A DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Utilization of low-coverage whole genome sequences for diversity analysis in plant genomes

BY

JUNKI LEE

AUGUST 2017

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

COLLEGE OF AGRICULTURAL AND LIFE SCIENCES

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Utilization of low-coverage whole genome sequences for diversity analysis in plant genomes

JUNKI LEE

Department of Plant Science
The Graduate School of Seoul National University

GENERAL ABSTRACT

Recently, a rapid progress was achieved for whole-genome shotgun sequencing (WGS) based on support of next generation sequencing (NGS) technology. Although NGS greatly enhanced WGS, obtaining the complete reference genome sequence is a big challenge and is usually geared towards model plants or major crops based on massive WGS data and a variety of supporting data. Meanwhile, genomes of most resource plants are still unrevealed. In this study, a multi-directional approach was taken to understand the genome of resource plants, which have no previous genomic data, with only low coverage WGS data. This research was conducted to obtain a variety of fundamental information and nucleotide diversity on the inter- or intra-species level: assembly of complete chloroplast (cp) genome and nuclear ribosomal DNA (nrDNA) sequences, development of a tool to

detect polymorphic single sequence repeat (pSSR) markers, and quantification of major repeats among relative species in various plant genomes.

In the first chapter, complete cp genome and nrDNA sequences for two genotypes of *Euonymus hamiltonianus*, a medicinal and ornamental plant, were obtained. Furthermore, a pipeline to identify WGS reads harboring simple sequence repeat (SSR) motifs and develop pSSR by systematic comparison of two WGS reads was developed. The pipeline is composed of several steps which include end joining of paired reads, isolation of WGS reads harboring SSR motifs, identification of SSR reads derived from unique non-repetitive regions, identification of pSSR via comparison of counterpart WGS reads derived from another individual plant, design of pSSR primer sets and validation. Phylogenetic analysis using assembled and complete 157,360 bp of cp genome and 5,824 bp of 45S nrDNA was conducted. A total of 161 pSSR contigs showing polymorphism were identified between the two different *E. hamilotnianus* genotypes. Among them, 20 primer pairs were designed and seven were validated as real pSSR markers.

In the second chapter, the pipeline for pSSR was applied to *Peucedanum japonicum* which is an indigenous medicinal and edible plant in Korea. A total of 452 pSSR candidates were identified between two *P. japonicum.* Among them, ten primer pairs were designed, nine of which were validated as real pSSR markers for seven *P. japonicum* genotypes.

In the third chapter, the genome proportion of the major repeats was measured using small amount of WGS data for five *Panax* species and a related species *Aralia elata*. The diploids *P. japonicus, P. vietnamensis,* and *P. notoginseng* and the tetraploids *P. ginseng* and *P.*

quinquefolius possess 2.0 to 4.9 Gb genome sizes for the haploid genome equivalent. About 39-52% of the genome is comprised by four long terminal repeat retrotransposon (LTR-RT) family members, *PgDel, PgTat, PgAthila,* and *PgTork*. In particular, *PgDel1* LTR-RT occupied 23-35% of the *Panax* genomes and directly impacted their genome size variation. The genome size difference between *P. quinquefolius* (4.9 Gb) and *P. ginseng* (3.6 Gb) is explained by a burst of 0.9 Gb of *PgDel* during environmental adaptation after migration from Asia to North America one million years ago. The genome proportion of *PgDel2* LTR-RT is 2.5% in tetraploids and approximately 5% in diploids. Fluorescence *in situ* hybridization analysis of *PgDel1* and *PgDel2* supported the *in silico* estimation across three *Panax* species, *P. notoginseng, P. ginseng* and *P. quinquenfolius*. Our data revealed the role of four major repeats, which occupy almost 50% of the genome proportion, for evolution of the *Panax* genus. These results suggest that the study of only small amount of WGS could lead to a better understanding of the genome of plants including non-model plants and minor crops.

Keywords: low genomic coverage whole genome sequence (WGS), *Euonymus hamiltonianus,* polymorphic simple sequence repeats, *Peucedanum japonicum, Panax* genus, long terminal repeat retrotransposon, major repeats.

Student number: 2012-30301

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

WGS	Whole genome sequence
NGS	Next generation sequencing
dnaLCW	*de novo* assembly using low coverage of WGS
cp	Chloroplast
nrDNA	Nuclear ribosomal DNA
SSR	Simple sequence repeat
PBS	Primer binding site
dpsLCW	detection of polymorphic SSR using low coverage WGS
LTR-RT	Long terminal repeat retrotransposon
TRs	Tandem repeats
TEs	Transposable elements
GP	Genomic proportion
R-GP	RepeatMasker-based genomic proportion
M-GP	Mapping-based genomic proportion
C-GP	Clustering-based genomic proportion
BAC	Bacterial artificial chromosome
FISH	Fluorescence *in situ* hybridization
CV	Coefficient of variation
MYA	Million years ago

GENERAL INTRODUCTION

Over the past decade, high-throughput parallel DNA sequencing popularly called next-generation sequencing (NGS) technologies have become widely accessible and available, decreasing the cost of DNA sequencing (Metzker 2010). The NGS technologies have been evolving rapidly including the robust development of protocols for generating sequencing libraries and building effective solutions to data analysis (Shendure *et al.* 2008). The comprehensive plant genome analysis by NGS has the potential to widen the scope in plant biological research. Beginning with the genome project of *Arabidopsis thaliana*, hundreds of plants have been examined and analyzed (Arabidopsis Genome 2000, Michael *et al.* 2015). However, due to financial and technological challenges, extensive genomic analysis for understanding non-model plants is still limited.

In many plant nucleus genomes, repetitive elements (REs) could make up a massive proportion. REs can be divided into two large families: dispersed repeats (DRs) and tandem repeats (TRs). DRs contain tRNA genes, paralogous genes, retro genes, and transposons which can be further divided into class I.1 long terminal repeat (LTR) retrotransposons, class I.2 non-LTR retrotransposons, and class II DNA transposons. TRs contain nrDNA, tandem paralogs, and satellite DNA which includes satellites, minisatellites, and microsatellites (Piégu *et al.* 2015, Richard *et al.* 2008). In addition, a plant cell generally has hundreds of chloroplasts (cp) genome copies and tens of mitochondria (mt) genome copies (Bendich 1987, KUROIWA *et al.* 1992). For this reason,

cp and mt could be considered as different kinds of REs. Repeats contribute to plant evolution, gene regulation, adaptation, and genome size variation (Feschotte *et al.* 2007, Oliver 2013, Volff 2006). Moreover, due to genome-specific variations, repeats are used as molecular markers to characterize individual genomes (Kalendar *et al.* 2006).

This study aims to develop extended approaches for the analysis of REs in plant genomes. I tried to develop mining of essential information with high-proportion genomic contents in a cell such as cp and other repeats using small amounts of NGS data. Three non-model plants were analyzed in this study. To identify genomic diversity and evolution, cp genome, 45S nuclear ribosomal DNA, and polymorphic simple sequence repeat sequences (pSSRs), were characterized by a simple and effective pipeline with small amounts of whole genome NGS data. To investigate the polymorphisms in other non-model plants, *Peucedanum japonicum* which is a valuable medicinal and edible plant, the pipeline for pSSRs was applied and validated. To explore the genomic and evolutionary roles of major repeats, the genomic composition of the repeats was estimated in the *Panax* genus, which has been used as traditional medicine for many centuries (Shishtar *et al.* 2014), using low coverage WGS data. Calculation of RepeatMasker-based genome proportion and cytogenetic analysis by fluorescence *in situ* hybridization were performed to compare species of the *Panax* genus and a related species *Aralia elata* (Smit *et al.* 2013-2015).

REFERENCES

Arabidopsis Genome I 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815.

Bendich AJ 1987. Why do chloroplasts and mitochondria contain so many copies of their genome? BioEssays 6: 279-282.

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41: 331-368.

Kalendar R, Schulman AH. 2006. IRAP and REMAP for retrotransposon-based genotyping and fingerprinting. Nat. Protoc. 1: 2478-2484.

Kuroiwa T, Fujie M, Kuroiwa H. 1992. Studies on the behavior of mitochondrial DNA. Journal of Cell Science 101: 483-493.

Metzker ML. 2010. Sequencing technologies the next generation. Nat. Rev. Genet. 11: 31-46.

Michael TP, VanBuren R. 2015. Progress, challenges and the future of crop genomes. Curr. Opin. Plant Biol. 24: 71-81.

Oliver KR, McComb JA, Greene WK. 2013. Transposable elements: powerful contributors to angiosperm evolution and diversity. Genome Biol. Evol. 5: 1886-1901.

Piégu B, Bire S, Arensburger P, Bigot Y. 2015. A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. Molecular phylogenetics and evolution 86: 90-109.

Richard GF, Kerrest A, Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol. Biol. Rev. 72: 686-727.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26: 1135-1145.

Shishtar E, Sievenpiper JL, Djedovic V, Cozma AI, Ha V, Jayalath VH, Jenkins DJ, Meija SB, de Souza RJ, Jovanovski E, Vuksan V. 2014. The effect of ginseng (the genus *Panax*) on glycemic control: a systematic review and meta-analysis of randomized controlled clinical trials. PLoS ONE 9: e107391.

Smit A, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays 28: 913-922.

CHAPTER I

Chloroplast genomes, nuclear ribosomal genes and polymorphic SSR markers derived from two whole genome sequences of *Euonymus hamiltonianus* genotypes

ABSTRACT

Euonymus hamiltonianus, belonging to the Celastraceae family, is a tree with value for ornamental and medicinal plants. WGS data of 950 Mb from two *E. hamiltonianus* genotypes based Illumina MiSeq platform was produced. Complete chloroplast (cp) genome sequence of 157,360 bp and complete 45S nuclear ribosomal DNA (nrDNA) transcription unit of 5,824 bp were obtained. Annotation of cp genome revealed 113 genes of 76,469 bp. The 45S nrDNA has protein coding regions of 5,370 bp. However, the cp genome and the 45S nrDNA sequences were identical for both genotypes. A sophisticate protocol in order to identify polymorphic simple sequence repeat (pSSR) was further developed based on *in silico* cross-comparison of homologous two WGS reads. The protocol named as detection of pSSR uning low coverage WGS (dpsLCW) covering several steps to filter real pSSRs among the WGS reads. Application of dpsLCW identified a total of 161 pSSRs using the WGS of two genotypes. PCR were conducted for 20 candidate targets and seven markers were validated as real pSSRs. In careful sequence comparison revealed that the false pSSRs were derived from paralog sequence pairs. The cp genomes, 45S nrDNA, and pSSRs will be valuable resources for *E. hamiltonianus* and the dpsLCW will be a valuable tool for identification of large volume of pSSRs for non-model plants.

Keywords: simple sequence repeats, whole genome sequence, *Euonymus hamiltonianus.*

INTRODUCITON

The high throughput productivity of next-generation sequencing (NGS) technology has been very useful in terms of understanding plant evolution and speciation (Varshney *et al.* 2009). The NGS technologies have been evolving rapidly including the robust development of protocols for generating sequencing libraries and building effective approaches to data analysis (Shendure *et al*. 2008). Recently, the pipeline with small amount of NGS data and named *de novo* assembly using low coverage whole genome sequences (dnaLCW) was developed (Kim *et al.* 2015) to obtain complete chloroplast (cp) genome and 45S nuclear ribosomal DNA (nrDNA) units.

