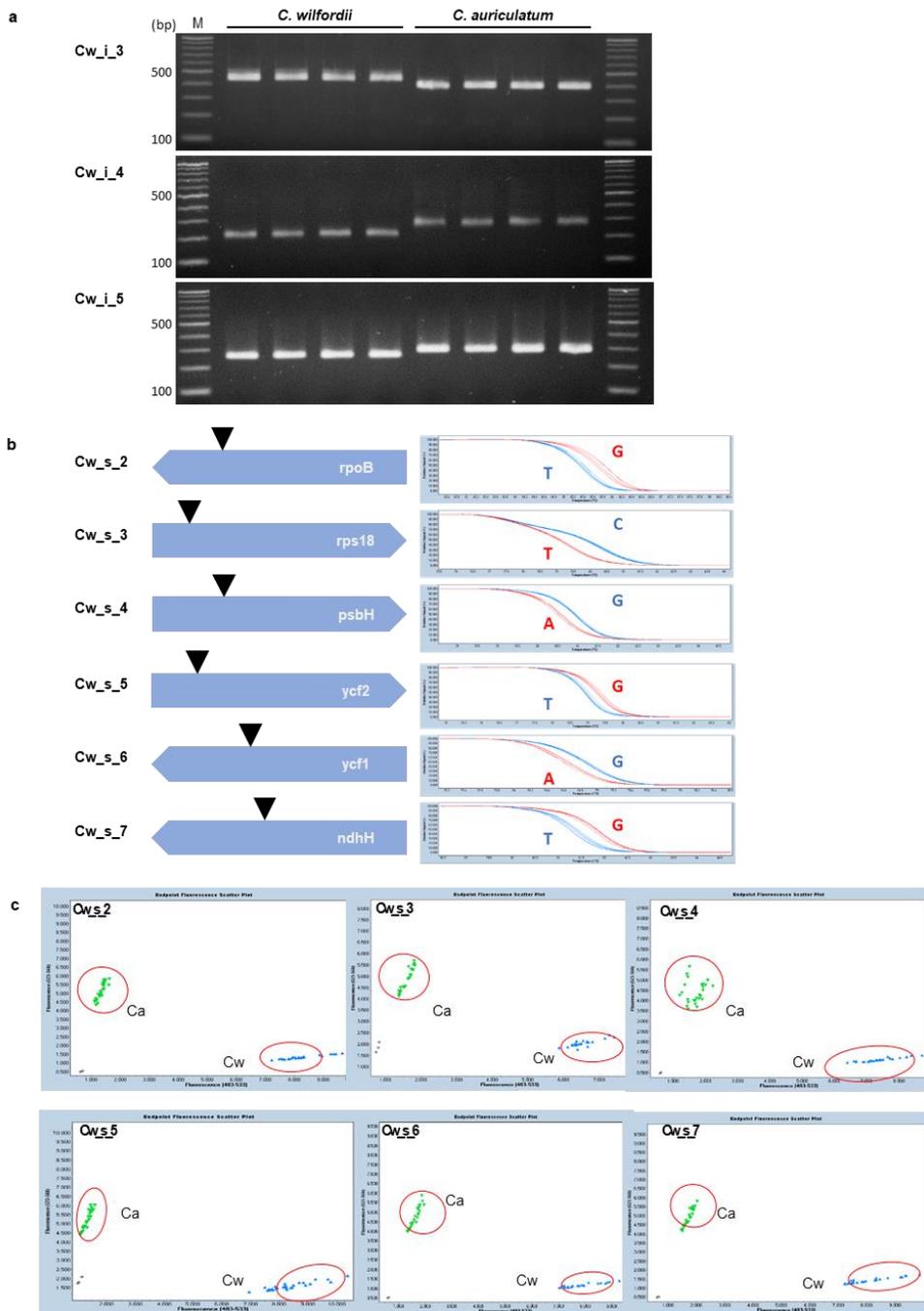


Figure 1-9. Nine additional plastid markers without MTPT counterparts for authentication of *C. wilfordii* (Cw) and *C. auriculatum* (Ca). (a) Agarose gel electrophoresis of four co-dominant markers for authentication of *Cynanchum*. (b) Genic SNPs between two *Cynanchum* species and application of HRM analysis for authentication. (c) Scatter plot

of KASP markers derived from the SNPs of (b) applied to the Cw and Ca populations

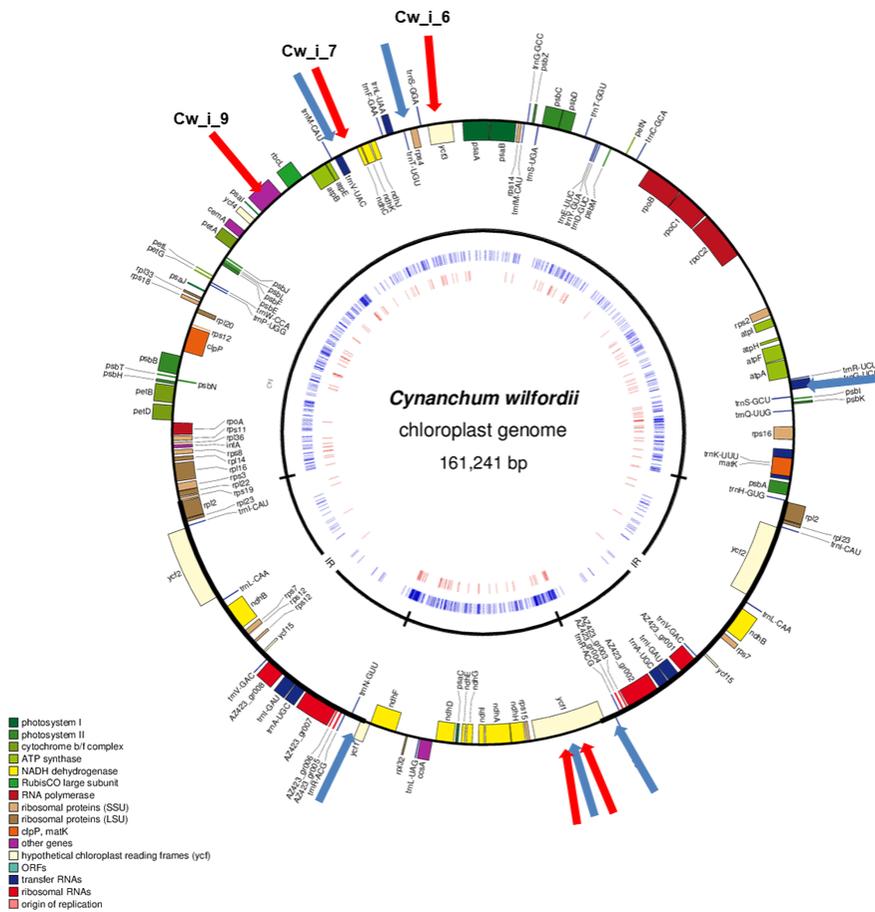


Figure. 1-10. Plastid map and intra- and inter-species polymorphism of *C. wilfordii*. Intraspecies diversities, InDels and SNPs, are denoted by red and blue arrows, respectively. Three InDel markers are denoted with marker names. Inter-species variations, InDels and SNPs, between Cw and Ca are shown at the inner circle with red and blue lines, respectively.

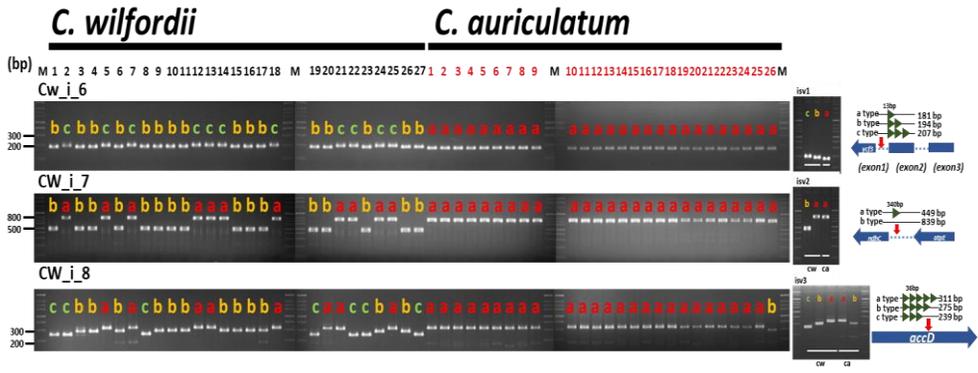


Figure 1-11. Genotyping of *C. wilfordii* and *C. auriculatum* collections for three plastid-derived markers

DISCUSSION

Plant mitochondrial genomes are extremely large, with sizes varying from 200 to 2000 kb (**Figure 1-12a**), 15–125 times larger than the conserved 16 kb animal mitochondrial genomes. In addition, horizontal genome flux between the plastid, mitochondrial and nuclear genomes is an established but underappreciated phenomenon in plants. *Cynanchum* species harbor a single type of plastid genome but multi-chromosomal mitochondrial genomes, a type of structure that in other plant species is known to be derived from tandem repeat-mediated recombination¹⁶⁻¹⁸. Here, I detected varying degrees of plastid genome flux, with additional lineage-specific patterns, in a study of more than 80 angiosperm mitochondrial genomes, showing that frequent plastid genome flux is very common in angiosperms. Intriguingly, a very recent transfer found in some species, including *C. annuum*, indicates that plastid flux into the mitochondrial genome occurs more frequently during plant evolution than was previously imagined¹⁹⁻²³. Such dynamic plastid-mitochondrial genome flux might contribute to diversifying the mitochondrial genome structure by contribution of 2–75 kb MTPTs in 81 plants (**Figure 1-12b**) Plastid DNA copy numbers vary in different species and in different tissues of plants. The plastid genome copy numbers decline rapidly, from 600 copies to fewer than 100 copies, during 5 days of dark treatment. The plastid DNA is degraded by the organelle exonuclease DPD1 and contributes to phosphorus relocation²⁴. When the plastid DNA is not completely degraded, the abundant fragments of plastid DNA can occasionally be horizontally transferred into other cellular genomes via a mechanism involving the double-strand break repair system of plant genome²⁵.

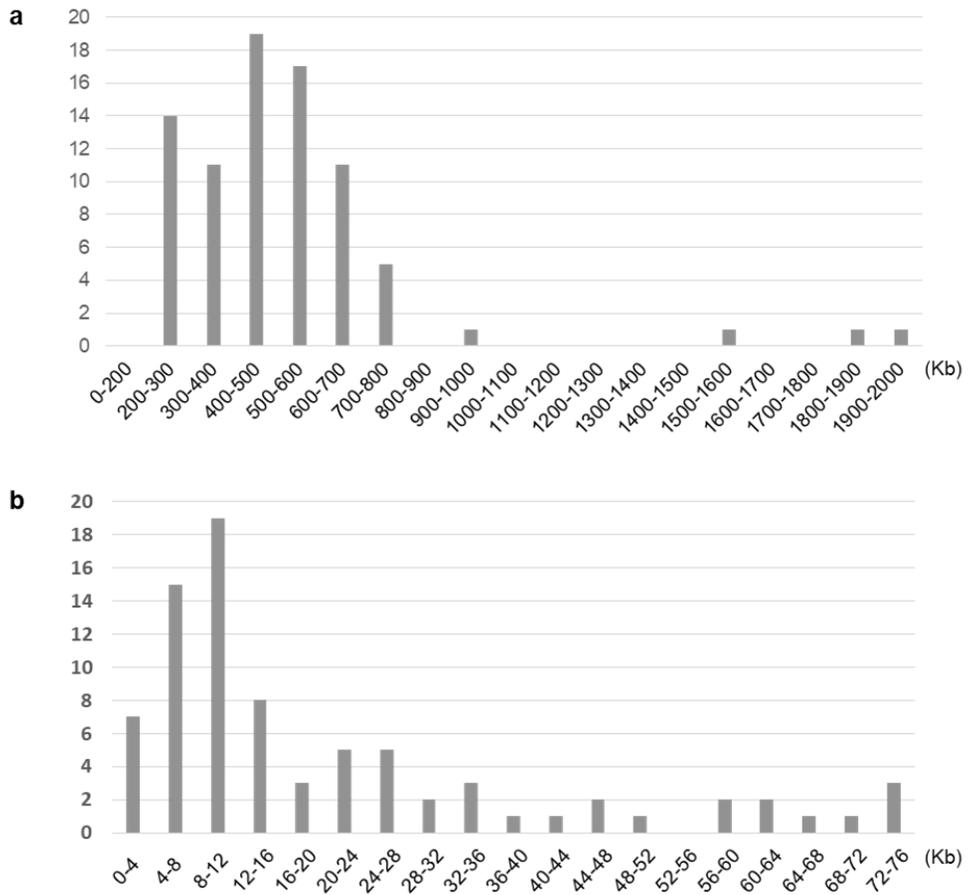


Figure 1-12. Mitochondrial genome sizes and MTPT amounts in the 81 plants used in this study. (a) Mitochondrial genome size distribution. (b) Distribution of MTPT lengths within the mitochondrial genomes.