A molecular marker is a particular DNA fragment which includes specific genetic information with differences in the genomic level (Agarwal *et al.* 2008). Molecular markers has been widely used as a valuable tool for assessing genetic variation and has highly improved the genetic analysis in crop plants (Varshney *et al.* 2005). Various genomic components of plant cells have been considered as targets of molecular markers. Among them, cp genome and 45S nrDNA are the key elements that are used to explain plant genetic diversity and evolution (Qiu *et al.* 1999, Soltis *et al.* 1999, Kim *et al.* 2015). The cp genomes are circular DNA molecules ranging from 120 to 217 kb in length with highly conserved structures and gene order. They are composed of large single copy (LSC), small single copy (SSC), and two copies of inverted repeats (IR). The nrDNA unit has a

fundamental genetic role of linking transcription to translation (Richard *et al.* 2008). The nrDNA unit contains the 28S large subunit, the 18S small subunit, the 5.8S gene, two internal transcribed spacers (ITS1 and ITS2), and large intergenic spacers (IGS) in the 45S nrDNA and 5S ribosomal RNA gene in the 5S rDNA (Long and Dawid 1980). They are generally high-copied and tandemly-repeated transcription units in the plant genome (Rogers and Bendich 1987).

Simple sequence repeats (SSRs, or called microsatellites) (Hiroe and Constance 1958, Jones *et al.* 2010) which are consisted of one to six or more nucleotide sequentially repetitive motifs in a head-to-tail structure (Kelkar *et al.* 2010) are also used as popular genetic markers. SSR based molecular markers are universally implemented for population genetic studies such as parentage analysis, fingerprinting, genetic structure analysis and genetic mapping (Mittal and Dubey 2009, Meng *et al.* 2014, Grover and Sharma 2016) due to copy number variations or polymorphic features with reproducibility. Recently, various softwares to find SSR motifs were developed such as MISA (http://pgrc.ipk-gatersleben.de/misa/), SSR Locator, and FullSSR (Rozen and Skaletsky 2000, da Maia *et al.* 2008, Abdelkrim *et al.* 2009, Metz *et al.* 2016). However, conventional SSR marker development and experimental authentication still require considerably long guided sequence information and a labor-intensive procedure to find polymorphisms (Ma *et al.* 2009).

Euonymus hamiltonianus, belonging to the Celastraceae family, is a tree that grows in Afghanistan, Russia, India, Nepal, Pakistan, Burma, Japan, China, and Korea

(https://npgsweb.ars-grin.gov). The bark of the dried stem has been used in traditional medicine to enhance blood circulation and treat cough, cramps, backache, and poisoning by lacquer (Sol 2010). *E. hamiltonianus* has been planted and cultivated as an ornamental tree because of autumn foliage. The berries also have attractive colors similar to those of the fall foliage. Despite its wide utilization, *E. hamiltonianus* has very limited genomic information.

In this study, cp genome and nrDNA sequences were obtained for two genotypes by dnaLCW method. A simple and effective protocol for the identification of polymorphic SSR motifs was also developed with small amounts of whole genome sequence (WGS) reads, named detection of polymorphic SSR (pSSR) using low coverage of WGS (dpsLCW), by *in silico* cross-comparison of two homolog sequences.

MATERIAL AND METHODS

Plant materials, genomic DNA extraction and NGS sequencing

Two wild genotypes of *E. hamiltonianus,* one with normal leaves (EH-n) and another with variegated leaves (EH-v), were sampled from Hantaek Botanical Garden (http://www.hantaek.co.kr, South Korea). The genomic DNA was extracted from leaves using a modified cetyltrimethylammonium bromide (CTAB) method (Allen *et al.* 2006). Among the genomic DNA of seven accessions, EH-n and EH-v were used to construct genomic libraries with insert sizes of about 500 bp, according to Illumina paired-end standard protocol (http://www.illumina.com) and sequenced using Illumina MiSeq genome analyzer at LabGenomics (www.labgenomics.co.kr).

Sequence assembly and phylogenetic analysis of cp genome and 45S nrDNA

Complete cp genomes and 45s nrDNA units were assembled by dnaLCW protocol using clc_novo_assemble (ver. 4.21.104315, CLC Inc, Aarhus, Denmark) and manual curation (Kim *et al.* 2015). The complete cp genome of *E. hamiltonianus* was aligned with ten related plants of complete cp genome using MAFFT 7 (Katoh *et al.* 2002). A phylogenetic tree was generated by maximum likelihood analysis using MEGA 7.0 with bootstrap values of 1,000 (Kumar *et al.* 2016).

Sequence preparation (quality control and paired end joining), identification of SSR motif and clustering

WGS reads of EH-n and EH-v were trimmed by trimmomatic (ver. 0.33) based on quality score and sequence length (set the minimum quality score: ≥20, read length: ≥70 bp) (Bolger 2014). Trimmed PE WGS reads of EH-n were assembled by clc_overlap_reads (ver. 4.21.104315, CLC Inc, Aarhus, Denmark) with a minimum overlapping length of 20 bp and 95% similarity (applying '-o 20 –s .95'). SSR motifs in jointed and non-jointed (≥250 bp) contigs were identified using microsatellite search module (MISA: http://pgrc.ipk-gatersleben.de/misa/) (minimum repeat number of 6, 5, 5, 5, 5, and 4 for di-, tri-, tetra-, penta-, hexa-, and hepta-nucleotides, respectively). The primary trimming of redundant contigs of reference species (EH-n) was fulfilled by sequence clustering using BLASTCLUST with a sequence-identity cutoff of 90% and length coverage threshold of 50% (Altschul *et al.* 1997). The contigs exceeding the median of the number of the clustered contigs were trimmed for further analysis. In this study, single clustered contigs were selected because the median value of each number of all clustered sequences was 1.

Discovery of polymorphic SSR between two WGS data

The trimmed WGS reads of EH-v were aligned to single-clustered SSR contigs of EH-n using clc_mapper with the matched length fraction of the 90% and sequence-similarity of 90% (ver. 4.21.104315, CLC Inc, Aarhus, Denmark). The second trimming of redundant

contigs of EH-n was done by removing the high-depth contigs that have more than 10 mapping coverage by WGS reads of EN-v, because highly mapped contigs could have possibly been derived from redundant DNA regions such as repeats of the genome of *E. hamiltonianus*. Variation sites that indicate a consistent difference between the contigs of EH-n and WGS reads of EH-v were found using clc_find_variation (ver. 4.21.104315, CLC Inc, Aarhus, Denmark). The information of the sequence variation file and SSR motif file from MISA (http://pgrc.ipk-gatersleben.de/misa/) were combined to estimate the polymorphic regions including microsatellite sites using in-house python program. Primer binding sites were designed using primer3 (Rozen and Skaletsky 2000).

PCR validation of designed SSR markers

Genomic DNA of two *E. hamiltonianus* accessions were used for PCR validation of SSR markers. PCR amplification was proceeded in a 25 μL reaction volume containing the following components: 20 ng of DNA template, 10 μM of primer set, 5 mM of dNTP, and one unit of *Taq* DNA polymerase (Vivagen, Seongnam, Korea). The amplification condition was as follows: 5 min at 94°C, 35 cycles of 94°C 20 sec, 58°C 20 sec, and 72°C 20 sec, and then 72°C for 7 min. PCR products were then separated by 12% polyacrylamide gel electrophoresis for two and a half hours to identify polymorphisms. The gel was stained with ethidium bromide and visualized under UV lamps for manual genotyping.

Results

Complete cp genome and 45s nuclear ribosomal DNA assembly

Two *E. hamiltonianus* were collected from Hantaek Botanical Garden in South Korea (http://www.hantaek.co.kr, South Korea). One has normal leaves (hereafter EH-n), but the other has unique leaves with variegated patterns (hereafter EH-v) (Fig. 1-1). Approximately 477 Mbp and 481 Mbp of WGS sequence data were obtained from EH-n and EH-v using Illumina Miseq platform, respectively.

Complete cp genome sequences of EH-n and EH-v were successfully assembled through dnaLCW using the high quality WGS reads with manual curation (Table 1-1 and Fig. 1-2) (Kim *et al.* 2015). Both EH-n and EH-v had identical cp genome sequences, which were 157,360 bp in length (GenBank: KY921875). The genome composed of four sections: a LSC of 86,399 bp, a SSC of 18,317 bp, and a pair of IRs of 26,322 bp. A total of 113 genes including 79 protein-coding genes, 30 transfer RNA genes, and 4 ribosomal RNA genes were annotated in the cp genome (Fig. 1-2). Phylogenetic analysis was done using complete cp genome sequence of *E. hamiltonianus* with 11 cp genomes of related species (Fig. 1-3). Three monophyletic groups were divided into three cohorts: rosids, asterids, and commelinids (Fig. 1-3). The 45s nrDNA sequences of 5,824 bp were also assembled for EH-n and EH-v, which includes complete transcriptional unit sequences (Genbank: KY926695) (Table 1-1 and Fig. 1-4).

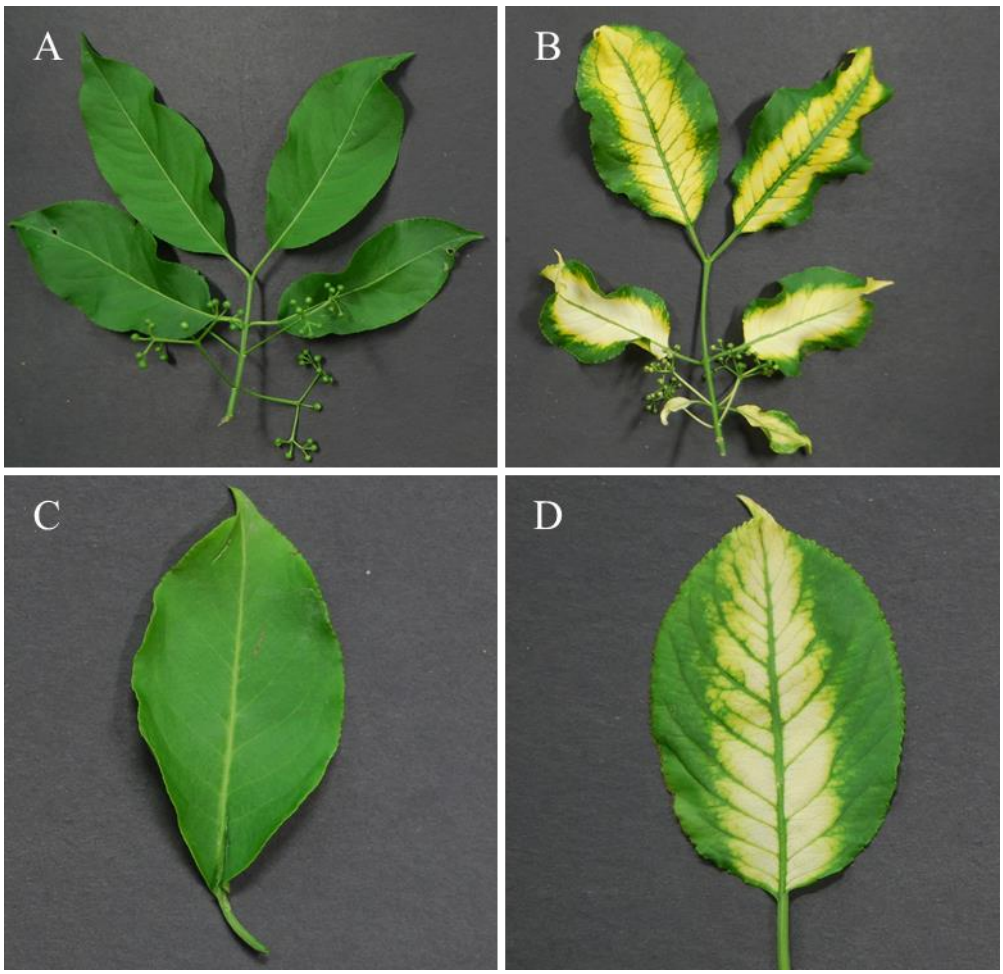

Figure 1-1. Leaf morphologies of two *E. hamiltonianus* genotypes. (A, C) Leaves and unopened flowers of normal plant. (B, D) Variegated leaves and unopened flowers of natural mutant plant which is being developed as an ornamental cultivar. The picture was taken from Hantaek Botanical Garden (2017-05-12).

Table 1-1. Status of chloroplast genome and nrDNA of two *E. hamiltonianus* genotypes.

Name	Chloroplast genome			nrDNA		
	Length (bp)	Coverage (x)	Alinged reads	Length (bp)	Coverage (x)	Alinged reads
E. hamiltonianus (normal)	157,360	72.72	46,187	5,824	381.69	9,500
E. hamiltonianus (variegated)	157,360	707.52	441,491	5,824	355.17	8,829

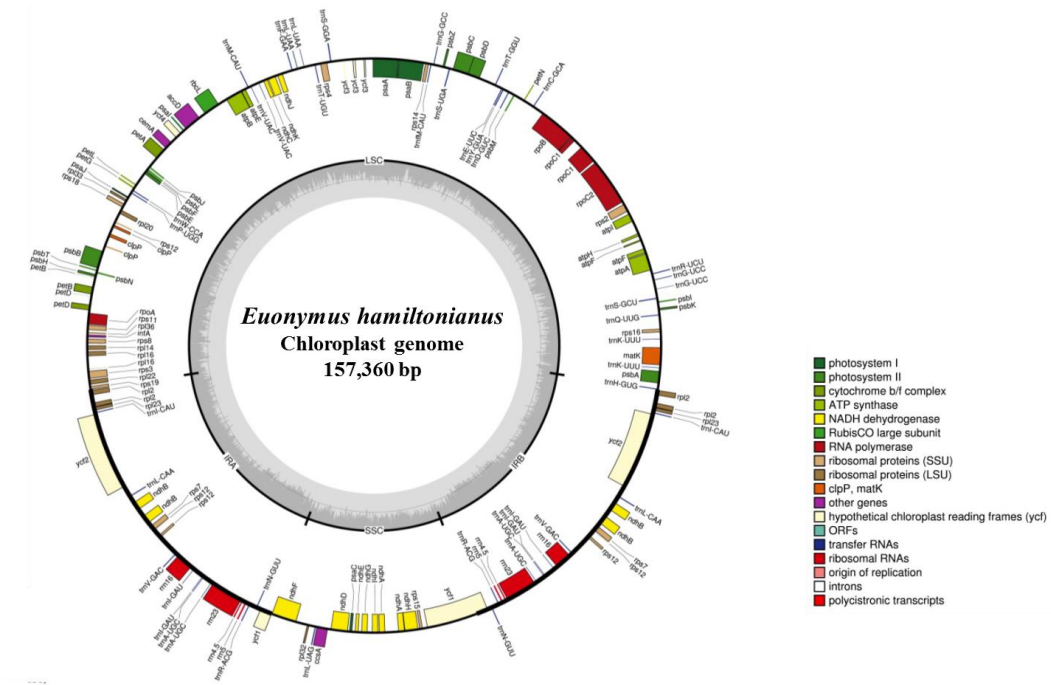

Figure 1-2. Chloroplast genome map of *E. hamiltonianus*. The complete cp genome sequence was annotated by the DOGMA program (http://dogma.ccbb.utexas.edu/). The map was generated using OGDRAW (http://ogdraw.mpimp-golm.mpg.de/). Genes in inner-circle and outer-circle were transcribed clockwise and anti-clockwise, respectively. The features of GC contents are displayed in the inner ring with internal blocks of cp genome, such as long single copy section (LSC), inverted repeat B (IRB), short single copy section (SSC), and inverted repeat A (IRA).

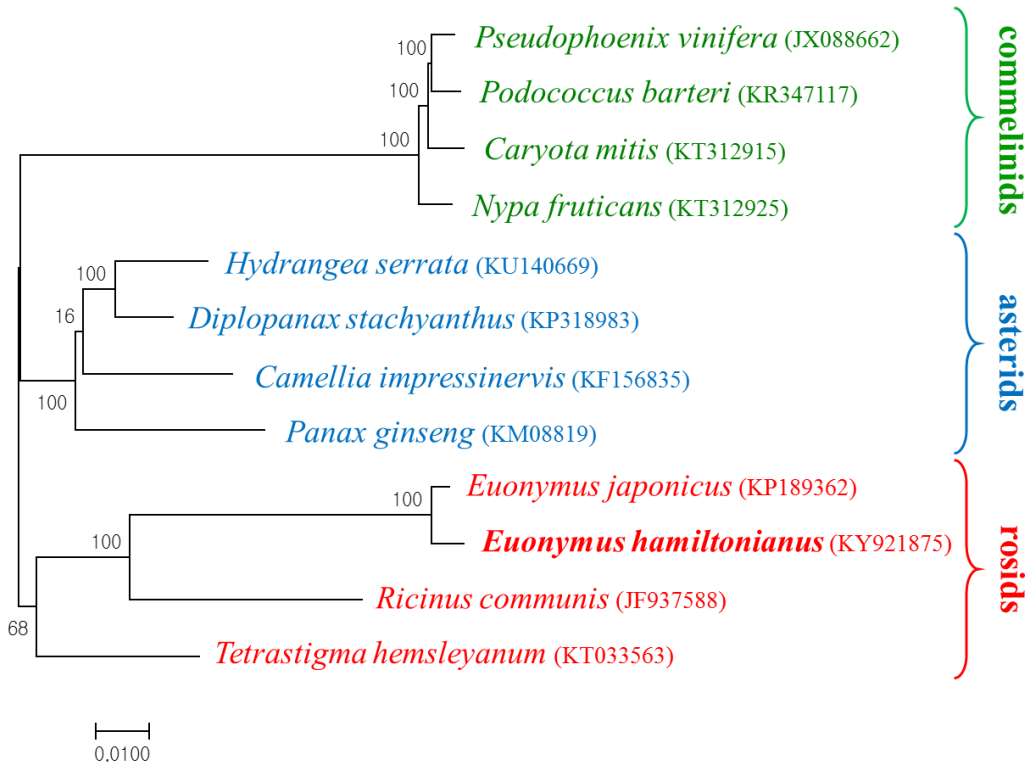

Figure 1-3. Phylogenetic analysis of *E. hamiltonianus* with eleven related species. Phylogenetic tree was prepared using complete cp genomes of 12 species. The green, blue, and red letters indicate the cohorts of commelinids, asterids, and rosids, respectively. The cp genome-based phylogenetic tree showed that *E. hamiltonianus* was grouped with other cohort rosids. The tree was generated by maximum likelihood using MEGA 7 (Kumar *et al.* 2016).

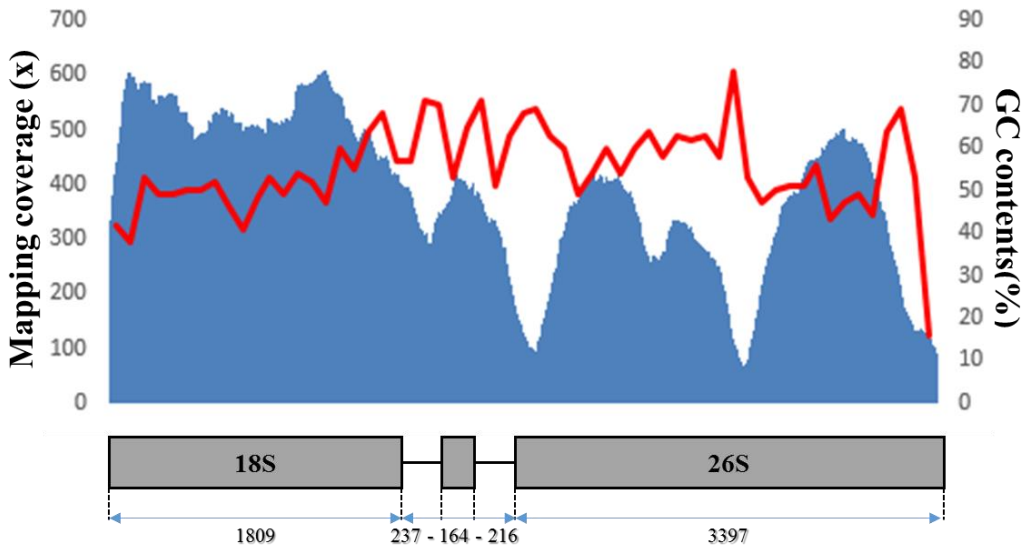

Figure 1-4. Schematic diagram of a complete 45S nrDNA unit of *E. hamiltonianus.* The WGS reads of EH-v were mapped again to the 45s nrDNA unit. GC content per 100 bp unit length is indicated by the red line.

Establishment of pipeline for detection of polymorphic SSR using low coverage of WGS (dpsLCW)

The dpsLCW contains 1) contig set construction by PE joining and singleton selection based on length of reference WGS data set, 2) identification of SSR motifs in contig set, 3) primary selection of potentially non-repetitive (NR) contigs by clustering, 4) secondary selection of potential NR contigs by alignment with WGS reads of alternative species, and 5) finding pSSR candidate contigs.

Firstly, at least two WGS data sets of closely related species (or intra-species) were required. The paired-end (PE) reads of both samples were trimmed based on quality score and combined (or assembled) into a contig by overlapping each PE read. Each sample was assigned to reference and alternative, respectively. Considerably long sequences of reference among overlapped contigs and non-overlapped singleton reads that are able to find primer binding sites (PBS) and can be compared with alternative WGS reads without difficulty were selected. In this study, the contigs under 250 bp in length were screened. The MiSeq platform was recommended with 500 bp library insert size which could easily join PE reads and also non-joined singleton could generally have a long sequence with an average of 300 bp in length. Secondly, the reference contigs were used to find robust SSR motifs using MISA (http://pgrc.ipk-gatersleben.de/misa/). Thirdly, NR candidates obtained among SSR motifs contigs were primarily selected by sequence clustering. Single clustered contigs were used. Fourthly, the selected NR contigs with SSR motifs of reference were aligned to trimmed WGS reads of the

alternative. High-mapping depth contigs were discarded because these contigs could have been derived from redundant genomic regions such as repetitive DNA sequences. Finally, pSSR contigs were selected using the information of alignment between reference and alternative from the previous step. All steps of this progress are portrayed in the schematic diagram in Figure 1-5.