To date, only 81 plant mitochondrial genome sequences have been reported, although thousands of plastid genomes have been characterized. This discrepancy is derived from the complex mitochondrial genome structure, with an abundance of MTPTs, size variance, and a high frequency of recombination. My results show that the mitochondrial genomes, including MTPTs, exhibited low nucleotide substitution rates, showing more than three times slower evolution compare to plastid genomes²⁶ (**Figure 1-2c**). Importantly, I demonstrated that plant barcoding targets are found in MTPTs (**Figure 1-4a**) at meaningful frequencies in a diverse array of plants. Overall, I have uncovered a notable shortcoming of the common practice of barcode-based authentication of plant products by demonstrating that a wide range of horizontal transfer events involve MTPTs, which are likely to result in co-amplification of unexpected bands (**Figure 1-7b,c**). These promiscuous DNAs could affect molecular taxonomy by causing mis-positioning of species if the plastid DNA sequences used were confused with those of MTPTs⁵.

EMA is estimated to have cost the herbal supplement industry 10–15 billion dollars per year since 2010². The impact of mis-authentication caused by the DNA marker paradox could have severe negative effects not only on the industry but also on all parties involved in the herbal supplement industry, from farmers to consumers and beyond, as demonstrated by the involvement of both the New York State attorney general’s office and the Korean Supreme Court in regard to *Cynanchum* products in 2015–2017. To avoid incorrect application of DNA markers and escape the DNA marker paradox, it is desirable to use multiple markers derived from different loci, which can credibly distinguish EMA from target plant products. However, the number of markers required to detect EMA does not need to be as high as that used in forensics to distinguish individuals of the same species (*Homo sapiens*).

Unlike major crops, herbal medicinal plants still have a wide range of natural intraspecies diversity because diverse undomesticated collections are wildcrafted⁶⁻⁹. I inspected haplotype diversity for 3 of the 11 intraspecies polymorphic sites revealed from a comparison of the five plastid genomes (**Figure 1-10, 1-11, 1-13, and Table 1-5**). Based on the genotypes for the three markers, 27 Cw collections were classified into 5 groups, unlike Ca individuals, which are derived from a few Chinese collections and thus showed narrow genetic diversity. Notably, variation of the markers would place 12 of the 27 Cw collections as Ca, and 1 of the 26 Ca collections as Cw, if genotyping were based on data from individual markers. These findings emphasize the importance of marker choice and of relying on more than one marker for authentication purposes.

To define the difficulty more generally, if the target species shows homogeneous genotypes for three loci, denoted *A*, *B* and *C*, the $A_1B_1C_1$ genotype can be clearly distinguished from the $A_2B_2C_2$ genotype. However, I found that many wild accessions have heterogeneous genotypes, such as $A_1B_1C_2$ or $A_2B_1C_2$, as a result of intraspecies plastid genome diversity (**Figure 1-13**) or plastid–mitochondrial genome flux (**Figure 1-7**)^{9,27}. If I were to focus only on the negative markers and assume that the heterogeneous genotypes represented counterfeit products (i.e., the negative detection method), the chance of detecting the counterfeits (the detection power) would be increased, but the possibility of a decision error that defines the genuine product as fake (the rate of false-positive, type I error) would also be increased. By contrast, if I were to focus on the positive markers and to assume that the heterogeneous genotypes represent the target product (i.e., the positive detection method), I could expect the opposite result: I would accept diverse genotypes as representing the target species and thereby reduce false-positive, but I would also reduce the detection power (**Figure 1-13**). If I applied the negative detection method to increase the detection power of an assay, the false-positive error rate would be increased; likewise, if I applied the positive detection method to reduce the false-positive error rate, the detection power would be reduced.

In general, medicinal plant resources are underdeveloped, and thus their genetic diversity is as yet unknown or underestimated. Given this necessary balancing act, as long as the counterfeit plant is nontoxic, I propose that the positive detection method should be applied to reduce false-positive errors, even if the detection power is reduced and the false-negative (Type II error) is thereby increased. This approach could reduce unforced errors resulting in the sanctioning of genuine products and thereby protect the industry. However, in cases where there are safety concerns, such as toxicity, associated with the EMA counterpart, the negative detection method should be applied to maximize the detection power even at the cost

of more false-positive errors, because trace amounts of adulterant could be a threat both to consumer health and to the long-term success of the industry.

The raw materials for most herbal products are heterogeneous because they are collected from natural habitats or cultivated from wild collections. Breeding of superior cultivars and the establishment of quality management systems should be encouraged as means to produce consistent functional foods. The production of functional foods from specific cultivars managed from seed to final products with traceability will benefit from the development of scientifically well-supported cultivar-specific markers, will minimize the damage from EMA and will thus promote the growth of the functional food and herbal supplement industry.

Materials and Methods

Assembly of plastid and mitochondrial genomes

We produced Illumina platform WGS Paired-End data using genomic DNA from individual plants for *Cynanchum wilfordii* (Cw) and *Cynanchum auriculata* (Ca). I assembled plastid genome sequence^{11,12} (GenBank accession, NC_029459, NC_029460) and mitochondrial genome sequence (GenBank accession MH931257, MH931258, MH931259, MH931260) using the WGS datasets. After trimming of low-quality reads, the assembly was done using CLC genome assembler (ver_4.01). The mitochondria-related contigs were extracted from the assembled contigs by comparison with the mitochondrial sequence of *A. syriaca* (NC_022796) using Nucmer²⁸. After removal of contigs that showed more than 99% identity with the plastid genomes, the extracted contigs were combined for each type of mitochondrial genome and validated by mapping of raw reads using the dnaLCW method^{6,7}.

Identification of MTPTs in mitochondrial genome

To detect MTPTs in plant mitochondrial genome sequences, I compared 81 mitochondrial genome sequences, including the newly assembled mitochondrial genomes of Cw and Ca and 79 published plant mitochondrial genomes (**Table 1-4**) retrieved from the NCBI database. Each mitochondrial genome sequence was compared with the plastid genome sequence of *A. thaliana* using BLASTN²⁹ based on more than 70% identity from strands of at least 30 bp with an expectation value of 1E-10. The regions of the plastid genome matching transfer RNA (tRNA) or ribosomal RNA (rRNA) were ignored and excluded from this study because tRNA and rRNA are conserved in both organelle genomes.

Plastid marker design and PCR amplification

The plastid sequences of the *Cynanchum* species were aligned using

MAFFT³⁰. After identifying polymorphic regions by pairwise alignment of the Cw and Ca chloroplast genomes, I mapped each raw NGS read to the corresponding chloroplast genome to detect promiscuous regions including MTPT. SNPs or InDels showing heterogeneous genotypes for more than 3% of each read depth were eliminated. Primers for codominant markers and high-resolution melting curve (HRM) analysis were designed using Primer-BLAST³¹. I selected InDel polymorphic regions whose length was greater than 20 bp and converted those regions to the codominant marker. In case of SNP, I filtered out adjacent SNPs less than 150 bp in length and designed primers flanking candidate SNPs with lengths from 100 to 150 bp.

The polymerase chain reactions (PCRs) was conducted with pre-amplification at 95°C for 5 min; 25–40 cycles of denaturation at 95°C for 30 sec, annealing at 58°C for 30 sec, and elongation at 72°C for 30 sec; and a final elongation at 72°C for 5 min. The mixture consisted of a total volume of 25 µL set up using the IncloneTM Taq DNA Polymerase Kit (Inclone, South Korea) with 5–80 ng of template DNA, 1× PCR reaction buffer, 0.2 mM of each dNTP, 0.2 pmol of each primer (Bioneer, South Korea) and 0.4 units Taq DNA polymerase. Amplification of Ca-specific markers was done following the procedure defined in an earlier paper¹⁵. High-resolution melting-curve (HRM) analysis was performed with detection by a LightCycler 480 real-time PCR machine (Roche Applied Science, Indianapolis, IN, United States) under pre-amplification at 95°C for 5 min followed by 45 cycles of 95°C for 30 sec, 58°C for 30 sec, and 72°C for 30 sec. The concentration of the mixture was 20 µL of reaction volume containing 20 ng of template DNA, 0.5 pmol of each primer (Bioneer, South Korea) and 1× Pre-mix of RealHelixTM Premier qPCR Kit (Nanohelix, South Korea).

The Kompetitive allele-specific PCR (KASP) markers were designed from the same SNP positions used for HRM analysis with Kraken software (LGC Genomics, Hoddson, UK). Thermocycling and endpoint

genotyping for the KASP assays were applied to the population of Cw and Ca using a Roche LC480 (Roche Applied Science, Indianapolis, IN, United States) following the manufacturer's instructions in a total volume of 10 μ L containing 100 ng of template DNA, 0.14 μ L of KASP assay mix and 5 μ L of KASP reaction mix (**Figure 1-9**).

Ks value calculation for genes in plastid and mitochondrial genomes.

We calculated the level of synonymous substitutions per synonymous site (Ks) between homologous genes in the plastid genomes (plastid genes, PT), mitochondrial genes (MT) and MTPTs of four Apocynaceae species. Pairwise alignments of the coding sequences for common genes derived from each organelle were conducted using webPRANK³² based on translated codons. After alignment, InDel regions and stop codons were trimmed using GBLOCKS³³, and Ks values were calculated using CODEML from the PAML package³⁴. Divergence time was estimated as $Ks/2\lambda$, where λ signifies the synonymous substitution rate of 2×10^{-9} for the plastid genomes and 0.6×10^{-9} for the mitochondrial genomes.

REFERENCES

- 1 Poornima, B. Adulteration and substitution in herbal drugs a critical analysis. *IJRAP* **1**, 8-12 (2010).
- 2 Johnson, R. Food fraud and economically motivated adulteration of food and food ingredients. Congressional Research Service Washington, DC (2014).
- 3 Chang, A., Kwak, B. Y., Yi, K. & Kim, J. S. The effect of herbal extract (EstroG-100) on pre-, peri-and post-menopausal women: a randomized double-blind, placebo-controlled study. *Phytotherapy research* **26**, 510-516 (2012)
- 4 Barnard-Kubow, K. B., McCoy, M. A. & Galloway, L. F. Biparental chloroplast inheritance leads to rescue from cytonuclear incompatibility. *New Phytologist* **213**, 1466-1476 (2017).
- 5 Song, H., Buhay, J. E., Whiting, M. F. & Crandall, K. A. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the national academy of sciences* **105**, 13486-13491 (2008).
- 6 Kim, K. *et al.* Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. *PloS one* **10**, e0117159 (2015).
- 7 Kim, K. *et al.* Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. *Scientific reports* **5**, 15655 (2015).
- 8 Massouh, A. *et al.* Spontaneous chloroplast mutants mostly occur by replication slippage and show a biased pattern in the plastome of *Oenothera*. *The Plant Cell* **28**, 911-929 (2016).
- 9 Joh, H. J. *et al.* Authentication of golden-berry *P. ginseng* cultivar ‘Gumpoong’ from a landrace ‘Hwangsook’ based on pooling method

- using chloroplast-derived markers. *Plant Breed Biotech* **5**, 16-24 (2017).
- 10 Nguyen, V. B. *et al.* Comprehensive comparative analysis of chloroplast genomes from seven *Panax* species and development of an authentication system based on species-unique SNP markers. *Journal of Ginseng Research* <https://doi.org/10.1016/j.jgr.2018.06.003> (2018).
- 11 Jang, W. *et al.* The complete chloroplast genome sequence of *Cynanchum auriculatum* Royle ex Wight (Apocynaceae). *Mitochondrial DNA Part A* **27**, 4549-4550 (2016).
- 12 Park, H.-S. *et al.* The complete chloroplast genome sequence of an important medicinal plant *Cynanchum wilfordii* (Maxim.) Hemsl.(Apocynaceae). *Mitochondrial DNA Part A* **27**, 3747-3748 (2016).
- 13 CBOL Plant Working Group. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* **106**, 12794-12797 (2009).
- 14 Enforcement Rule of the Health Functional Foods Act in *Ordinance of the Ministry of Food and Drug Safety*, Vol. 1386, 10 (2016).
- 15 Kim, J. H., Moon, J.-C., Kang, T. S., Kwon, K. & Jang, C. S. Development of cpDNA markers for discrimination between *Cynanchum wilfordii* and *Cynanchum auriculatum* and their application in commercial *C. wilfordii* food products. *Applied Biological Chemistry* **60**, 79-86 (2017).
- 16 Alverson, A. J., Rice, D. W., Dickinson, S., Barry, K. & Palmer, J. D. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. *The Plant Cell* **23**, 2499-2513 (2011).
- 17 Chen, Z. *et al.* Plant mitochondrial genome evolution and cytoplasmic male sterility. *Critical reviews in plant sciences* **36**, 55-

- 69 (2017).
- 18 Cho, K.-S. *et al.* The complete mitochondrial genome sequences of potato (*Solanum tuberosum* L., Solanaceae). *Mitochondrial DNA Part B* **2**, 781-782 (2017).
 - 19 Sloan, D. B., Müller, K., McCauley, D. E., Taylor, D. R. & Štorchová, H. Intraspecific variation in mitochondrial genome sequence, structure, and gene content in *Silene vulgaris*, an angiosperm with pervasive cytoplasmic male sterility. *New Phytologist* **196**, 1228-1239 (2012).
 - 20 Straub, S. C., Cronn, R. C., Edwards, C., Fishbein, M. & Liston, A. Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome Biology and Evolution* **5**, 1872-1885 (2013).
 - 21 Sloan, D. B. & Wu, Z. History of plastid DNA insertions reveals weak deletion and AT mutation biases in angiosperm mitochondrial genomes. *Genome biology and evolution* **6**, 3210-3221 (2014).
 - 22 Gandini, C. & Sanchez-Puerta, M. Foreign plastid sequences in plant mitochondria are frequently acquired via mitochondrion-to-mitochondrion horizontal transfer. *Scientific reports* **7**, 43402 (2017).
 - 23 Van de Paer, C., Bouchez, O. & Besnard, G. Prospects on the evolutionary mitogenomics of plants: a case study on the olive family (Oleaceae). *Molecular ecology resources* **18**, 407-423 (2018).
 24. Takami T, Ohnishi N, Kurita Y, Iwamura S, Ohnishi M, Kusaba M, Mimura T, Sakamoto W: **Organelle DNA degradation contributes to the efficient use of phosphate in seed plants.** *Nature plants* 2018:1.
 25. Bock R: **The give-and-take of DNA: horizontal gene transfer in plants.** *Trends in plant science* 2010, **15**:11-22.