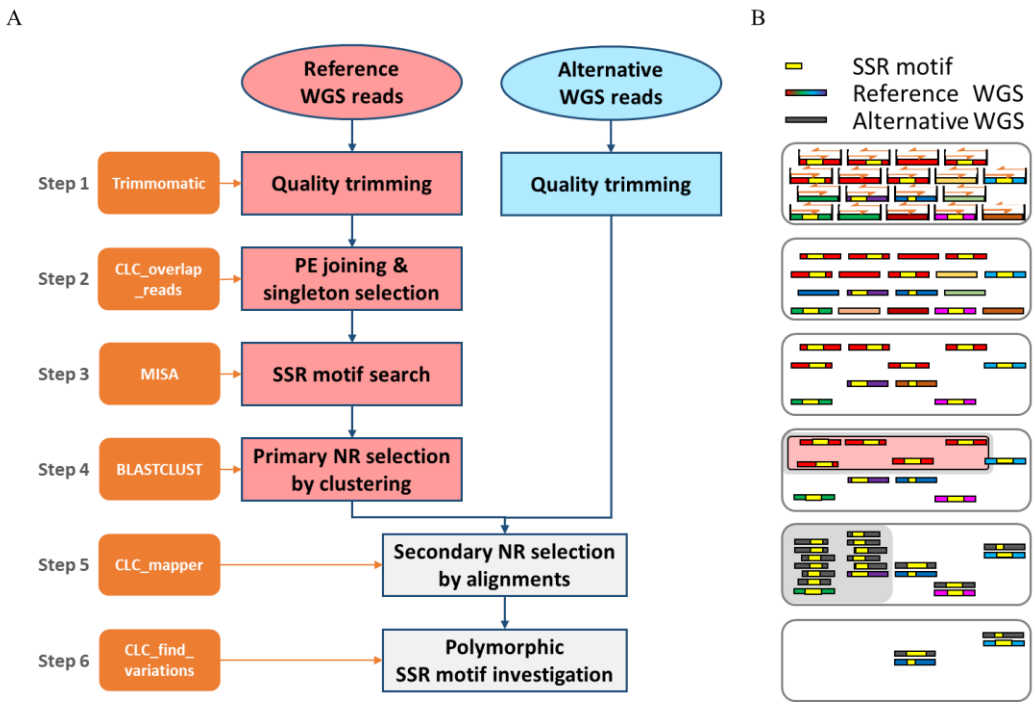

Figure 1-5. Pipeline for dpsLCW. (A) Steps for WGS reads of reference and alternative species are shown in different colored boxes, blue and red, respectively. The adopted programs in each protocol are shown in orange boxes. (B) Yellow colored, variously colored, and grey colored small rectangles indicate SSR motifs, reference WGS reads, and alternative WGS reads, respectively. Same colored rectangles of reference WGS represent homologous WGS reads. WGS reads in the pink box indicate one cluster in step 4. The WGS reads in light grey regions in step 4 and 5 were not used in further steps.

Verification of dpsLCW protocol with *E. hamiltonianus*

The dpsLCW was applied for the identification of intraspecific variations between EH-n and EH-v. A number of quality-trimmed sequences of 426 Mb and 423 Mb were prepared for EH-n and EH-v, respectively. 508,265 PE reads of EH-n were joined into single contigs by the alignment between each PE read (63.8%). A total of 457,385 PE reads in EH-v were also joined into single contigs (57.9%). Thus, EH-n as reference and EH-v as an alternative were used, because the number of jointed contigs of EN-c were smaller than those of EH-n. The contig data set and the remaining 596,964 non-joined singletons of EH-n were filtered out by a read length of up to 250 bp due to their capability of easily finding their potential PBS. Qualified 805,049 contigs were used to search SSR motifs. Within these contigs, MISA detected 19,053 contigs comprising of various SSR motifs. After this, sequence clustering was performed to investigate NR contigs using BLASTCLUST (Altschul *et al.* 1997). Single clustered contigs of 7,629 were chosen based on the median value of all clusters for further analysis (median value: 1) (Table 1-2).

To identify pSSR motifs between two samples, 1,600,199 WGS reads of EH-v were aligned to 7,629 contigs of EH-n. Only 46,284 reads of EH-v (2.89%) were mapped onto EH-n contigs with an average coverage of 2.78 (maximum coverage: 10243.3). A total of 116 contigs were discarded with more than ten mapping coverage by WGS reads of EH-v, because highly mapped contigs might have originated from repetitive DNA sequences. Moreover, 6,054 unmapped contigs were excluded for the unavailability of pSSRs. Among the remaining, 1,459 contigs which have been mapped under ten mapping coverage by

EH-v reads, 161 contigs of 60,859 bp sequences were finally selected for the detection of pSSR (Table 1-2). A total of 163 pSSR motifs were found between EH-n and EH-v, of which 1 contig had 3 SSR motifs including a dinucleotide and two trinucleotide SSR motifs. 117 contigs had dinucleotide SSR motifs, 31 contigs had trinucleotide SSR motifs, eight contigs had tetranucleotide SSR motifs, and six contigs had more than pentanucleotide SSR motifs (Fig. 1-6).

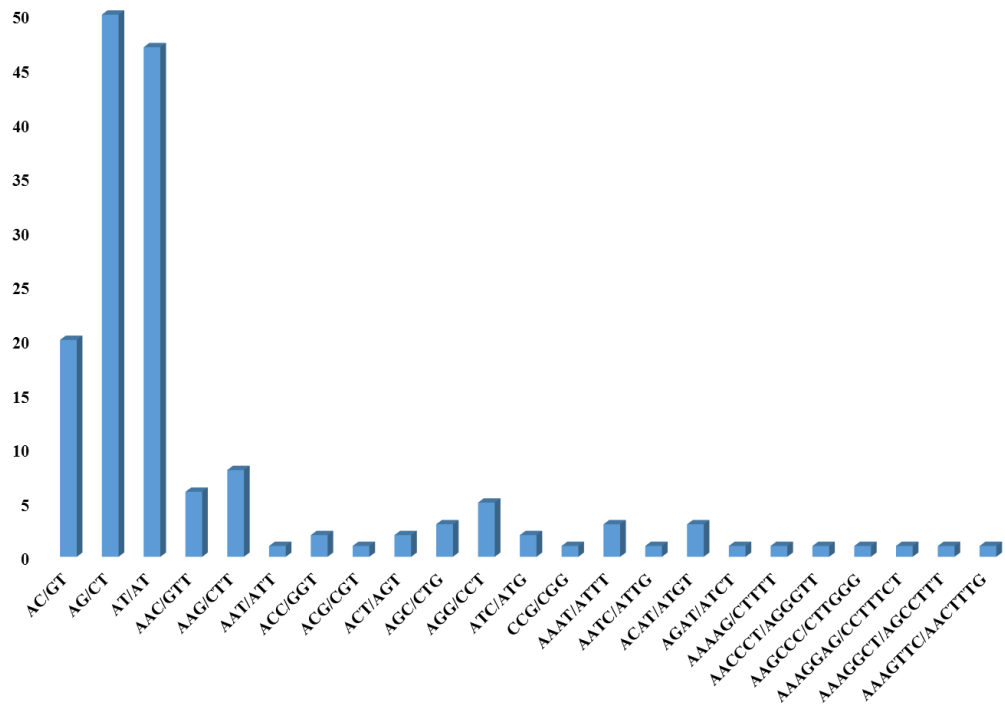

Figure 1-6. Distribution of candidates of polymorphic SSR motifs of *E. hamiltonianus*. Among the detected SSR, the AG/CT dinucleotide repeat motif was most abundant.

Validation of predicted pSSR for *E. hamiltonaunis*

To validate that identified pSSRs show actual variance between EH-n and EH-v, 20 contigs were randomly chosen and designed primer pairs using Primer3 (Koressaar and Remm 2007, Untergasser *et al.* 2012) (Table 1-3). The 20 primer pairs were successfully amplified (Fig. 1-7). Among them, seven polymorphic markers were developed including a dominant marker of EhSSR08 (Fig. 1-7). Using these markers, EH-n and EH-v can be distinguished from each other (Fig. 1-7). These SSR markers will help in the evaluation and classification of another *E. hamiltonaunis.*

Table 1-2. Status of WGS reads of two *E. hamiltonianus* accessions in dpsLCW

Phase	Contents of phase	*E. hamiltonianus* (normal)		*E. hamiltonianus* (variegated)	
		reads	basepairs	reads	basepairs
	Raw data	1,637,296	477,360,027	1,625,572	481,075,391
Step 1	Trimmed data	1,613,494	426,154,231	1,600,199	423,728,477
Step 2	PE joined & singleton (≥250)	805,049	284,940,618		
Step 3	SSR containing contigs	19,053	7,169,090		
Step 4	Primary filtered contigs	7,629	2,564,370		
Step 5	Secondary filtered contigs	1,459	564,592		
Step 6	pSSR containg contigs	161	60,859		

Table 1-3. Primer information of candidate pSSR markers evaluated in this study.

Marker ID	SSR motif		Contig length (bp)	Estimated product size		Primers	Description of blastx (e-value)
	EH-n	EH-v		EH-n (bp)	EH-v (bp)		
EhSSR01	(TC)17	(TC)6	488	254	232	F GAAATTGTGCACTCCCCTGTT R TCTCAAAATGCGAAGCGCAG	
EhSSR02	(TC)17	(TC)7	404	220	200	F CGGATCAACCAGTCGTCCAA R TACTGTGCTAGCCCAAACCG	XP_011035889 PREDICTED: probable methyltransferase PMT23 (4e-04)
EhSSR03	(AT)16	(AT)6	463	215	195	F GGTGCAGGTTCAGAAAGGCT R AGAGCCAAATCGACAAAAAGGG	
EhSSR04	(AGA)10	(AGA)4	393	280	262	F TCACTAACCTGCTTGCACCAA R GAGAGCGATGAAGATGCGTG	
EhSSR05	(GA)12	(GA)5	287	188	174	F TAGTAGTCGAGTGGGATGGGG R TCATGTGCCACCGAAATACCAA	
EhSSR06	(GAAAGGA)6	(GAAAGGA)4	284	207	193	F CCGAGCCGGATCTTGAAAGT R TGGATAGGTCCGGATTGCCT	
EhSSR07	(CT)26	(CT)19	379	298	284	F TGTGTGGCCAAGACACAAGT R ACTGGCAACTTTCCTAGACTGA	
EhSSR08	(GA)16	(GA)10	477	273	261	F CCAGCAAAAGCTTAAGGAAACGA R GCACATCTCCATTGCAAGTTCA	
EhSSR09	(TA)11	(TA)5	453	283	271	F GGCCTCGTTACTGCTATGCT R TGCCATCGTATTTGGGTCCT	XP_015382307 PREDICTED: patatin-like protein 2, partial (3e-59)
EhSSR10	(CTG)7	(CTG)3	430	232	220	F GCCATGGACTAATTGCTGCG R TGGGACCAACAAGCCAACAT	XP_002526966 PREDICTED: protein SIEVE ELEMENT OCCLUSION B (0.073)

Table 1-3. (continued)

Marker ID	SSR motif		Contig length (bp)	Estimated product size		Primers	Description of blastx (e-value)
	EH-n	EH-v		EH-n (bp)	EH-v (bp)		
EhSSR11	(GA)9	(GA)16	301	231	244	F ACGTCACATCCACCATGCAA R ATGGCATTCCGTCCGTGATT	
EhSSR12	(AT)10	(AT)6	371	236	228	F GAATGCATGCCACTCCAACA R ATAAGCAATTGGGGAACCTAGTA	XP013315502 hypothetical protein PV05_07244 (5.7)
EhSSR13	(AG)8	(AG)5	270	251	245	F TCAGGTCTTGCAGTCTCTGATTT R GAAGAAGGGGCAGAGGTTGTT	XP_015868974 PREDICTED: uncharacterized protein LOC107406380 (3e-13)
EhSSR14	(TGT)6	(TGT)4	270	179	173	F ACATACACGCACCTTAGGTCA R CAATCGCAGCAGCAACAGTATC	XP018845393 PREDICTED: DEAD-box ATP-dependent RNA helicase 8-like (4e-04)
EhSSR15	(TA)7	(TA)4	286	236	230	F AGTCCCCGCTAAGAGGCATA R AACACAGAGAAGTCTGCGGG	
EhSSR16	(ACA)6	(ACA)4	532	202	196	F AGGACAGACATGGCCTTTCAC R CCGAGAAGTTCGGAGGTTGT	XP_016689476 PREDICTED: homeobox protein knotted-1-like 3 (2e-07)
EhSSR17	(ATG)8	(ATG)6	407	220	211	F CCAAAGCGAGATGAGTGTGTTAAT R TCGTCCAGTTGGGGTCCTTT	XP_002301160 hypothetical protein POPTR_0002s12380g (9e-14)
EhSSR18	(GGT)5	(GGT)3	436	269	263	F GTTGGTTTATCTGGGTTGGCT R ATTGGGTGAGCAGCACTGTA	
EhSSR19	(ATAA)5	(ATAA)4	301	170	166	F TGCACAAGAGTTCTTTATTTCAGCA R GCAGTAGCTTAGCATGGGTCA	
EhSSR20	(ATGT)5	(ATGT)4		155	147	F AGCTTGGCTTGCCTTTTTCAG R ACAATTATGGATGCATTTGTTGTTT	