- 26 Sloan, D. B. *et al.* Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS biology* **10**, e1001241 (2012).
- 27 Nguyen, V. B. *et al.* Authentication markers for five major *Panax* species developed via comparative analysis of complete chloroplast genome sequences. *Journal of agricultural and food chemistry* **65**, 6298-6306 (2017).
- 28 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome biology* **5**, R12 (2004).
- 29 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410 (1990).
- 30 Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics* <https://doi.org/10.1093/bib/bbx108> (2017).
- 31 Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC bioinformatics* **13**, 134 (2012).
- 32 Löytynoja, A. & Goldman, N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC bioinformatics* **11**, 579 (2010).
- 33 Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology* **56**, 564-577 (2007).
- 34 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591 (2007).

CHAPTER 2

Development of metabarcoding method for unknown plant mixture

ABSTRACT

Plant kingdom is extremely wide and diverse. DNA barcoding has contributed to classification of plants with rapid and accurate authentication. NGS sequencing progress promoted metabarcoding analysis, which can define the kinds and the relative amount of plant species in mixed materials. Here, I selected highly conserved sequence regions by comparative analysis of 23 chloroplast genomes of four taxonomically distant family in dicot, Araliaceae, Apiaceae, Papaveraceae, Cannabaceae and designed the multiple primer pairs for metabarcoding of dicot plant. The amplicon of each primer had enough sequence diversity to be differentiated at genus level which was represented by phylogenetic relationship. After the *in silico* PCR at plant order level and *in vitro* DNA amplification using various plant samples, I filtered the candidate primers successfully amplifying both of monocot and dicot plant as much as possible with appropriate size of single band for NGS library construction. I also included two barcoding targets from nuclear ribosomal DNA. Overall, I selected four universal barcode targets from chloroplast genome and included 18S and ITS2 targets which are in nuclear genome and common for microorganism and animal genome. NGS application of six amplicons revealed the rough estimation of ingredient in mixed samples based on read depth analysis, although additional methods should be needed to reduce the PCR bias between samples. These primer pairs and NGS application can be a useful guide for defining plant materials in a mixed sample based on metabarcoding quantification.

INTRODUCTION

DNA barcoding is a technique using a DNA sequence information for characterizing and authentication of species. It was started to overcome the limitation of conventional identification methods. Short DNA fragments amplified from several standardized barcoding regions are widely used, for example, *COI* region for animals, 16s rDNA region for microbiomes. For plants, the combination of *matK* and *rbcL*¹ and ITS2² region have been widely accepted universal barcoding regions. Plant DNA barcoding is widely used not only for species identification in taxonomy, but also practical purpose like quality control of raw materials in herbal product, protection of specific cultivars in horticulture and food crops. Chinese groups constructed ITS2 sequence library for medicinal plant used in traditional Chinese medicine (TCM) for standardization of their source and prevention of adulteration.

Conventional DNA barcoding process is started from collecting and extracting the DNA from individual samples. After PCR amplification using conventional or mini barcoding primers, sequence of fragment or PCR band itself become used for species authentication. These process is usually time and labor consuming and inappropriate from the mixed materials. To overcome these limitations, DNA barcoding has been combined with NGS platform and the reference sequence database, which is called the “Metabarcoding”. In this technique, the each NGS read from PCR amplicon of mixed samples serves as individual DNA barcode through the comparison with constructed barcoding database such as NCBI. This technique has been widely used for population analysis like intestinal floral³, rainforest biodiversity⁴, and dietary investigation⁵ using 12s or 18S rDNA for animal and 16s rDNA or ITS for microbiome.

However, in the case of plant, using the conventional universal barcoding primers for metabarcoding has several limitations. First, the

length of conventional barcoding regions for plants are more than 500bp, which are inappropriate for making NGS library. Next, there is no single locus to be able to cover all of the species diversity for plants like 16s rDNA or *COI*, so there have to be multiple markers for plant barcoding. Proposed locus up to date have different limitations such as low sequence recovery rate in *matK*, low sequence diversity in *rbcL*, heterogeneous nature of ITS2 due to immature concerted evolution of ribosomal gene. Combinations of two or more barcoding regions like *matK* with *rbcL*¹ were still failed to overcome their inborn limitations, so additional barcoding locus proper for metabarcoding needs to be developed.

The plants species in Apiales order have been traditionally utilized for their medicinal effect. Many of well-known herbal plants belong to this order such as *Panax ginseng*, *Panax quinquefolius*, *Angelica gigas*, *Peucedanum japonicum*, and *Foeniculum vulgare*. Due to complex taxonomy in this order, molecular markers are needed to authenticate these species properly but most of them are still lack of genome information, while only from carrot and ginseng, whole genome sequences of were published in 2016⁶ and 2018⁷ respectively. *Papaver* and *Cannabis* are genus of nationally regulated narcotics plants popular with opium poppy and cannabis. The draft genome sequence of them were reported in 2011 for *C. sativa*⁸ and 2018 for *P. somniferum*⁹. Several studies using molecular markers for discrimination were published for *Papaver*^{10,11} and for *Cannabis*¹²⁻¹⁴.

Here I designed new primer pairs for plant metabarcoding from multiple alignment of complete chloroplast genome of 23 species and conventional barcoding regions. Amplifying ability of primer pairs were evaluated by actual PCR to various real plant DNA and *in silico* primer specificity investigation with chloroplast genomes in Genbank. Finally, NGS analysis using four candidate primers from chloroplast two nuclear primer and was conducted with mixed DNA for application of them.

RESULTS

Selection of conserved regions in chloroplast genomes and primers design

I selected 23 chloroplast genome of 5 genera including *Panax*, *Peucedanum*, *Angelica*, *Cannabis*, *Papaver* species. Three genus, *Panax*, *Peucedanum*, and *Angelica*, belonged to Apiales order where various important medicinal species included. The other two genus, *Cannabis* and *Papaver*, belonged to Rosales and Ranunculales order respectively, where the regulated drug species were included (**Table 2-1**).

The consensus sequences was generated from multiple alignment result of the 23 chloroplast genomes and the variable regions were masked with N. Therefore, only conserved nucleotides among all of the 23 chloroplast genomes were remained, so that the primers designed from this template sequence are perfectly common in all of the chloroplast used. Initially, 30 primer pairs were designed under the length condition of 500bp and the primer pairs amplifying the product with less than 350bp at least one species were filtered, which resulting in 22 primer pairs (**Figure 2-1, Table 2-2**).

Table 2-1. List of chloroplast genomes used for primer design.

#	Species	Note
1	<i>Panax ginseng</i>	KM088019.1
2	<i>Panax quinquefolius</i>	KT028714
3	<i>Panax quinquefolius</i>	NC_27456.1
4	<i>Panax notoginseng</i>	
5	<i>Panax notoginseng</i>	KJ566590.1
6	<i>Panax notoginseng</i>	KT001509.1
7	<i>Panax japonicas</i>	NC_028703.1
8	<i>Panax vietnamensis</i>	NC_028704.1
9	<i>Panax bipinnatifidus</i>	KX247146.1
10	<i>Panax stipuleanatus</i>	NC_030598.1
11	<i>Angelica gigas</i>	NC_29393..1
12	<i>Peucedalum japonicum</i>	NC_34644.1
13	<i>Peucedalum japonicum</i>	IM150921_6
14	<i>Peucedalum japonicum</i>	IM150601_19
15	<i>Papaver somniferum</i>	NC_29434.1
16	<i>Papaver somniferum</i>	
17	<i>Papaver somniferum</i> var <i>setigerum</i>	
18	<i>Cannabis sativa</i>	KY084475.1
19	<i>Cannabis sativa</i>	
20	<i>Cannabis sativa</i>	Northernlight
21	<i>Cannabis sativa</i>	NC_26562.1
22	<i>Cannabis sativa</i>	KR779995.1
23	<i>Humulus lupulus</i>	NC_28032.1

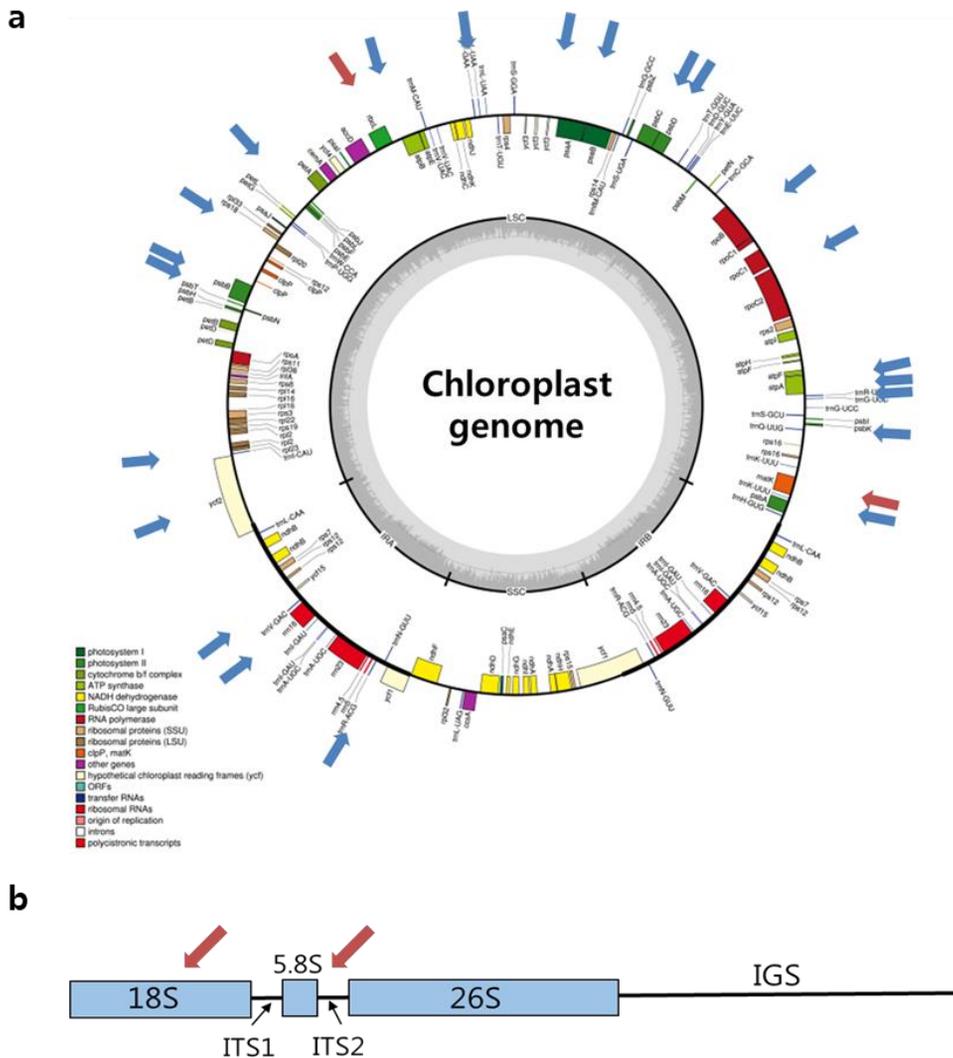


Figure 2-1. Location of designed primer pairs on genome map of chloroplast (a) and 45s nuclear ribosomal DNA (b). The blue arrows denoted the positions of newly designed amplicon in this study on the representative chloroplast genome map. The red arrows indicated markers from universal barcoding regions, *matK* and *rbcL* in chloroplast and 18S rDNA and ITS2 in 45s rDNA.

Table 2-2 Information of designed primer pairs

Name	Start*	END*	Product (bp)	Location
M1-L	924	1347	445~427	<i>psbA</i>
M2-H	1344	1772	447~504	<i>psbA - trnK-UUU</i>
M3-M	8206	8659	470~487	<i>trnQ - psbK</i>
M4-H	10978	11259	282~456	<i>trnG - trnR</i> intergenic region
M5-M	11237	11661	422~460	<i>trnR - atpA</i>
M6-M	34964	35436	473	<i>psbD</i>
M7-M	35591	36019	429	<i>psbD - psbC</i>
M9-M	43628	44017	379~390	<i>psaA - ycf3</i> intergenic
M10-M	50048	50488	351~441	<i>trnL-UAA - trnF-GAA</i>
M12-M	71667	72006	325~340	<i>rpl20 - clpP</i>
M13-M	74443	74967	525	<i>psbB</i>
M15-M	88234	88692	454~471	<i>trnI-CAU - ycf2</i> intergenic region
M18-M	110206	110669	454~529	<i>trnN-GUU - ycf1</i> intergenic region
M8-L	41331	41740	410	<i>psaB - psaA</i> intergenic region
M11-H	66007	66334	316~404	<i>psbJ - psbF</i> intergenic region
M14-M	76089	76517	405~448	<i>psbB - psbH</i> intergenic region
M16-L	93657	94035	343~388	<i>ycf2</i>
M17-M	100546	100838	273~398	<i>rps12 - trnV-GAC</i> intergenic region
M19-M	15465	15847	402	<i>atpF</i>
M20-M	24712	25057	369	<i>rpoC2 - rpoC1</i> intergenic region
M21-M	28683	29073	414	<i>rpoB</i>
M22-M	65934	66265	344	<i>atpB - rbcL</i> intergenic region
M23	2251	2678	428	Upstream of <i>matK</i>
M24	2251	2683	433	Upstream of <i>matK</i>
M25	2738	3136	399~417	Downstream of <i>matK</i>
M26	2738	3130	393~411	Downstream of <i>matK</i>
M27	2737	3137	401~419	Downstream of <i>matK</i>
M28	2738	3130	393~411	Downstream of <i>matK</i>
M29	57365	57708	344	Upstream of <i>rbcL</i>
M30	57682	58148	467	Downstream of <i>rbcL</i>
M31	57627	58151	525	Downstream of <i>rbcL</i>
18S	1272	1639	368	Coding region of 18S rDNA
ITS	3337	3731	395	ITS2 region

*Start and end positions were based on chloroplast genome and 45s rDNA of *Panax ginseng* cv. Chunpoong (KM088019.1, KM036295.1)

Designing primers from conventional universal barcoding primer with reduced amplicon size

Besides newly designed 22 primer pairs, additional eight primers were picked from *matK* and *rbcL*, where recommended combination target for universal plant barcoding from CBOL group¹. I divided their amplicon region into half to amplify shortened products less than 500bp, proper to further NGS analysis, and used the one side of conventional primer as it (Figure 2-2). In detail, the forward primer of M23 and M24 was the same as conventional *matK* barcoding forward primer and the reverse primer is the newly designed primer in this research. Amplicon region of this primer was overlapped with upstream region of conventional *matK* barcoding region, the size of which was from 399bp to 428bp in the 23 chloroplast genomes. From *rbcL* region, three primer pairs M29, M30, M31 were designed. The amplicon of M29 was overlapped with upstream region of conventional *rbcL* barcoding region and those of M30 and M31 were matched with downstream region of which

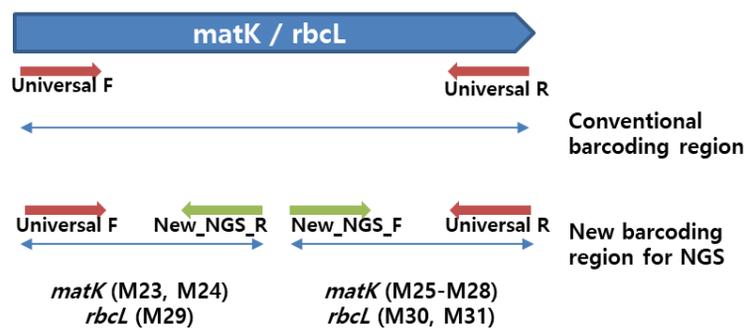


Figure 2-2. Schematic representation of primer designing strategy from two universal barcoding regions, *matK* and *rbcL*. Conventional primers were symbolized with red arrows and newly designed primers in this study were indicated by the green arrows.

Phylogenetic relationship for evaluating sequence diversity

For evaluating the discrimination resolution, phylogenetic relationships of primer pairs were constructed based on sequence variations of amplicons. From the tree morphology, the resolutions of each marker were represented more clearly (**Figure 2-3**). The pair M3 was derived from intergenic region of *trnQ* and *psbK*, the pair M5 from intergenic region of *trnR* and *atpA*, the pair M10 from intergenic region of *trnL-UAA* and *trnF-GAA*, the pair M9 from intergenic region of *psbJ* and *psbF*, the pair M13 from *psbB*, the pair M14 from intergenic region of *psbB* and *psbH*, the pair M15 from intergenic region of *trnI-CAU* and *ycf2*, the pair M17 from intergenic region of *rps12* and *trnV-GAC*, and the pair M18 from intergenic region of *trnN-GUU* and *ycf1*. Those primer pairs above showed family level distinguishable resolution which was represented by four clades in each of phylogenetic relationship, corresponding to *Panax* (Araliaceae), *Angelica* and *Peucedanum* (Apiaceae), *Papaver* (Papaveraceae), *Cannabis* (Cannabaceae) and genus level resolution between *Angelica* and *Peucedanum*, *Cannabis* and *Humulus*. The primer pair M2 from intergenic region of *psbA* and *trnK-UUU* and the pair M6 from CDS of *psbD* were also showed family level resolution but not distinguishable between *Angelica* and *Peucedanum*. The primer pairs M1 and M16 were picked from *psbA* and *ycf2* region respectively. Their amplicons were divided into four clade corresponding to family level and showed genus level resolution in Apiaceae and species level resolution in *Panax* but no polymorphism between *Cannabis* and *Humulus*.

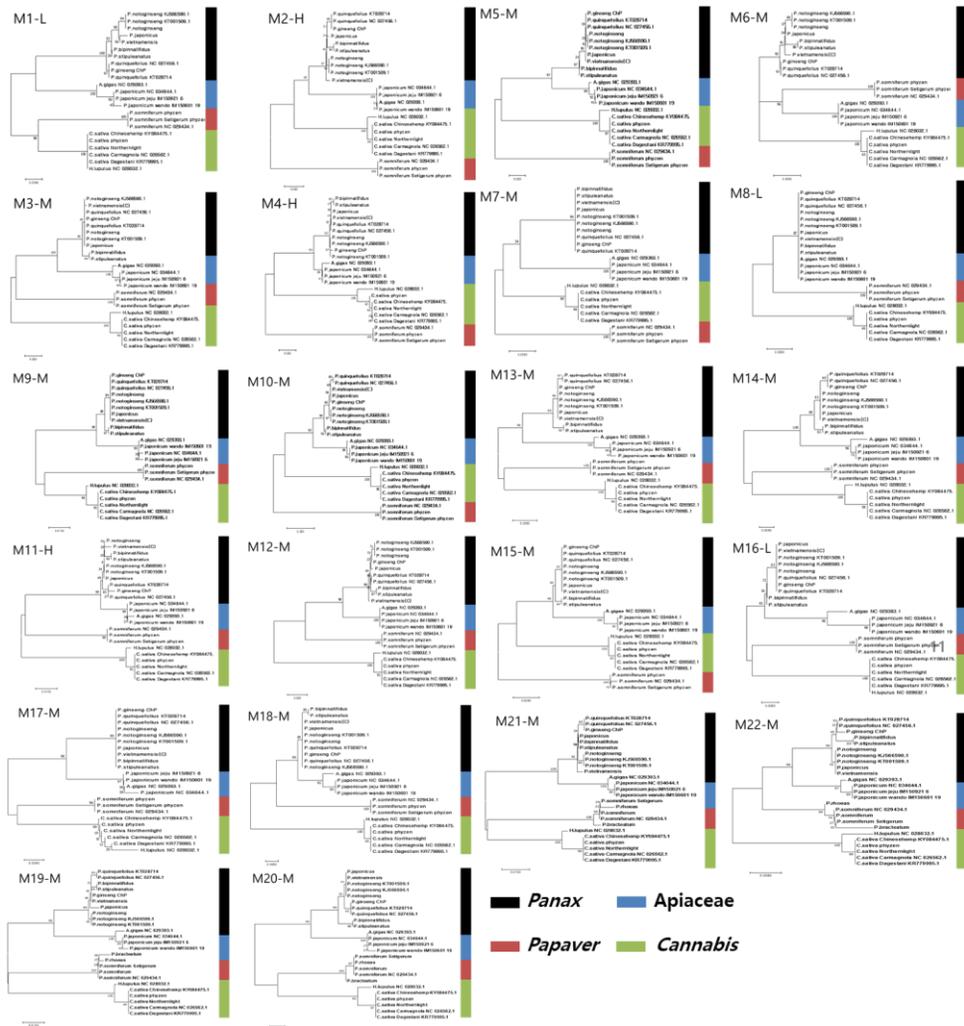


Figure 2-3. Phylogenetic relationship of amplicon from each primer pairs. The trees of each primer pairs were constructed after multiple alignments of each amplicons. The colored bar on the right side of tree indicates each of taxonomical group.

There were primers showed broad level of resolution like M7 with all identical sequences within *Panax* or M8 with monophyletic group of Apiales species together, although both of them showed genus level resolution in Cannabaceae. Based on these resolution of each primer pairs, I divided them in to three group based on the extent of resolution range, H for high diversity even at species level, M for middle level diversity, and L for

low diversity like order level (**Table 2-1**).

Overlapping of amplicon with mitochondrial plastid DNA (MTPT)

Total 22 newly picked chloroplast markers were distributed to entire chloroplast genome region (**Figure 2-1**). Most of their amplicons were amplified from intergenic region where showed low frequency of mitochondrial transfer from the previous result (**Figure 1-4, Chapter 1**). In case of genic region derived primer pairs like M1, M16, M13, M19, the possible error from co-amplification of MTPT counterpart seemed to be low because the transfer frequency of their genes were around 10 times.

***In silico* PCR analysis for extended application of markers to overall plant lineage**

For further extended application of the primer pairs as universal barcoding markers, I checked the specificity of the primer pairs to 32 dicot (**Table 2-3**) and 11 monocot (**Table 2-4**) selected as representative for each order. When targeting dicot plants, newly designed primer pairs were successfully amplified from at least 25 species except M4 pair from 18 species (**Table 2-5**). Eight primer pairs from *matK* and *rbcL* were also showed positive amplification result more than 28 species except M25 pair of 21 species, M27 pair of 22 species and M30 pair of 12 species. However, specificity checking from monocot plants showed lower success rate with median value of 72% compared to 90% of dicot plants. Nine primer pairs showed successful *in silico* amplification result with 90% in dicot and 72% in monocot; M1, M2, M19, M7, M10, M23, M26, M28, M29.

Table 2-3. List of dicot order and representative species used for *in silico* PCR

Order	Species	Accession No.
Celastrales	<i>Euonymus schensianus</i>	NC_036019.1
Cucurbitales	<i>Cucumis sativus</i>	MF095790.1
Fabales	<i>Senna tora</i>	NC_030193.1
Fagales	<i>Quercus baronii</i>	KT963087.1
Malpighiales	<i>Byrsonima crassifolia</i>	NC_037192.1
Oxalidales	<i>Averrhoa carambola</i>	KU569488.1
Rosales	<i>Pyrus pyrifolia</i>	NC_015996.1
Zygophyllales	<i>Larrea tridentata</i>	NC_028023.1
Brassicales	<i>Brassica nigra</i>	KT878383.1
Geraniales	<i>Pelargonium alternans</i>	NC_023261.1
Huerteales	<i>Tapiscia sinensis</i>	MF926267.1
Malvales	<i>Hibiscus syriacus</i>	KP688069.1
Myrtales	<i>Plinia trunciflora</i>	NC_034801.1
Sapindales	<i>Xanthoceras sorbifolium</i>	KY779850.1
Buxales	<i>Buxus microphylla</i>	EF380351.1
Proteales	<i>Platanus occidentalis</i>	DQ923116.1
Ranunculales	<i>Aconitum longecassidatum</i>	NC_035894.1
Trochodendrales	<i>Trochodendron aralioides</i>	KC608753.1
Caryophyllales	<i>Colobanthus quitensis</i>	NC_028080.1
Santalales	<i>Erythralum scandens</i>	NC_036759.1
Saxifragales	<i>Sinowilsonia henryi</i>	NC_036069.1
Apiales	<i>Glehnia littoralis</i>	KT153022.1
Aquifoliales	<i>Ilex latifolia voucher</i>	KX426465.1
Asterales	<i>Artemisia frigida</i>	NC_020607.1
Dipsacales	<i>Kolkwitzia amabilis</i>	NC_029874.1
Cornales	<i>Camptotheca acuminata</i>	KY511612.1
Ericales	<i>Diospyros lotus</i>	NC_030786.1
Garryales	<i>Eucommia ulmoides</i>	KU204775.1
Gentianales	<i>Cynanchum wilfordii</i>	NC_029459.1
Icacinales	<i>Iodes cirrhosa</i>	NC_036304.1
Lamiales	<i>Mentha spicata</i>	NC_037247.1
Solanales	<i>Capsicum frutescens</i>	KR078312.1

Table 2-4. List of monocot order and representative species used for *in silico* PCR