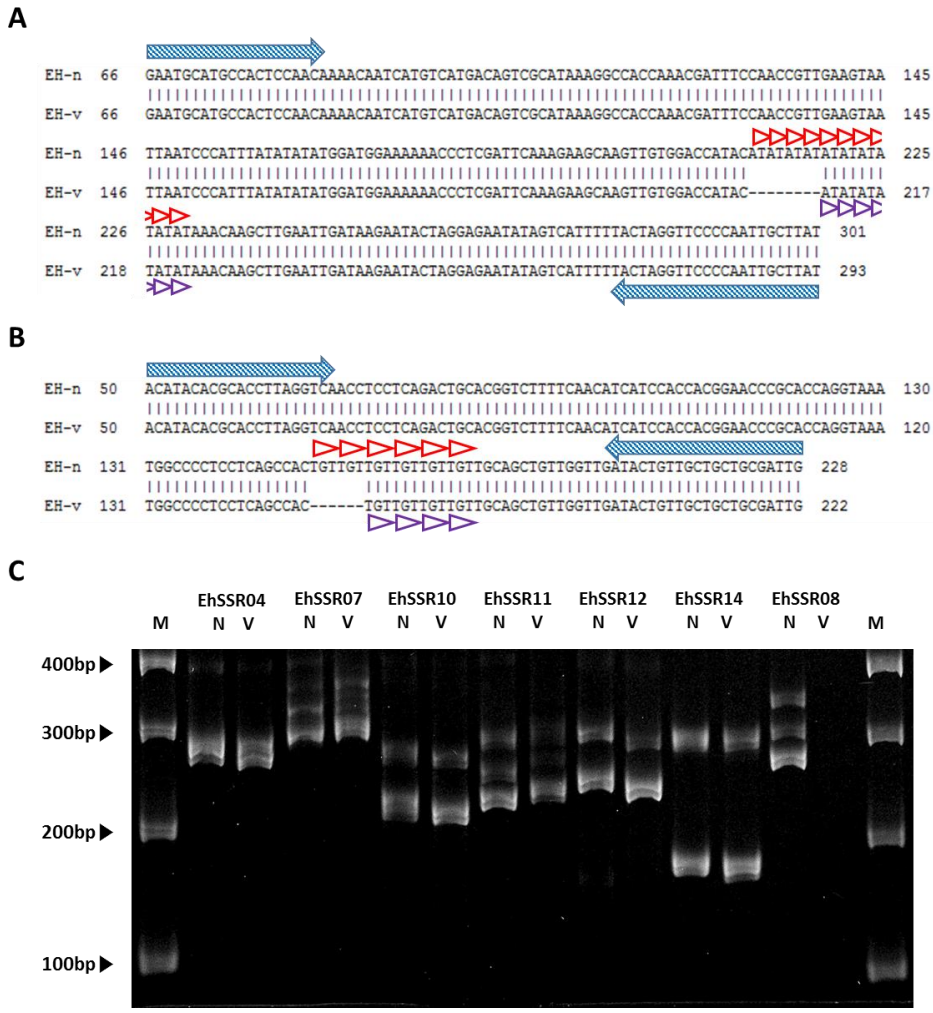

Figure 1-7. PCR validation of seven EhSSR markers in two *E. hamiltonianus* genotypes. A. diagram of EhSSR12 between EH-n and EH-v. B. diagram of EhSSR14 between EH-n and EH-v. Red and purple supine triangles describe SSR motifs in EH-n and EH-v, respectively. C. N and V indicate EH-n and EH-v, respectively. M indicates 100-bp ladders.

Discussion

Complete cp genomes and 45S nrDNA sequences of *E. hamiltonianus*

The non-model or underutilized plants are plants with potential value but are not widely grown (Park *et al.* 2009). Although much more research for the non-model plants is needed, there is a lack of genetic resources including the information of applicable molecular markers compared to major plants. Under these circumstances, dnaLCW which could help to improve knowledge and availability of non-model plants was previously invented (Kim *et al.* 2015). Using this, complete cp genome and 45S nrDNA of *E. hamiltonianus* were generated (Fig. 1-2 and 1-3). The phylogenetic tree of cp genome of *E. hamiltonianus* with 11 related species showed that rosids and asterids clades of core eudicots were distinguished from commelinid clade of monocots. As expected, *E. hamiltoniaunus* was placed closely to the *E. japonicus* (KB189362) species in the same genus with a nucleotide similarity of 97% in rosids clades. Interestingly, the generated EH-n and EH-v of *E. hamiltonianus* did not have any polymorphic sites in the complete cp genome and 45S nrDNA sequences.

False pSSRs derived from dpsLCW in *E. hamiltonianus*

SSR markers are one of the most applicable markers for the genetic evaluation and

utilization of crops due to their benefits: 1) high-reproducibility, 2) hyper-variable nature, and 3) co-dominant nature (Park *et al.* 2009). However, developing SSR markers was not easy in non-model plants due to insufficient genomic information (Gong *et al.* 2008, Ma *et al.* 2009). Therefore, an efficient protocol was devised in this study, named dpsLCW, to identify pSSR motifs for marker development to support fundamental research on non-model plants.

The dpsLCW was simulated to find pSSRs using two *E. hamiltonaunis* genotypes (EH-n and EH-v). Twenty SSR primers designed by dpsLCW were successfully amplified (EhSSR01-20). Among them, polymorphisms were observed in seven primer pairs including a dominantly amplified marker of EhSSR08 (35% of designed EhSSR primers). Those primer sets were almost consistent with the estimated size of PCR products (Fig. 1-7). The non-amplification of EhSSR08 in EH-v might be caused by nucleotide polymorphisms in primer pairs. The other unexpected non-polymorphic PCR products may have been due to paralogous sequences in *E. hamiltonianus.* Candidates of paralogous pairs were observed in some sequences with polymorphic sequences at flanking regions of SSR motifs (Fig. 1-8). NGS sequencing or assembly errors during the "PE joining & Singleton selection" step of dpsLCW could be one of the reason for unexpected non-polymorphic PCR products. However, 20 SSR markers by dpsLCW showed polymorphisms with a high success rate (7 out of 20). The seven markers and other 142 contigs with candidates of pSSR will provide fundamental information on the evaluation of useful genetic resources and breeding in *E. hamiltonaunis*. The dpsLCW could be a

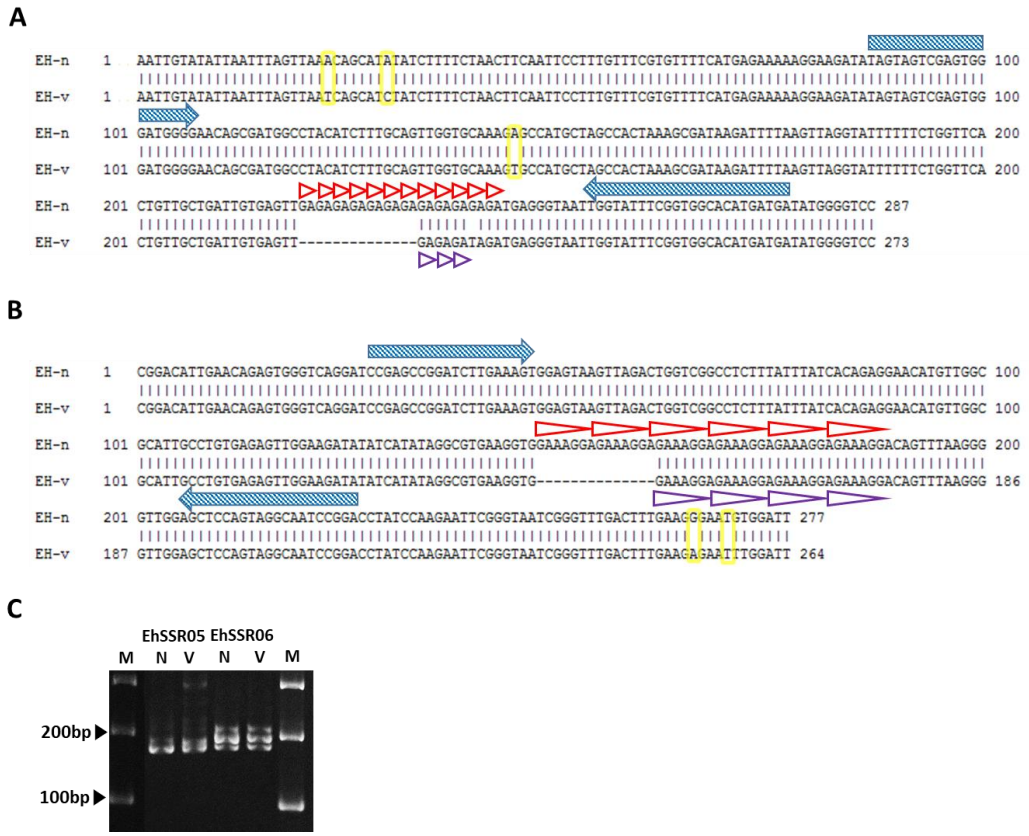

Figure 1-8. False pSSR products in *E. hamiltonianus.* A. diagram of EhSSR05 between EH-n and EH-v. B. diagram of EhSSR06 between EH-n and EH-v. Red and purple supine triangles describe SSR motifs in EH-n and EH-v, respectively. Yellow box indicate discordant nucleotides between EH-n and EH-v. C. N and V indicate EH-n and EH-v, respectively. M indicates 100-bp ladders.

considerably useful pipelines for polymorphic marker development and could be applied to other non-model plants.

The advantages of dpsLCW

The dpsLCW is a relatively economical method. The dpsLCW only requires two small scale WGS data, allowing it to be applied to non-sequenced or non-model plants regardless of the presence of reference or long sequence information. Recently, single molecule real time sequencing that can significantly increase the length of reads based on the sequencing of a single DNA molecule was developed (Eid 2009), but NGS technology could be more compatible with dpsLCW because of its inexpensive, wide genomic range coverage, and high-throughput productivity. In addition, a large number of plants could be also simultaneously investigated through the multiplexed sequencing technology (Smith *et al.* 2010).

The dpsLCW is considerably precise for detecting authentic pSSR. In this study, 35% (7 of 20) of the polymorphic SSR markers were successfully identified (Fig. 1-3). Unlike conventional researches, two factors may play a role in achieving high success rates with low rates of multi bands: (1) The dpsLCW can select pSSR candidates through the direct comparison between two WGS data sets. Conventional research for pSSR identification required time consuming wet experiments due to low polymorphic rates (with less than 5 % of success rate) from the SSR candidates because they utilized sequence

harboring the SSR motif from one genotype (Choi *et al.* 2011, Izzah *et al.* 2014, Kim *et al.* 2012). (2) The dpsLCW could remove highly abundant sequences such as repetitive DNA sequences through the filtering steps “Primary NR filtering by clustering” and “Secondary NR filtering by alignments”. Repetitive DNA, especially retrotransposons, could also be used as molecular markers. For example, inter-retrotransposons amplified polymorphisms (IRAP), sequence-specific amplified polymorphisms (S-SAP), and retrotransposon-microsatellite amplified polymorphism (REMAP) (Kalendar and Schulman 2006). However, most retrotransposon based molecular markers usually produce multiple bands with a wide range of lengths of amplicon and were seriously affected by genomic tendency of targeted retrotransposons. Additionally, advanced information of retrotransposons in the plant subject is needed. Moreover, a considerable portion of the dual filtered contigs might be composed of NR regions or genic regions like EST in the genome. To verify whether the dual trimmed contigs of EH-n were related to genic regions or not, the mapping with the 3,279,262 ESTs of *E. alatus* by 454 platform (SRA #: SRA025080) that belongs to the same genus was used. Among the 7,513 contigs, 1,334 (18%) contigs were mapped with RNA reads of *P. praeruptorum*. It may be deduced that at least one of the five filtered contigs through dpsLCW was related to the genic region. However, it is hard to say that the rest of contigs were not related to genic regions because the number of targeted ESTs of *E. alatus* was too small.

The dpsLCW has scalability. The designed SSR primer sets could successfully be applied to related species due to the variability of SSR motifs in plants. Furthermore,

another molecular marker system such as cleaved amplified polymorphic sequences (CAPS) or derived cleaved amplified polymorphic sequences (dCAPS) caused by SNP or INDELs in genome sequences might be applied through the dpsLCW. And of course, if there is WGS data of plant species registered in public databases, our protocol could be applied.

The dpsLCW has some limitations. In this study, Illumina Miseq platform were used with an average length of 300 bp and a library insert size of 500 bp. Thus, dpsLCW was able to easily select long sequences and to develop SSR markers with a potential PBS. When using the relatively short Illumina sequence of HiSeq, NextSeq, or other platforms, long contigs could be generated by the *de novo* assembly process or control of WGS library size that can be possibly joined into one long contig using both PE reads. The other limitation is that the step of NR filtering by clustering and alignment in dpsLCW could not perfectly filter out all repeat rich regions in the genome. This can be compensated by increasing the amount of sequencing data or by filtering repetitive DNA of relative species using acceptable public database.

Reference

Abdelkrim J, Robertson B, Stanton JA, Gemmell N. 2009. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques 46: 185-192.

Agarwal M, Shrivastava N, Padh H. 2008. Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep. 27: 617-631.

Allen G, Flores-Vergara M, Krasynanski S, Kumar S, Thompson W. 2006. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat. Protoc. 1: 2320-2325.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FI, Costa de Oliveira A. 2008. SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation. Int. J. Plant Genomics 2008: 412696.

Choi HI, Kim NH, Kim JH, Choi BS, Ahn IO, Lee JS, Yang TJ. 2011. Development of reproducible EST-derived SSR markers and assessment of genetic diversity in *Panax ginseng* cultivars and related species. J. Ginseng Res. 35: 399-412.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. 2009. Real-time DNA sequencing from single polymerase molecules. Science 323: 133-138.

Gong L, Stift G, Kofler R, Pachner M, Lelley T. 2008. Microsatellites for the genus Cucurbita and an SSR-based genetic linkage map of *Cucurbita pepo* L. Theor. Appl. Genet. 117: 37-48.

Grover A, Sharma PC. 2016. Development and use of molecular markers: past and present. Crit. Rev. Biotechnol. 36: 290-302.

Hiroe M, Constance L. 1958. Umbelliferae of Japan. University of California publications in botany 144.

Izzah NK, Lee J, Jayakodi M, Perumal S, Jin M, Park BS, Ahn K, Yang TJ 2014. Transcriptome sequencing of two parental lines of cabbage (*Brassica oleracea* L. var. capitata L.) and construction of an EST-based genetic map. BMC Genomics 15: 149.

Jones AG, Small CM, Paczolt KA, Ratterman NL. 2010. A practical guide to methods of parentage analysis. Mol. Ecol. Resour. 10: 6-30.

Kalendar R, Schulman AH. 2006. IRAP and REMAP for retrotransposon-based

genotyping and fingerprinting. Nat. Protoc. 1: 2478-2484.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30: 3059-3066.

Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, Makova KD. 2010. What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. Genome Biol. Evol. 2: 620-635.

Kim K, Lee SC, Lee J, Lee HO, Joh HJ, Kim NH, Park HS, Yang TJ. 2015. Comprehensive Survey of Genetic Diversity in Chloroplast Genomes and 45S nrDNAs within *Panax ginseng* Species. PloS ONE 10: e0117159.

Kim K, Lee SC, Lee J, Yu Y, Yang K, Choi BS, Koh HJ, Waminal NE, Choi HI, Kim NH, Jang W, Park HS, Lee J, Lee HO, Joh HJ, Lee HJ, Park JY, Perumal S, Jayakodi M, Lee YS, Kim B, Copetti D, Kim S, Kim S, Lim KB, Kim YD, Lee J, Cho KS, Park BS, Wing RA, Yang TJ. 2015. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. Sci. Rep. 5: 15655.

Kim NH, Choi HI, Ahn IO, Yang TJ. 2012. EST-SSR marker sets for practical authentication of all nine registered ginseng cultivars in Korea. J. Ginseng Res. 36: 298-307.

Koressaar T, Remm M. 2007. Enhancements and modifications of primer design program Primer3. Bioinformatics 23: 1289-1291.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis

Version 7.0 for Bigger Datasets. Mol. Biol. Evol. 33: 1870-1874.

Long EO, Dawid IB. 1980. Repeated genes in eukaryotes. Annu. Rev. Biochem. 49: 727-764.

Ma KH, Kim NS., Lee GA, Lee SY, Lee JK, Yi JY, Park YJ, Kim TS, Gwag JG, Kwon SJ. 2009. Development of SSR markers for studies of diversity in the genus *Fagopyrum.* Theor. Appl. Genet. 119: 1247-1254.

Meng W, Fei X, Peng Y, Duan XY, Zhou YL, Shen CY, Zhang GZ, Wang BT. 2014. Development of SSR Markers for a Phytopathogenic Fungus, *Blumeria graminis* f. sp. tritici, Using a FIASCO Protocol. J. Integr. Agr. 13: 100-104.

Metz S, Cabrera JM, Rueda E, Giri F, Amavet P. 2016. FullSSR: Microsatellite Finder and Primer Designer. Adv. bioinformatics 2016: 4.

Mittal N, Dubey AK. 2009. Microsatellite markers-A new practice of DNA based markers in molecular genetics. Phcog. Rev. 3: 235.

Park YJ, Lee JK, Kim NS. 2009. Simple sequence repeat polymorphisms (SSRPs) for evaluation of molecular diversity and germplasm classification of minor crops. Molecules 14: 4546-4569.

Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402: 404-407.

Richard GF, Kerrest A, Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol. Biol. Rev. 72: 686-727.

Rogers SO, Bendich AJ. 1987. Heritability and Variability in Ribosomal RNA Genes of

Vicia faba. Genetics 117: 285-295.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132: 365-386.

Soltis PS, Soltis DE, Chase MW. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402: 402–404.

Seol YJ, Lee TH, Park DS, Kim CK. 2016. NABIC: A New Access Portal to Search, Visualize, and Share Agricultural Genomics Data. Evol. Bioinform. Online 12: 51-58.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26: 1135-1145.

Smith AM, Heisler LE, St Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N, Nislow C. 2010. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Res. 38: e142.

Sol m. 2010. An illustrated guide to Korean medicinal plants. NesusBOOKS

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3--new capabilities and interfaces. Nucleic Acids Res. 40: e115.

Varshney RK, Graner A, Sorrells ME. 2005. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 23: 48-55.

Varshney RK, Nayak SN, May GD, Jackson SA. 2009. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 27: 522-530.

CHAPTER II

High-throughput development of polymorphic simple sequence repeat markers using two whole genome sequence data in *Peucedanum japonicum*

Abstract

Resource plants are important and have strong potential for a variety of utilities as crops or pharmaceutical materials. However, most resource plants remain wild and thus their utility for breeding and biotechnology is limited. Molecular markers are useful to initiate genetic study and molecular breeding for these understudied resource plants. Various wild *Peucedanum japonicum* which is indigenous resource plant utilized as oriental medicine and leafy vegetable in Korea were collected. In this study, two independent whole genome sequences (WGSs) were produced from two collections and identified large scale polymorphic simple sequence repeat (pSSR) based on our pipeline to develop SSR markers based on comparison of two WGSs. A total of 452 candidate pSSR contigs were identified. To confirm the accuracy and utility of pSSR, ten SSR primer pairs were designed and were successfully applied those to seven collections of *P. japonicum.* The WGS and pSSR candidates identified in this study will be useful resource for genetic research and breeding purpose for the valuable resource plant, *P. japonicum.*

Keywords: molecular markers, polymorphisms, simple sequence repeats, whole genome sequence, *Peucedanum japonicum*.

INTRODUCTION

Indigenous resource plants have important values for a variety of utilities such as crops or pharmaceutical materials although there is limited research. Molecular markers are actively applied in many major crops, but the development of marker on the minor crops and resource plants is still limited. A molecular marker is a particular DNA fragment which includes specific genetic information with genome-level polymorphisms (Agarwal *et al.* 2008). Molecular markers have been widely used as valuable tools to measure genetic variation and thereby highly enhanced the genetic analysis on plants (Varshney *et al.* 2005).

Simple sequence repeats (SSRs, otherwise called as microsatellites) (Hiroe *et al.* 1958, Jones *et al.* 2010) which consist of sequential repeat of one to six or more nucleotide motifs in a head-to-tail structure (Kelkar *et al.* 2010) have been used often in genetic studies. SSR markers are further universally implemented for intraspecies variation and population genetic studies such as parentage analysis, fingerprinting, genetic structure analysis and genetic mapping (Grover *et al.* 2016, Meng *et al.* 2014, Mittal *et al.* 2009) due to variable polymorphic features and high reproducibility. Because of the utility and advantage of SSR markers, various programs have been developed to identify SSR motifs such as MISA (http://pgrc.ipk-gatersleben.de/misa/), SSR Locator, and FullSSR (Abdelkrim *et al.* 2009, da Maia *et al.* 2008, Metz *et al.* 2016). However, conventional SSR marker development and experimental verification still require long guided sequence information and a labor-intensive trial and error procedure by repetitive wet experiments

to find polymorphic SSR markers (Ma *et al.* 2009).

Peucedanum japonicum is a perennial herb that belong to the Apiaceae family and grows on the cliff or between rocks in coastal regions of Taiwan, Japan, China and Korea (Buwalda 1949, Hiroe and Constance 1958, Seo *et al.* 2001). The aerial part of *P. japonicum* such as young buds, leaves, and berries are used as foodstuff whereas the underground roots have been used as oriental herbal medicine in the treatment of headache, anemia, apoplexy, paralysis, common cold and respiratory illnesses (Amano 1987, Hisamoto *et al.* 2003, Kang *et al.* 2013). *P. japonicum* has a variety of utilities. It is commonly used as a leafy vegetable and an herbal medicine, and a growing interest in this plant has increased framer's cultivation in Korea. However, despite this growing interest, genomic information to advance breeding efforts on *P. japonicum* is limited (Han et al. 2017, Seo *et al.* 2001).