Order	Species	Accession No.
Acorales	<i>Acorus calamus</i>	AJ879453.1
Alismatales	<i>Epipremnum aureum</i>	KR872391.2
Asparagales	<i>Asparagus officinalis</i>	NC_034777.1
Arecales	<i>Cocos nucifera</i>	KX028884.1
Commelinales	<i>Hanguana malayana</i>	NC_029962.1
Poales	<i>Oryza sativa</i>	NC_031333.1
Zingiberales	<i>Musa balbisiana</i>	NC_028439.1
Dioscoreales	<i>Dioscorea elephantipes</i>	EF380353.1
Liliales	<i>Lilium lancifolium</i>	NC_035589.1
Pandanales	<i>Carludovica palmata</i>	NC_026786.1
Petrosaviales	<i>Japonolirion osense</i>	NC_036154.1

Table 2-5. Summary of *in silico* PCR success rate from order level of monocot and dicot and *in vitro* PCR result

Primer name	<i>in silico</i> PCR		<i>in vitro</i> PCR**	
	Dicot order (32)	Monocot order (11)	Dicot plants	Monocot plants
M1	32	10	S	S
*M2	32	11	S	S
M3	29	5	M	S
M4	18	11	M	P
M5	30	7	S	S
M6	30	3	S	S
M7	30	9	S	S
M9	32	5	S	P
M10	30	11	M	M
M12	25	9	M	S
M13	27	7	M	S
M15	32	8	S	M
M18	29	10	S	S
*M19	29	10	S	S
M20	31	10	S	P
*M21	30	6	S	S
M22	30	4	S	S
M23	28	9	S	P
M24	30	2	S	P
M25	21	4	S	P
M26	28	9	S	P
M27	22	6	S	P
M28	29	9	S	P
*M29	29	11	S	S
M30	12	10	S	P
M31	29	4	M	P

* Selected primers for further NGS application

**S, single band from all species; M, partial multiple band amplification, P, no PCR product from several samples.

PCR validation of primers from dicot and monocot plant DNA

I chose 17 newly designed primers and 9 conventional region derived ones from the **Table 2-2** considering there amplicon size and diversity for PCR validation. Each primers were tested using total 31 plants DNA (**Table 2-6**) with 24 dicot and 7 monocot plants under three annealing thermal condition (54°C, 56°C, 58°C degree). All of the 17 primer pairs were successfully amplified the product from the dicot samples under three different thermal

conditions. However, from primer pair M4 in Apiaceae and M3, M10, M12, M13 in *P. somniferum*, unexpected multiple bands were amplified which might reflect interspecies level variation of those plants. From monocot samples, 15 primer pairs were successfully amplified the product but M4, M9, M15, M18, M20 pairs were failed (**Figure 2-4 and 2-5**).

In case of nine primers from *matK* and *rbcL*, they were also well amplified from dicot plants (**Figure 2-6**) under annealing temperature of 54 and 56 degree, so optimal annealing temperature of all primers for amplifying dicot plants was set to 56 degree. However, only M29 pair showed all positive result from monocot samples and the other 8 pairs were not well amplified from monocot plants.

Table 2-6. List of plant DNA used for PCR Test

#	Scientific name	Sample code	#	Scientific name	Sample code
S1	<i>Cannabis sativa</i>	16-CA-1	S17	<i>Panax notoginseng</i>	Pn
S2	<i>Cannabis sativa</i>	16-CA-1	S18	<i>Panax japonicas</i>	Pj
S3	<i>Cannabis indica</i>	16-CA-15	S19	<i>Panax vietnamensis</i>	Pv
S4	<i>Cannabis indica</i>	16-CA-16	S20	<i>Panax bipinnatifidus</i>	Pb
S5	<i>Humulus japonicus</i>	Hj_phar	S21	<i>Angelica gigas</i>	Ag
S6	<i>Papaver somniferum</i>	17-399-1	S22	<i>Peucedanum japonicum</i>	Pej1
S7	<i>Papaver somniferum</i>	16-133-4-1	S23	<i>Peucedanum japonicum</i>	Pej2
S8	<i>Papaver somniferum</i> subsp. <i>setigerum</i>	SPO-1J	S24	<i>Miscanthus sinensis</i>	M87
S9	<i>Papaver somniferum</i> subsp. <i>setigerum</i>	SPO-2J	S25	<i>Miscanthus sinensis</i>	M367
S10	<i>Papaver somniferum</i> subsp. <i>setigerum</i>	17-308-1	S26	<i>Oryza sativa</i>	416027
S11	<i>Papaver rhoeas</i>	16-nPA-5	S27	<i>Avena sativa</i>	As1
S12	<i>Papaver somniferum</i> subsp. <i>setigerum</i>	14-596-8	S28	<i>Asparagus officinale</i>	Ao1
S13	<i>Cannabis sativa</i>	Cs1	S29	<i>Allium cepa</i>	Ac1
S14	<i>Cannabis sativa</i> northern light	Cs_n1	S30	<i>Allium tuberosum</i>	At1
S15	<i>Panax ginseng</i>	Pg	S31	<i>Papaver orientale</i>	Po
S16	<i>Panax quinquefolius</i>	Pq			

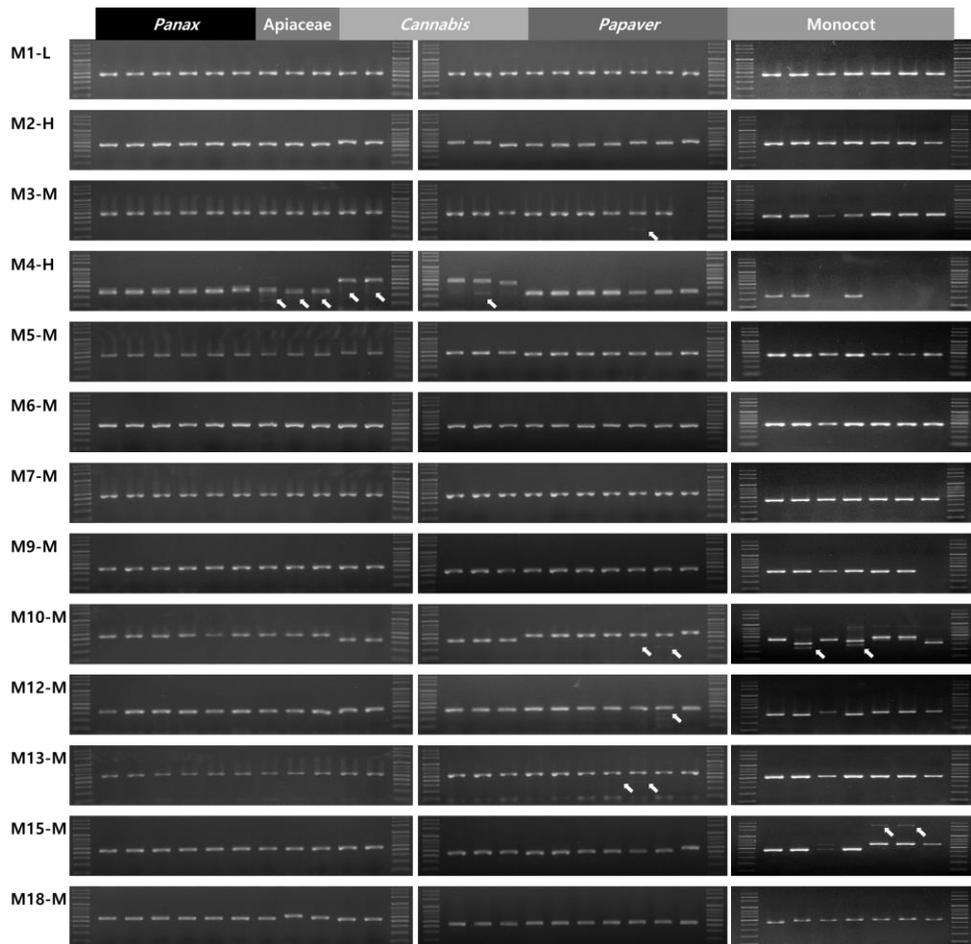


Figure 2-4. PCR amplification of chloroplast specific primers (M1 – M18) from dicot plants of four families. List of samples used are in Table 2-6

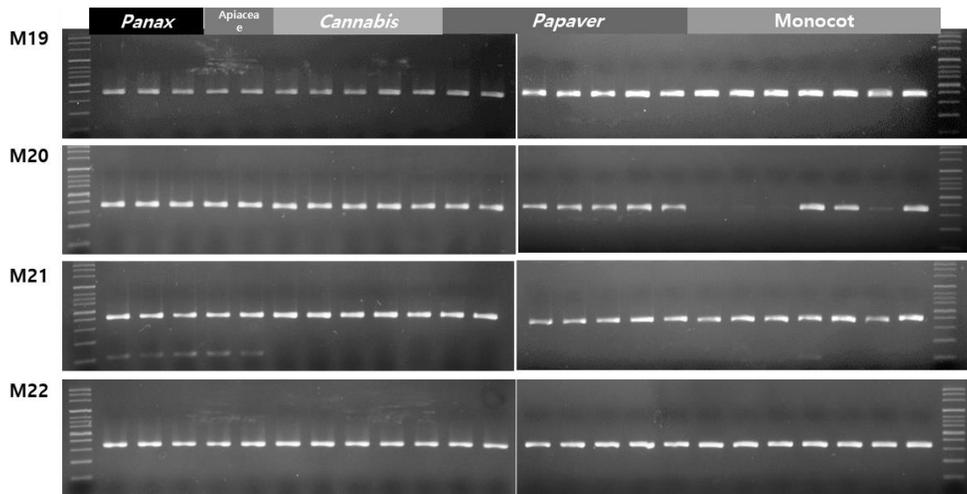


Figure 2-5. PCR amplification of chloroplast specific primers (M19 – M22). Both of dicot and monocot samples were used in **Table 2-6**.

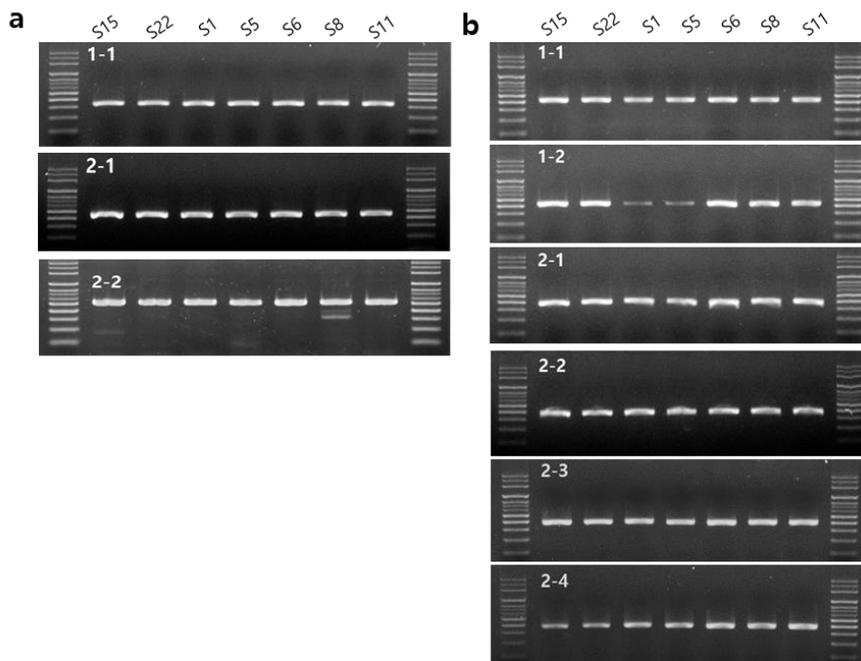


Figure 2-6. PCR amplification of nine universal primers (M23 – M31) (a) Three primers of *matK* (M23-M25) from 6 dicot samples (b) Six primers of *rbcL* (M26-M31) from 6 dicot samples.

NGS application of PCR products from mixed DNA template

From 31 primer pairs, three chloroplast specific primers, M2-H, M19-M, M21-M, and one universal barcoding primer, M23-L (*rbcL*), were selected for candidate of NGS application based on their species discrimination resolution and PCR amplification success in all of the sample tested (**Table 2-5 and 2-7**). M2-H was highly variable markers with species level discrimination. M19-M and M21-M was middle level variable markers with Genus level resolution. M23-L derived from *rbcL* gene was chose for using currently available database.

In the first NGS application, four chloroplast targeting markers above and two markers from nuclear genome region 18S rDNA¹⁵ and ITS2¹⁶ which were previously reported to common for animal genome and microorganism were used for constructing NGS libraries from the DNA mixture with equal amount of four plant species (**Table 2-7**). After quality trimming, NGS reads with the number from 53,866 to 334,394 were obtained from two replication of each of primers (**Table 2-8**). However, the M19-M amplicon of the second replicate failed to generate enough NGS read, so this dataset was not used for further analysis.

When the reads of each markers were mapped to the reference amplicon sequences in parallel, they were rightly located to their own amplicon sequence. There were mapped reads on 18S and ITS2 amplicon from plants even though they mainly targeted non-plant materials. In M2-H, all of the four species showed similar read depth ratio from 17% to 34% with a range of fluctuation from species to species (**Figure 2-7**). However, the number of reads from single species were different from marker to marker as in the case of *P. ginseng* with much low number (5%) in M19-M and M23-L and much high number in 18S or ITS2, or *O. sativa* with high number in M19-M and M23-L but almost no or less amplification from M21-M, 18S and ITS2. The results of multiplex PCR of five primer from

mixed DNA showed similar pattern with independent primer PCR results (Figure 2-8).