Next-generation sequencing (NGS) technology is widely used because of its high-throughput productivity (Choi *et al.* 2011, Kim 2012, Varshney *et al.* 2009). Recently, A high-throughput method to assemble complete sequences of chloroplast genome and nuclear ribosomal DNA (nrDNA) simultaneously using low coverage WGS data were developed, coined as *de novo* assembly using low coverage WGS (dnaLCW) (Kim *et al.* 2015). The dnaLCW method was successfully applied to develop markers for genetic diversity for resource plants. However, the information is limited to the chloroplast genome and nrDNA which are highly conserved genomic regions. Therefore, large-scale SSR markers derived from nuclear genomes are necessary for further genetics and genomics approach for breeding. Here, I would like to expand our approach to find large-scale

polymorphic SSR markers using WGS data and provide useful genomic information and genetic pSSR markers that can be utilized for genetics and breeding of the valuable resource plant, *P. japonicum*.

MATERIALS AND METHODS

Plant materials, genomic DNA extraction and NGS sequencing

Various wild *P. japonicum* collections were collected in the Southern coastal area in Korea. These collections is maintained in Seoul National University farm, Suwon, in Korea. Among them, seven *P. japonicum* collections were used in this study. Three wild individuals were collected from Jeju Island (hereafter PJ #1), Geumo Island (hereafter PJ #2) and Wan Island (hereafter PJ#3) in Korea. The other two individuals (hereafter PJ#4 and PJ#5) were provided from Hantaek botanical garden (http://www.hantaek.co.kr, South Korea) and other two individuals (hereafter PJ#6 and PJ#7) were provided from Medicinal herb garden of college of Pharmacy in Seoul National University (http://www.snuherb.ac.kr). Hereafter All *P. japonicum* sample will be refer to their collection number. The fresh leaves of seven *P. japonicum* individuals were collected for DNA preparation. Each leaf sample was grinded separately using liquid nitrogen. The genomic DNAs were extracted using a modified cetyltrimethylammonium bromide (CTAB) method (Allen *et al.* 2006). Among seven plants, PJ#1 and PJ#2 were used to construct genomic libraries with insert size of 500 bp, according to Illumina paired-end (PE) standard protocol (http://www.illumina. com) and sequenced using Illumina MiSeq genome analyzer from LabGenomics (www.labgenomics.co.kr). The raw sequences of PJ#1 and PJ#2 have been uploaded at NABIC database and registered as NN-2130-000001 and NN-2132-000001, respectively (http://nabic.rda.go.kr/) (Seol *et al.* 2016).

Sequence preparation (quality control and paired end joining), identification of SSR motif and clustering

WGS reads of PJ#1 and PJ#2 were trimmed off using Trimmomatic (ver. 0.33) based on quality score and sequence length; condition is set to the minimum quality score: ≥20, read length: ≥70 bp (Bolger 2014). Trimmed PE reads of PJ#1 were further assembled for joining of both forward and reverse PE sequences for each read by clc_overlap_reads (ver. 4.21.104315, CLC Inc, Aarhus, Denmark). Forward and reverse sequences for each PE reads were assembled and were made one PE-joined contigs (contigs) by combining of both forward and reverse reads based on the overlap. In the contig sets, contigs greater than 250 bp including SSR motifs were identified using microsatellite search module (MISA, http://pgrc.ipk-gatersleben.de/misa/) with the minimum repeat number of 6, 5, 5, 5, 5, and 4 for di-, tri-, tetra-, penta-, hexa-, and hepta-nucleotides, respectively. The contigs of reference species (PJ#1) were clustered based on homology using BLASTClust (https://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html, Altschul *et al.* 1997) and the clusters with single contigs were selected for further analysis because these sequence were considered as derived from non-repeat genomic regions.

Discovery of pSSRs between two WGS reads

The WGS reads of PJ#2 were aligned with the contigs of PJ#1 in previous step by using clc_mapper (ver. 4.21.104315, CLC Inc, Aarhus, Denmark). Afterward, the candidates of pSSR have been located through comparison between contigs of PJ#1 and WGS reads of PJ#2 using clc_find_variation (ver. 4.21.104315, CLC Inc, Aarhus, Denmark). The

information of the sequence variation file and the SSR motif file from MISA were combined to estimate the polymorphic regions including microsatellite sites using in-house python program.

PCR validation of SSR markers

Using sequences information with the candidate pSSR sites, SSR markers were designed by Primer3 (Rozen *et al.* 2000, Koressaar *et al.* 2007, Untergasser *et al.* 2012). Genomic DNAs of seven *P. japonicum* collections were used as templates for PCR validation of SSR markers. PCR amplification was carried out in a 25-μL reaction volume containing the following components: 20 ng of DNA template, 10 μM of primer set, 5 mM of dNTP, and one unit Taq DNA polymerase (Vivagen, Seongnam, Korea). For six primer sets, PjSSR01 to PjSSR06, the amplification condition was as follows: 5 minutes at 94°C, 35 cycles of 94°C 20 s, 58°C 20 s, and 72°C 20 s and then 72°C for 7 min. For other primer sets, PjSSR07 to PjSSR10, the amplification condition was the same as above except annealing condition of 54°C for 20 s. PCR products were then separated by 9% polyacrylamide gel electrophoresis for identification of polymorphisms. Gel was stained with ethidium bromide and visualized under UV lamps for manual genotyping.

RESULTS

Identification of pSSR using two WGS in *P. japonicum*

Diverse germplasm for *P. japonicum* which is an indigenous plant in Korea were collected. Among the collections, WGS sequencing was carried out for two individuals, PJ#1 and PJ#2, which were collected from Jeju Island and Geumo Island, respectively. Raw WGS of PJ#1 and PJ#2 were approximately 1.1 Gbp and 1.2 Gbp, respectively. After sequence quality filtering, a total amount of 870 Mb and 980 Mb of WGS data remained for PJ#1 and PJ#2, respectively (Table 2-1). MiSeq sequences usually provide average 300 bp for forward and reverse sequence of each PE read. Our PE library was constructed from 500 bp insert size DNA fragments. To obtain relatively long reads and reduce the redundant finding from same sequences derived from forward and reverse sequence of same PE read, forward and reverse sequences were assembled for each PE reads to make PE-joined contigs (contigs) based on the overlapped sequences. The high quality 3.4 million reads were reduced to 1.6 M contigs by pair joining. Using the contigs, following steps for *in silico* identification of pSSRs were conducted. First, SSR motif from PJ#1 was selected through MISA program. Second, considerably non-redundant contigs were chosen as candidates by sequence clustering because high copy reads might be derived from repetitive sequence regions. SSR motifs were identified from 25,814 contigs, of which 12,206 were chosen from the solitary clustered contigs (Table 2-1). To identify pSSR motifs between PJ#1 and PJ#2, WGS reads of PJ#2 were aligned onto 12,206 contigs of

Table 2-1. Sequencing status of two *P. japonicum* accessions.

Contents of phase	*P. japonicum* acc. #1 (PJ#1)		*P. japonicum* acc. #2 (PJ#2)	
	reads	basepairs	reads	basepairs
Raw data	3,550,678	1,063,845,547	3,937,550	1,183,084,443
Trimmed data	3,438,470	869,332,349	3,817,637	978,732,808
Contigs assembly & selection	1,646,551	595,380,248		
SSR containing contigs	25,814	9,891,314		
Primary filtered contigs	12,206	4,565,371		
Secondary filtered contigs	4,698	1,878,413		
pSSR containg contigs	452	179,609		

PJ#1. During alignment, the 1,376 contigs with high mapping depth (≥10) and 6,132 contigs without any counterpart mapped reads were excluded. The remaining 4,698 contigs were then compared with SSR motif of PJ#2 and only 452 contigs contains pSSR motif were selected for further analysis. Among them, 371, 60, 21, and 7 contigs have di-, tri-, tetra- and more than penta-nucleotide SSR motifs, respectively (Fig. 2-1). Among the 452 pSSR contigs, five contigs had more than 2 SSR motifs.

Validation of pSSR and application for *P. japonicum* germplasm

To validate the pSSRs, ten contigs were randomly chosen and were designed primer pairs to amplify the SSR regions using Primer3 (Rozen *et al.* 2000, Koressaar *et al.* 2007, Untergasser *et al.* 2012) (Table 2-2). Five more *P. japonicum* plants (PJ#3-7) were included for genotyping each collections using the primer pairs. All ten primer sets (PjSSR01 ~ 10) successfully amplified each of *P. japonicum* collections. Nine of ten produced the estimated polymorphic SSR genotypes derived from the copy numbers of SSR units among *P. japonicum* accessions (Fig. 2-2). Meanwhile, one primer, PjSSR08, did not show any polymorphism among genotypes which differ with the expectation (Table 2-3). The genotypes of the *P. japonicum* accessions for each marker is represented in Table 2-3. The mean allele number of developed markers was around 2.7, whereas the mean genetic diversity and polymorphic information contents (PIC) were around 0.47 and 0.42, respectively. The developed pSSR may be valuable resources and the ten SSR markers will beneficial to evaluate and classify *P. japonicum* genotype for genetic study. Blastn analysis (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was carried out against non-redundant (nr)

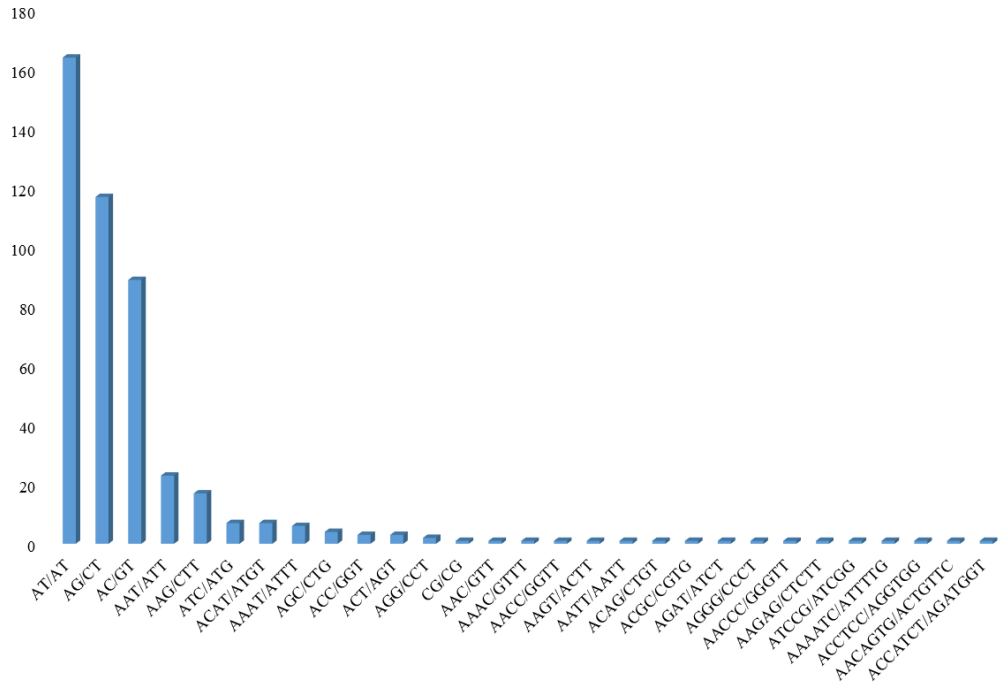

Fig 2-1. Classification of pSSR candidates based on the SSR motif in *P. japonicum*. The horizontal axis represent the type of SSR motif and vertical axis represent the number of SSR motifs.