Table. 2-7. List of samples and Primers used for 1st NGS application

	Code	Sample name
Sample list	S15	<i>Panax ginseng</i>
	S6	<i>Papaver somniferum</i>
	S31	<i>Papaver orientale</i>
	S26	<i>Oryza sativa</i>
	Target region	Primer name
Primer	Chloroplast	M2
		M19
		M21
		M23 (<i>rbcL</i>)
	Nuclear	18S
		ITS2

Table 2-8. The 1st NGS result of two replication from each of markers

Marker	Replication 1		Replication 2	
	Total read count	Mapped read count (%)	Total read count	Mapped read count (%)
M2-H	101,492	71,594 (70.54%)	60,395	47,929 (79.36%)
M23-L	226,117	139,939 (61.89%)	334,494	250,186 (74.80%)
M19-M	118,485	87,809 (74.11%)	294	233 (79.25%)
M21-M	113,363	81,577 (71.96%)	109,865	71,083 (64.70%)
18S	150,581	56,828 (37.74%)	194,763	102,276 (52.51%)
ITS2	53,866	28,889 (53.63%)	54,279	35,093 (64.65%)

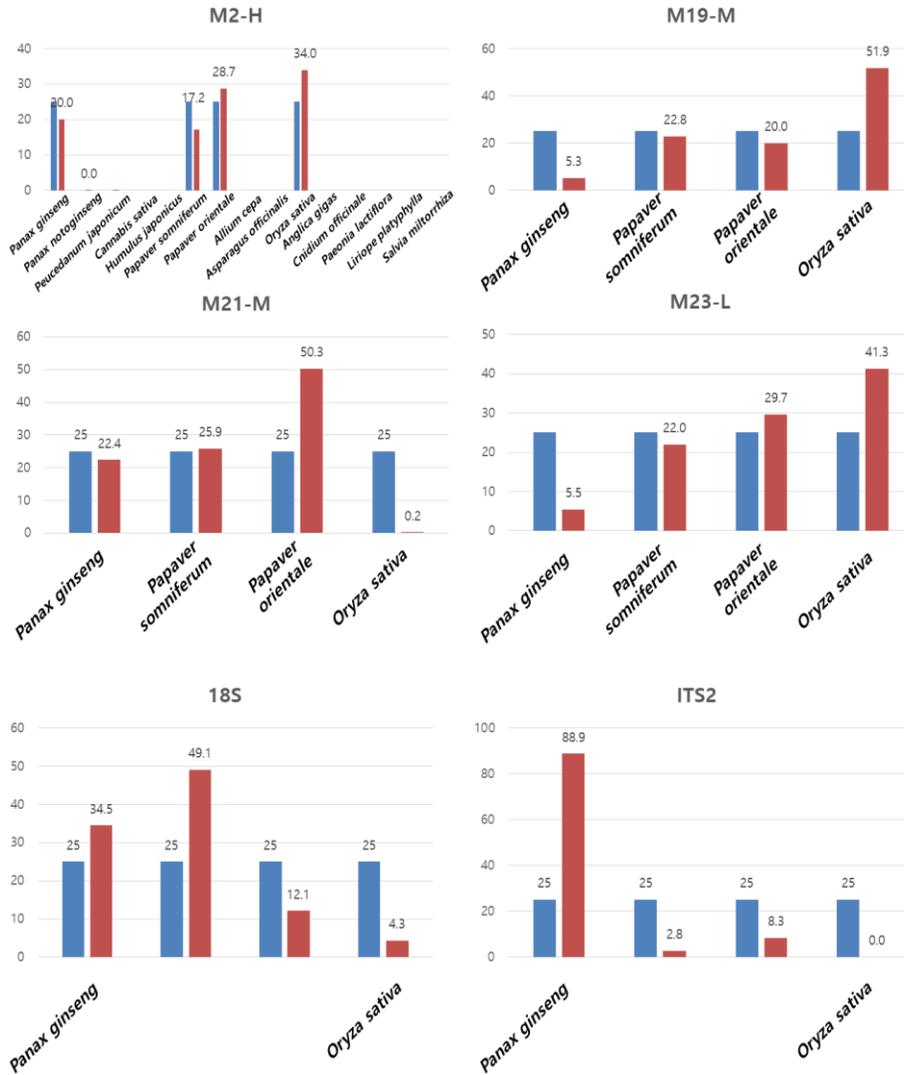


Figure 2-7. Relative NGS read depth of four plant species from each amplicon (1st NGS analysis)

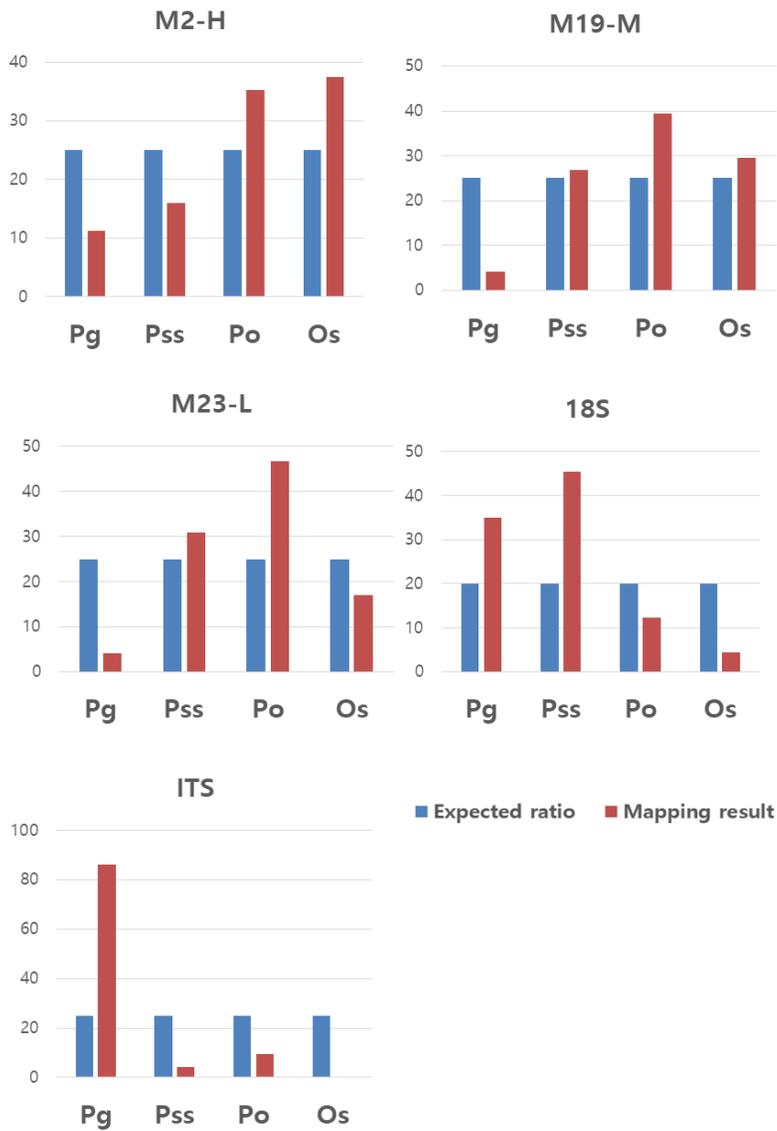


Figure 2-8. Relative NGS read depth of four plant species from each amplicon of multiplex PCR

The second NGS application was conducted with more enlarged species. Totally 11 plant species were mixed with same and different ratio at DNA level (**Table 2-9**). After quality trimming, NGS reads with the number from 93,764 to 221,586 were obtained from each of primers (**Table 2-10**). When the reads of each amplicons were mapped to their sequences, the reads depth of each species showed more biased pattern compared to the first NGS application result (**Figure 2-9**). Even in mixed DNA with same ratio, read depths of several species including *C. sativa*, *A. cepa*, *A. officinalis* were dominated from all of the four markers. In M21-M amplicon, no reads were mapped to *O. sativa* same with the result from the 1st NGS analysis indicating different PCR efficiency from species to species in mixed DNA. Additionally, in mixed DNA with different ratio, actual read depths per samples were different with expected values in several species. The read depth of *C. sativa* showed much more abundant read depth even though the input DNA from all of the three markers except M2-H. Likewise, *A. officinale* also showed dominated read abundance from three markers except M19-M. Therefore, Read depth analysis in these dataset could provide rough estimation of which species in mixture but quantitative portion of each species should be carefully considered. PCR efficiency of each species in mixed state were different marker to marker, so the relative portion of some species could be overestimated.

Table 2-9. List of samples and primer used for the second NGS application

	Code	Name	Ratio	
			Mixture1	Mixture2
Sample list	S6	<i>Papaver somniferum</i>	1	15
	S31	<i>Papaver orientale</i>	1	7
	S13	<i>Cannabis sativa</i>	1	5
	S15	<i>Panax ginseng</i>	1	30
	S16	<i>Panax quinquefolius</i>	1	7
	S17	<i>Panax notoginseng</i>	1	3
	S22	<i>Peucedanum japonicum</i>	1	3
	S26	<i>Oryza sativa</i>	1	9
	S28	<i>Asparagus officinale</i>	1	3
	S29	<i>Allium cepa</i>	1	1
	S31	<i>Humulus japonicus</i>	1	1
	Type	Primer name		
Primer	Single	M2		
		M19		
		M21		
		M23 (<i>rbcL</i>)		

Table 2-10. The second NGS results of each amplicon from mixed DNA with 11 species.

Marker	Same ratio		Different ratio	
	Total read count	Mapped read count (%)	Total read count	Mapped read count (%)
M2-H	93,764	72,354 (77.17%)	166,369	138,025 (82.96%)
M23-L	343,522	183,179 (53.32%)	289,076	169,146 (58.51%)
M19-M	253,638	193,428 (76.26%)	256,264	195,503 (76.29%)
M21-M	315,041	265,990 (84.43%)	221,586	140,471 (63.39%)

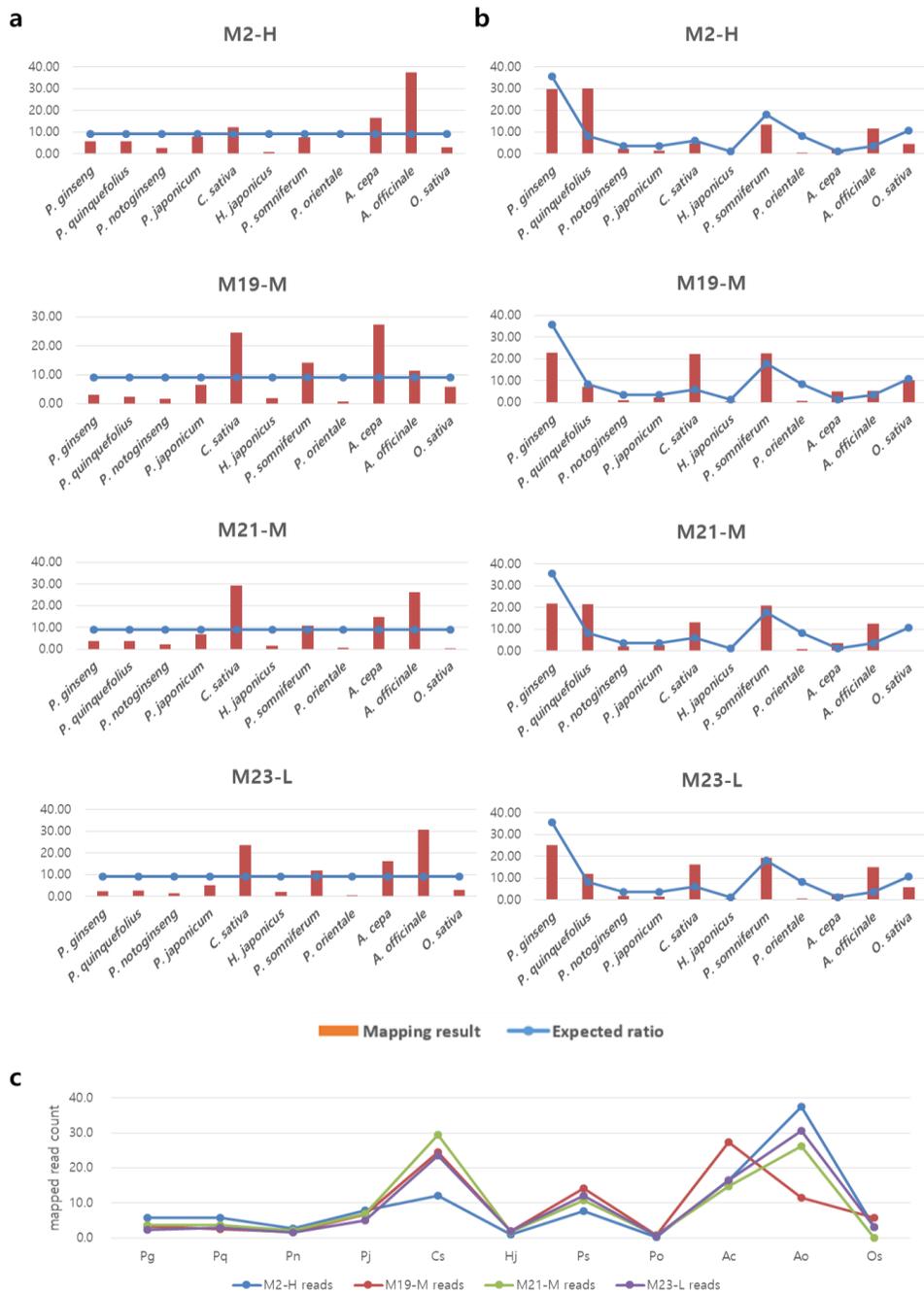


Figure 2-9. Relative read depth of 11 plant species from each amplicon (2nd NGS analysis) (a) DNA mixture with same ratio. (b) DNA mixture with different ratio. (c) Summary of read depth of each samples from DNA mixture with same ratio.

DISCUSSION

Metabarcoding of plants has been applied in diverse research field as diet research^{17,18}, Biodiversity^{4,19}, and even an trace of illegal trade^{20,21} mainly using single locus marker like *rbcL*, ITS, *trnL* were applied with diverse population. In most cases, the metabarcoding results provided cost-effective and relatively reliable identification for each purpose. But it wasn't until recently that metabarcoding of plants with multiple loci was conducted. Application of multiple markers in plant barcoding should be expanded as resolution and PCR reaction of each marker are variable like in this study and both organelle and nuclear genome need to be considered simultaneously to reflect overall aspect of plant species diversity.

Chloroplast genome is well known to be conserved across the plant lineage. To date, more than 9,000 chloroplast genomes of angiosperm have been registered in the GenBank. Therefore, enlarged comparative genomics of chloroplasts makes possible to discover conserved region across the far distant taxonomy such as order level like in this study. As the primers in this study derived from these highly conserved sequence and didn't have degenerative sequence, it was possible to amplify the PCR products with high intensity from all targeted samples. Considering that one of the limitations of conventional barcoding primers is different PCR efficiency species to species due to mismatch of primer with template^{1,22}, they have advantages of reducing one of the possible PCR bias. Because target regions of each primer were dispersed to overall chloroplast genome (**Figure 2-1**), these primers could be used for scanning of entire chloroplast genome diversity. Moreover, since their amplicons harbor abundant SNPs and InDels, they could be used to advantage in covering natural biodiversity.

From the actual PCR test with target species and monocot plants, those primers were successfully amplified with a few exceptions like M4 pairs showing unexpected multiple bands. It was also fortunate that *in silico*

primer specificity check to representative species presented high success rate in dicot orders and moderate rate in monocot orders. Thus, these primer sets are expected to cover wide range of plant kingdom.

Two primers derived from nuclear regions, 18s and ITS2, were succeeded in amplifying the NGS read from plant samples. As these primers originally intended to amplify non-plant materials, using them have beneficial to identify both of plant and non-plant material at one time. The rDNA regions have several advantage as they are multi-copies, found in both of plant and non-plant organisms and reflecting actual genome level diversity. The ITS2 region is the most commonly used barcoding region for plants since 2010²³. However, 18s rDNA are not sufficiently diverse at interspecies level in several taxa²⁴ and heterogeneous copies within an individual have been found both of 18s rDNA and ITS2 region because of ploidy level of genome²⁵ or imperfect synchronization by the concerted evolution^{26,27} or.

Co-amplification of mitochondrial plastid DNA (MTPT) has been underestimated in previous plant barcoding. However, it could be a serious noise in taxonomical analysis with highly sensitive methods such as NGS²⁸. In this study, most of primers were designed from intergenic regions to exclude MTPT amplification as much as possible because those fragments were usually derived from genic region of chloroplast discussed in Chapter 1. Nevertheless, the *rbcL*, the most frequently transferred MTPT region was included to utilize the current barcoding database and only limited number of mitochondrial genome sequences were available compared to the number of those from chloroplast, so interpreting the NGS data should be carefully done about the MTPT.

NGS application of four selected primers with two additional nuclear primer in this study suggest possible rough species estimation from mixed DNA. Overall read depth trend of each species was almost similar among primers (**Figure 2-9c**) and multiplex PCR using five primer mixture

showed similar read mapping depth for each of amplicon with independent PCR (**Figure 2-8**). Multiplexing of primers has advantage when the quantity of samples is too small to get enough DNA for PCR like forensic analysis and reducing the cost for sequencing with large number of samples. However, if the mixed primers compete with each other, no amplification from one primer could be observed. In this study when the M19-M and M21-M was mixed, no PCR product was amplified from M21-M. In some cases, additional NGS data polishing steps is needed as the more increased number of chimera reads is generated in multiplex PCR condition²⁹

Not only primers, bias among samples were observed in both of two times of NGS application. No mapped reads were noticed from *O. sativa* in M21-M and overestimated read depth were observed from *C. sativa*, *A. cepa*, and *A. officinale* (**Figure 2-9**). The copy number of chloroplast and ribosomal DNA could be variable among species or samples in the same amount of total DNA. Because the genome size of each plant species is different, the smaller the genome size is, the more copies of chloroplast and ribosomal DNA exist. Additionally, the quantity of chloroplast and total length of nuclear ribosomal DNA are also different among species and for chloroplast, even from tissue types³⁰⁻³². The estimated copy number of chloroplast genome is >1,000 copies in a plant cell as 1,241 copies in *C. wilfordii* from the chapter 1, but it could be variable depending on developmental stage or condition of leaves like in *A. thaliana* less than 200 copies in senescent leaves or dark treatment³⁰. PCR efficiency also cause the mapping bias, resulting from primer binding site competition, GC contents of internal amplicon sequence, or DNA quality of each samples. Therefore, alternative methods reducing or skipping the PCR step are needed to minimize the PCR based bias as much as possible for NGS library construction of metabarcoding.

Materials and Methods

Selection of barcoding candidate regions

I used 23 plant chloroplast genome for discovering candidate universal barcoding regions including ten of *Panax* species, five of Apiaceae species, five of *Papaver* species, five of *Cannabis* species, and one *Humulus* species (**Table 1**). Most of the sequences were retrieved from Genbank except one *Cannabis sativa*, one *Papaver somniferum*, and *P. somniferum* var. *setigerum*, *P. bracteatum*, *P. rhoeas* which were provided from Supreme Prosecutor's Office (Republic of Korea).

Discovering conserved region in chloroplast genome

Total 23 chloroplast sequences were conducted to multiple alignment using MAFFT³³ with default parameter. Consensus nucleotide sequence was generated and polymorphic regions of InDel and SNP, represented to IUPAC codes or dash character, were masked with "N" not to remain any non-nucleotide characters. This masked consensus sequence was further used for designing primers by Primer-BLAST³⁴ with maximum amplicon size of 500 bp. Then I selected primer pairs satisfying two criteria simultaneously, one is that only a single region should be amplified from each species and the other is that actual amplicon size from each species should be between 400bp to 500bp considering average read length of Illumina Miseq platform, 300bp for each direction.

Sequence diversity of amplicon and phylogenetic analysis

Sequence diversity of amplicon of each primers were evaluated by counting the number of polymorphic site specific for each species, genus and family level. Phylogenetic relationships were also constructed using MEGA 7³⁵ with a neighbor-joining model 1000 bootstrap replications. Considering the number of polymorphic site and phylogenetic relationship, the primers were divided into three groups, high (H), middle (M), and low (L), according to

their amplicon diversity resolution.

Designing of primers from conventional barcoding region

Among conventional universal plant barcoding primers, *matK* and *rbcL*¹ were selected for compatibility with existed international sequence database like Genbank. The conventional forward primer of *matK* was used as it and only reverse primers were newly designed for each of them to reduce the amplicon size to less than 500bp. Similarly, new forward primers for conventional reverse primer of *matK* were designed. For *rbcL* region, the primers were newly picked, not conventional primers, but for their amplicon regions to be overlapped with widely used barcoding position as much as possible.

Validation of primers by Polymerase chain reaction (PCR)

Selected primers considering sequence diversity were proceeded to polymerase chain reaction (PCR). Genomic DNA of each samples were extracted using Qiagen DNeasy Mini Kit following manufacture's instruction and the concentration of which were calculated by spectrophotometer (ND-1000, Thermo Fisher Scientific). PCR was conducted in mixture with a total volume of 25 ul consist of 10 ng of template DNA, 1x PCR reaction buffer (Inclone, South Korea), 0.2mM of each dNTP (Inclone, South Korea), 0.2pmol of each primer (Bioneer, South Korea) and 0.4 units of Taq DNA polymerase (Inclone, South Korea) under 35 thermal cycles of denaturation at 95°C for 30 sec, annealing at 54°C to 58°C for 30 sec, and elongation at 72°C for 30 sec followed by a final elongation at 72°C for 5 min. The amplified PCR products were visualized by electrophoresis on 2% concentration of agarose gel or 9% of polyacrylamide gel under 100V and 100mA conditions for 30 minutes.

NGS application of individual PCR product

From 31 primer pairs, three chloroplast specific primers, M2, M19, M21, and one universal barcoding primer, M23 (*rbcL*), were selected for candidate of 1st NGS application based on their species discrimination resolution and PCR amplification success in all of the sample tested (**Table 2-7**). Equal amount of DNA mixture from four species were used for PCR template. NGS libraries were generated through two step of PCR amplification (**Figure 2-10**). The first round of PCR reaction was conducted with barcoding primers which were combined with Illumina multiplexing sequence following the protocol for sequencing of 16S rDNA from individual samples. After then, each PCR products from same primer were mixed with equal volume and used for second PCR for tagging of Illumina NGS adapter sequence. NGS running of Miseq was conducted following manufacture's protocol. The raw reads were demultiplexed and trimmed with clc quality trimmer with parameter of minimum length of 200bp and quality score more than 35. For calculation of read depth for each species, trimmed reads were mapped to the amplicon sequences derived from each chloroplast genomes using the clc mapper (ver, 4.21.104315) with parameter of both length fraction and similarity of 0.98.

With increased samples and primer combinations, additional second NGS application was conducted. Total 11 samples were mixed with both equal ratio (Mixture1) and different ratio (Mixture2) (**Table 2-8**). The experimental steps were proceeded equally with 1st NGS application from the library construction to calculation of read depth based on mapping.

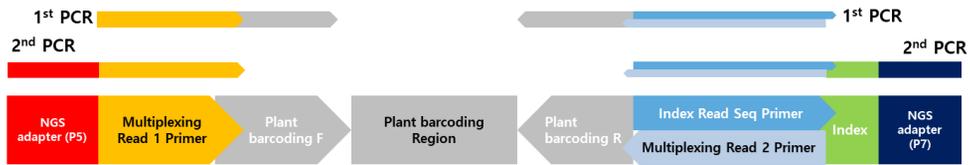


Figure 2-10. Diagram of metabarcoding primers for constructing NGS library based on two step PCR. The gray box indicated plant barcoding sequences designed from this study.

Table 2-11. Sequence of Primers used in this study

Name	Location	Forward	Reverse
M1-L	<i>psbA</i>	CCGCCGAATACACCAGCTA	TTGATGGTATTTCGTGAACCTGT
M2-H	<i>psbA - trnK-UUU</i>	GAAACAGGTTACGAAATACCATC	ATCCGACTAGTTCGGGTT
M3-M	<i>trnQ - psbK</i>	CCAAAACCCGTTGCCTTACC	CCTCGGGCAATTGGCG
M4-H	<i>trnG - trnR</i> intergenic	CTTCCAAGCTAACGATGCGG	TTGCGTCCAATAGGATTTGAACC
M5-M	<i>trnR - atpA</i>	GGTTCAAATCCTATTGGACGCA	TCAATTGGCAAGAGGTCAACG
M6-M	<i>psbD</i>	GGAGGGACCGTTTCGTTTTTG	TCGAAATATAGCTGCTACACCA
M7-M	<i>psbD - psbC</i>	TCCGTGCTTTTAACCCAACCTC	ATTCGCCGCCCAACG
M9-M	<i>psaA - ycf3</i> intergenic	AGATCTCTCCAAATCACTGGT	TCCGAACACTTGCCCCG
M10-M	<i>trnL-UAA - trnF-GAA</i>	GGGTTCAAAGTCCCTCTATCCC	TGAACTGGTGACACGAGGATT
M12-M	<i>rpl20 - clpP</i>	GAACGAGTCGCACATACACC	GTCAGCAACAGAAGCCCAAG
M13-M	<i>psbB</i>	CTGTCCATATAATGCATACAGCTC	CCGGAACAAAAGGATCAAACC
M15-M	<i>trnI-CAU - ycf2</i> intergenic	GGCGCTTTAACCATTAGCC	AATTCATAGGTATAGGAAGAAGCC
M18-M	<i>trnN-GUU - ycf1</i> intergenic	TCTACCCTGAGCTACTGAGGA	GCCCTATGGAGAATGTGGTCA
M19-M	<i>atpF</i>	CCTTGTAAAGGCTTGTGGAAAAC	TGAGCGTGAGAGCCAAATGA
M20-M	<i>rpoC2 - rpoC1</i> intergenic	TTAGGATTCATTTCTTGTGCAAA	GAATATGGAGGTACTTATGGCAGA
M21-M	<i>rpoB</i>	CCCGGATACTCGGTTCAAATA	TATTGATGTGAGATGGATCCAGAA
M22-M	<i>atpB - rbcL</i> intergenic	AATTCATGTCGAGTAGACCTTGT	TTCTCCAGCAACGGGCTC
M23	Upstream of <i>matK</i>	CGTACAGTACTTTTGTGTTACGAG	TGTCATTTTACTGTGGTCTCA
M24	Upstream of <i>matK</i>	CGTACAGTACTTTTGTGTTACGAG	GGCAATGTCAATTTTACCTGTGG
M25	Downstream of <i>matK</i>	AAGCGAGAATTGATTTTCCTTGA	TACCCATCCAGTCCATCTG
M26	Downstream of <i>matK</i>	AAGCGAGAATTGATTTTCCTTGA	ACCCAGTCCATCTGGAAATCTT
M27	Downstream of <i>matK</i>	GAAGCGAGAATTGATTTTCCTTGA	ATACCCTACCCAGTCCATCTG
M28	Downstream of <i>matK</i>	AAGCGAGAATTGATTTTCCTTGA	ACCCAGTCCATCTGGAAATCT
M29	Upstream of <i>rbcL</i>	TGTACCCACAACAGAGACT	GTTAGTAACAGAACCTTCTTCAAA
M30	Downstream of <i>rbcL</i>	CTTTTTGAAGAAGGTTCTGTACT	ACTCCAATCTCTGGCAA
M31	Downstream of <i>rbcL</i>	GCCCCTTGTCTGGAGAAGA	GGAACTCCCAATCTCTGGC
18S	Coding region of 18S rDNA	GGTGGTGCATGGCCGTCTTAGTT	TACAAAGGGCAGGGACGTAAT
ITS	ITS2 region	GCATCGATGAAGAACGCAGC	TCCTCCGCTTATTGATATGC
	Universal adapter F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	
	Universal adapter R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	

REFERENCES

- 1 Group, C. P. W. *et al.* A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* **106**, 12794-12797 (2009).
- 2 Gao, T. *et al.* Identification of medicinal plants in the family Fabaceae using a potential DNA barcode ITS2. *Journal of ethnopharmacology* **130**, 116-121 (2010).
- 3 Hamad, I. *et al.* Metabarcoding analysis of eukaryotic microbiota in the gut of HIV-infected patients. *PloS one* **13**, e0191913 (2018).
- 4 Hibert, F. *et al.* Unveiling the diet of elusive rainforest herbivores in next generation sequencing era? The tapir as a case study. *PLoS One* **8**, e60799 (2013).
- 5 Vesterinen, E. J. *et al.* What you need is what you eat? Prey selection by the bat *Myotis daubentonii*. *Molecular ecology* **25**, 1581-1594 (2016).
- 6 Iorizzo, M. *et al.* A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nature genetics* **48**, 657 (2016).
- 7 Kim, N. H. *et al.* Genome and evolution of the shade-requiring medicinal herb *Panax ginseng*. *Plant biotechnology journal* (2018).
- 8 Van Bakel, H. *et al.* The draft genome and transcriptome of *Cannabis sativa*. *Genome biology* **12**, R102 (2011).
- 9 Guo, L. *et al.* The opium poppy genome and morphinan production. *Science* **362**, 343-347 (2018).
- 10 Hosokawa, K., Shibata, T., Nakamura, I. & Hishida, A. Discrimination among species of *Papaver* based on the plastid *rpl16* gene and the *rpl16-rpl14* spacer sequence. *Forensic science international* **139**, 195-199 (2004).
- 11 Zhou, J. *et al.* Complete chloroplast genomes of *Papaver rhoeas* and *Papaver orientale*: Molecular structures, comparative analysis, and

- phylogenetic analysis. *Molecules* **23**, 437 (2018).
- 12 Kohjyouma, M. *et al.* Intraspecific variation in *Cannabis sativa* L. based on intergenic spacer region of chloroplast DNA. *Biological and Pharmaceutical Bulletin* **23**, 727-730 (2000).
- 13 Kojoma, M., Seki, H., Yoshida, S. & Muranaka, T. DNA polymorphisms in the tetrahydrocannabinolic acid (THCA) synthase gene in “drug-type” and “fiber-type” *Cannabis sativa* L. *Forensic Science International* **159**, 132-140 (2006).
- 14 Kojoma, M., Iida, O., Makino, Y., Sekita, S. & Satake, M. DNA fingerprinting of *Cannabis sativa* using inter-simple sequence repeat (ISSR) amplification. *Planta Medica* **68**, 60-63 (2002).
- 15 Treonis, A. M. & Wall, D. H. Soil nematodes and desiccation survival in the extreme arid environment of the Antarctic Dry Valleys. *Integrative and Comparative Biology* **45**, 741-750 (2005).
- 16 Mullin, P. G., Harris, T. S. & Powers, T. O. Phylogenetic relationships of Nygolaimina and Dorylaimina (*Nematoda: Dorylaimida*) inferred from small subunit ribosomal DNA sequences. *Nematology* **7**, 59-79 (2005).
- 17 Kartzinel, T. R. *et al.* DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. *Proceedings of the National Academy of Sciences*, 201503283 (2015).
- 18 Erickson, D. L. *et al.* Reconstructing a herbivore’s diet using a novel *rbcL* DNA mini-barcode for plants. *AoB Plants* **9** (2017).
- 19 Drummond, A. J. *et al.* Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience* **4**, 46 (2015).
- 20 Arulandhu, A. J. *et al.* Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *GigaScience* **6**, 1-18 (2017).
- 21 De Vere, N. *et al.* Using DNA metabarcoding to investigate honey bee foraging reveals limited flower use despite high floral

- availability. *Scientific reports* **7**, 42838 (2017).
- 22 Li, X. *et al.* Plant DNA barcoding: from gene to genome. *Biological Reviews* **90**, 157-166 (2015).
- 23 Chen, S. *et al.* Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PloS one* **5**, e8613 (2010).
- 24 Pawlowski, J. *et al.* CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS biology* **10**, e1001419 (2012).
- 25 Boutte, J. *et al.* Haplotype detection from Next-Generation Sequencing in high-ploidy-level species: 45S rDNA gene copies in the hexaploid *Spartina maritima*. *G3: Genes, Genomes, Genetics* **6**, 29-40 (2016).
- 26 Kim, K. *et al.* Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. *Scientific reports* **5**, 15655 (2015).
- 27 Parveen, I., Gafner, S., Tehen, N., Murch, S. J. & Khan, I. A. DNA barcoding for the identification of botanicals in herbal medicine and dietary supplements: strengths and limitations. *Planta medica* **82**, 1225-1235 (2016).
- 28 Song, H., Buhay, J. E., Whiting, M. F. & Crandall, K. A. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the national academy of sciences* **105**, 13486-13491 (2008).
- 29 De Barba, M. *et al.* DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources* **14**, 306-323 (2014).
- 30 Takami, T. *et al.* Organelle DNA degradation contributes to the efficient use of phosphate in seed plants. *Nature plants*, 1 (2018).
- 31 Rauwolf, U., Golczyk, H., Greiner, S. & Herrmann, R. G. Variable

- amounts of DNA related to the size of chloroplasts III. Biochemical determinations of DNA amounts per organelle. *Molecular Genetics and Genomics* **283**, 35 (2010).
- 32 Rosato, M., Álvarez, I., Feliner, G. N. & Rosselló, J. A. High and uneven levels of 45S rDNA site-number variation across wild populations of a diploid plant genus (*Anacyclus*, Asteraceae). *PloS one* **12**, e0187131 (2017).
- 33 Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics* (2017).
- 34 Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC bioinformatics* **13**, 134 (2012).
- 35 Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular biology and evolution* **33**, 1870-1874 (2016).

ABSTRACT IN KOREAN

부정 원재료 혼입(Economically motivated adulteration, EMA)는 약용식물 산업에 있어서 큰 비중을 차지하는 문제로, 연간 100 억에서 150 억불의 비용이 이를 해결하기 위해 소비되고 있다. 약용식물 산업의 지속적이고 안정적인 성장을 위해서는 올바른 판별법에 기초한 신뢰할 수 있는 원재료가 공급되는 것이 중요하며, DNA 바코딩은 이를 위한 효과적인 해결책으로 각광받고 있다. 그러나 대부분의 약용 식물들은 순화나 육종을 거치지 않고 야생에서 채집되어 유통되거나 재배되고 있어 넓은 범위의 생물 다양성을 그대로 가지고 있는 경우가 많다. 또한 미토콘드리아, 엽록체에서 일어나는 양친 유전이나, 핵 계놈을 포함하는 세 유전체 간의 수평 유전체 교환 (Horizontal genome transfer)등과 같이 식물 유전체가 가지고 있는 복잡한 특성들이 미치는 영향들은 과소평가 되고 있다. 그러나 이러한 특징들을 고려하지 않고 무분별하게 DNA 바코딩을 적용할 경우 중 판별에 있어서 더 큰 혼란을 초래할 수 있으며 산업 전반을 침체시킬 수 있다. 실제로 2015 년 미국에서는 DNA 바코딩 결과를 토대로 뉴욕 검찰이 주요 건강 기능성 식품 회사에 판매중지 가처분 신청을 내렸으나 곧 번복하였으며, 한국에서도 백수오를 원료로 하는 건강기능성 제품에 근연종 이엽우피소가 혼입되었다는 보도로 인하여 큰 혼란을 겪은 적이 있다.

첫 번째 장에서는 앞서 언급한 백수오와 이엽우피소의 엽록체와 미토콘드리아 비교 유전체 연구와 올바른 마커 적용에 대한 제언에 대해 다루고자 한다. 먼저 차세대 유전체 분석 기술을 이용하여 백수오와 이엽우피소의 엽록체 및 미토콘드리아 완전장을 해독하였으며 이 두 유전체를 비교한 결과 엽록체의 35%에 해당하는 서열이 미토콘드리아에서 상동성을 가지고 있는 것을 알 수 있었다. 또한 백수오와 이엽우피소의 미토콘드리아에서 발견되는 엽록체 조각(Mitochondrial Plastid DNA, MTPT)들은 두 종간에 매우 유사하였으며, 현재 엽록체 서열들과의 유사성을 토대로 이들의 삽입

시기를 추정해본 결과, MTPT 서열들이 백수오와 이엽우피소가 분화되기 이전의 공통 조상에서부터 삽입 되었으며 미토콘드리아의 매우 느린 진화 속도로 인하여 현재까지도 매우 잘 보존되어 있는 것으로 추정되었다. 또한 81 개의 속씨 식물 미토콘드리아 게놈을 조사하여 MTPT 가 이들간에 매우 다양하게 존재하지만 대체적으로 분류군 별로 특이적인 패턴을 보여주는 것을 알 수 있었으며, 이들이 식물 미토콘드리아 유전체의 복잡성에 기여했을 것이라 추정하였다. 또한 실제 백수오와 이엽우피소에서 개발된 마커를 토대로 MTPT 와 엽록체의 PCR 기반 동시 증폭이 종 판별에 있어 역설적인 오류를 줄 수 있음을 밝혔으며, 이를 근거로 분자 마커를 이용한 종 판별을 수행할 때의 올바른 적용 원칙을 제시하고자 하였다.

두 번째 장에서는 식물에서 불특정 혼합물에서의 차세대 유전체 분석 기술 기반 메타바코딩을 위한 엽록체 유전체 마커를 개발하고자 하였다. 기존에 구축되어 있는 국제 생물 정보 데이터 베이스의 정보를 적극적으로 이용하기 위하여 현재 식물 종 판별을 위해 널리 사용되고 있는 *matK* 와 *rbcL* 유전자 지역으로부터 NGS 에 적합한 크기의 PCR 산물을 만들어 낼 수 있는 마커를 디자인하였다. 또한 주요 약용식물 및 마약성 식물들도 보다 효과적으로 판별할 수 있도록 두릅나무과, 미나리과, 양귀비과, 삼과를 포함한 4 개 과로부터 23 개의 엽록체 유전체 정보를 확보하고, 비교유전체 분석을 이들 4 개과의 엽록체 서열과 완벽하게 일치하면서도 다양성을 보이는 다수의 마커를 개발하였다. 이들은 현재 NCBI 에 등록되어 있는 전체 식물군을 대상으로 진행한 가상 환경에서의 중합효소연쇄반응 (*in silico* PCR)에서도 쌍떡잎 식물에서 80%, 외떡잎 식물에서 70% 정도의 성공률을 보여주었으며, 실제 DNA 를 이용한 PCR 실험에서도 성공적인 결과를 얻을 수 있었기 때문에 새로운 범식물마커로 이용할 수 있을 것으로 기대된다. 4 개의 엽록체 기반 마커와 2 개의 핵 게놈 기반 마커 및 식물 혼합물 DNA 를 이용한 두 번의 NGS 적용 실험에서도, 혼합물에 어떤 시료가 들어있는 지를 보여주는 정성분석은 가능한

것으로 판단되었다. 다만, 정량분석의 경우 혼합 DNA 에서 각 시료별 PCR 효율이 마커마다 차이가 있어 특정 종이 우점하는 사례가 발견되었기 때문에 보다 신중하게 해석해야 하며, PCR 에 의한 오류를 줄이기 위한 추가적인 방법이 더 적용될 필요성이 있을 것으로 생각된다.