Table 2-2. Primer information of ten SSR markers

Marker ID	SSR motif		Contig Length (bp)	Estimated PCR product (bp)		Primers	Discription of blastn (e-value)
	PJ#1	PJ#2		PJ#1	PJ#2		
PjSSR01	(GTAGATG)4	(GTAGATG)3	543	203	196	F AGGTTGTGGGTTCAATTCCCA R GTCAGAAGTTGGCCGTCAGA	
PjSSR02	(TTG)7	(TTG)5	521	246	240	F GTGTTATTCACTCTCCAAGGAAGG R TCGCGTCCAAACGAACCTTA	
PjSSR03	(TC)17	(TC)10	520	102	88	F TGGCGTTAACAATGATTCACCT R CTATGCTTTGCTGCTGCAGTT	XM_017381863.1 Daucus carota serine/threonine-protein kinase, mRNA (2e-14)
PjSSR04	(GAA)8	(GAA)5	519	213	204	F GAAGAAGTTGAAGGGGAGGGT R CGTTCTCTCAGTCCGCTCATT	XM_014765223.1 Glycine max receptor-like protein kinase 2-like, transcript variant X1, mRNA (5e-10)
PjSSR05	(AGAAG)5	(AGAAG)4	515	220	215	F TGGTGAACGACGGAGAAGTG R CTTGCTGACATGGCGGATTT	XM_017363412.1 Daucus carota cytochrome P450 84A1-like, mRNA (1e-16)
PjSSR06	(GGGA)5	(GGGA)4	514	137	133	F GCGGAAATGATGGTGGTTGG R AGATAGATGGTCCCAGCCCA	
PjSSR07	(GTAA)8	(GTAA)7	509	245	241	F AAACCGTTTTGTCCCCACTT R TGCTATTTGGTTGAGCTTTTTGGT	
PjSSR08	(TAA)6	(TAA)8	541	257	263	F TGGGCTCACATCAACCAACT R TCGAGCTCTCTCGGAATAGA	XM_010656117.2 Vitis vinifera uncharacterized, transcript variant X2, mRNA (4e-04)
PjSSR09	(TA)11	(TA)7	525	244	236	F ACACACAAATAGATAGACACGCTG R CCGAGTCTTTCTCGCAGGTT	
PjSSR10	(AGA)5	(AGA)3	517	290	284	F GAGTGATGGGAGAGGAAAGCAG R TCTCTGGAGCTTTGGAAACCAT	

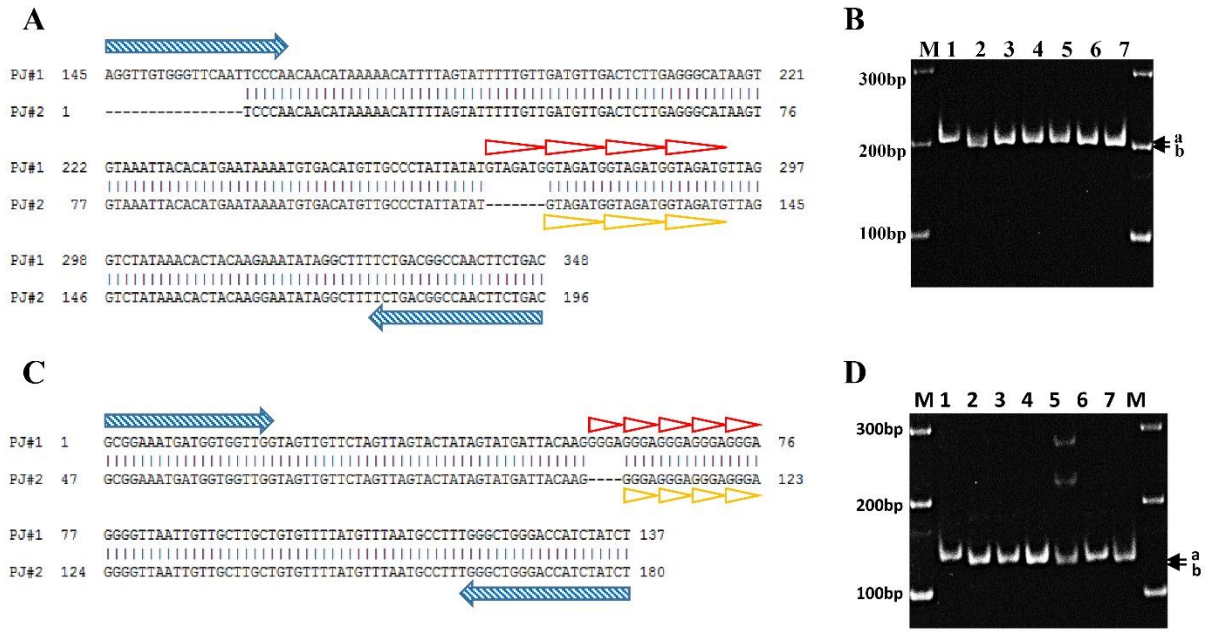

Figure 2-2. *In silico* prediction and PCR validation of PjSSR01 and PjSSR06 primers. **(A, C)** Sequence alignment between contigs of PJ#1 and the counterpart WGS reads of PJ#2 and the primer regions, PjSSR01 and PjSSR06, respectively. Blue arrow pointing left and right indicate the forward and reverse primers, and the colored triangle boxes, red and yellow, represent the SSR motifs for PJ#1 and PJ#2, respectively. **(B, D)** Gel electrophoresis of PCR products for PjSSR01 and PjSSR06, respectively. The number 1 to 7 indicate seven *P. japonicum* accessions, PJ#1 to PJ#7, respectively. M indicates 100-bp ladder. Arrows with alphabets (a,b) indicate different alleles among seven collections.

Table 2-3. Genotypes and allele diversity of 10 markers among seven collections of *P. japonicum*

Primers	PJ#1	PJ#2	PJ#3	PJ#4	PJ#5	PJ#6	PJ#7	No. of alleles	Genetic diversity[a]	PIC[b]
PjSSR01	a	b	a	a	a	a	b	2	0.408	0.325
PjSSR02	a	b	b	b	b	b	c	3	0.449	0.406
PjSSR03	a	c	b	c	c	c	c	3	0.449	0.406
PjSSR04	b	a	d	a	a	c	c	3	0.694	0.641
PjSSR05	a	b	a	c	c	a	b	3	0.653	0.580
PjSSR06	a	b	b	b	b	a	a	2	0.490	0.370
PjSSR07	b	c	d	c	c	c	a	4	0.612	0.570
PjSSR08	a	a	a	a	a	a	a	1	0.000	0.000
PjSSR09	a	b	b	b	b	b	b	2	0.245	0.215
PjSSR10	a	a	b	c	b	c	d	4	0.735	0.685
Mean								2.7	0.473	0.420

[a]Genetic diversity is the probability to show difference between two randomly chosen alleles from the population. [b]PIC: polymorphism information content. Genetic diversity and PIC values were calculated based on genotype of individuals by PowerMarker ver. 3.0 (Liu *et al.* 2005).

nucleotide sequence database and identified that four of ten *P. japonicum* primer regions were similar to genic regions of other plant species (Table 2-2).

DISCUSSION

Resource plants have a variety of useful potentials and some resource plants can be developed as minor food crops or pharmaceutical materials. Minor crops are crops with high value but not widely studied, hence it is also known as an 'orphan crops' or 'underutilized crops'. In addition, genomic or genetic information of minor crops is rarely reported. SSR marker is one of the most applicable markers used the genetic evaluation and utilization of crops due to its high-reproducibility, hyper-variable and co-dominant nature (Park *et al.* 2009). Therefore, many efforts were conducted to develop SSR markers by construction and sequencing of SSR-rich genomic libraries for the minor crops and by utilization of EST sequence (Choi *et al.* 2011, Izzah *et al.* 2014, Kim 2012). However, previous researches required the time consuming wet experiments to find polymorphic SSR markers due to low polymorphism rate (with less than 5 % of success rate) from the SSR candidates because they utilized sequence harboring the SSR motif from one genotype (Choi *et al.* 2011, Izzah *et al.* 2014, Kim 2012). In this study, two WGSs were explored to discover polymorphic SSR markers for the valuable resource plants *P. japonicum* as leafy vegetable and functional foods. Moreover, the designed SSR primer sets were confirmed and could be successfully applied in other collections. Ten SSR primers (PjSSR01~10) were successfully amplified, with high polymorphism in all collections (Table 2-2 and 2-3). Some showed difference with our expectation. The PCR products of both PjSSR04 and PjSSR10 resulted to a smaller product size than our

estimated sizes and PjSSR08 did not show any polymorphism. Despite the unexpected results, however, our SSR markers displayed polymorphic amplicons with high success rate (9 out of 10). The 452 pSSR candidates will provide fundamental information on evaluation of useful and fundamental genetic resources for breeding in *P. japonicum* (Table 2-2).

REFERENCES

Abdelkrim J, Robertson B, Stanton JA, Gemmell N. 2009. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques 46: 185-192.

Agarwal M, Shrivastava N, Padh H. 2008. Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep. 27: 617-631.

Allen G, Flores-Vergara M, Krasynanski S, Kumar S, Thompson W. 2006. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nature Protoc. 1: 2320-2325.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

Amano T. 1987. Ryukyuretto syokubutu hougensyu (The botanical dialects in the Ryukyus). Shinseitosho Shuppan: Naha

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Choi HI, Kim NH, Kim JH, Choi BS, Ahn IO, Lee JS and Yang TJ. 2011. Development of reproducible EST-derived SSR markers and assessment of genetic diversity in *Panax ginseng* cultivars and related species. J. Ginseng Res. 35: 399.

da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FI, Costa de Oliveira A.

2008. SSR Locator: Tool for simple sequence repeat discovery integrated with primer design and PCR Simulation. Int. J. Plant Genomics 2008: 412696.

Grover A, Sharma PC. 2016. Development and use of molecular markers: past and present. Crit. Rev. Biotechnol. 36: 290-302.

Han BX, Yuan Y, Huang LQ, Zhao Q, Tan LL, Song XW, He XM, Xu T, Liu F, Wang J. 2017. Specific PCR identification between *Peucedanum praeruptorum* and *Angelica decursiva* and identification between them and adulterant using DNA barcode. Pharmacogn. Mag. 13: 38.

Hiroe M, Constance L. 1958. Umbelliferae of Japan. University of California publications in botany 144.

Hisamoto M, Kikuzaki H, Ohigashi H, Nakatani N. 2003. Antioxidant compounds from the leaves of *Peucedanum japonicum* Thunb. J. Agric. Food Chem. 51: 5255-5261.

Izzah NK, Lee J, Jayakodi M, Perumal S, Jin M, Park BS, Ahn K, Yang TJ. 2014. Transcriptome sequencing of two parental lines of cabbage (*Brassica oleracea* L. var. capitata L.) and construction of an EST-based genetic map. BMC Genomics 15: 149.

Jones AG, Small CM, Paczolt KA, Ratterman NL. 2010. A practical guide to methods of parentage analysis. Mol. Ecol. Resour. 10: 6-30.

Kang SY, Oh TW, Kim JW, Park YK. 2013. Effect of the water extract of *Peucedani Japonici* Radix on ovalbumin-induced allergic asthma in mice. Kor. J. Herbology 28: 1-7.

Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, Makova KD. 2010.

What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. Genome Biol. Evol. 2: 620-635.

Kim K, Lee SC, Lee J, Lee HO, Joh HJ, Kim NH, Park HS, Yang TJ. 2015. Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. PLoS ONE 10: e0117159.

Kim K, Lee SC, Lee J, Yu Y, Yang K, Choi BS, Koh HJ, Waminal NE, Choi HI, Kim NH, Jang W, Park HS, Lee J, Lee HO, Joh HJ, Lee HJ, Park JY, Perumal S, Jayakodi M, Lee YS, Kim B, Copetti D, Kim S, Kim S, Lim KB, Kim YD, Lee J, Cho KS, Park BS, Wing RA, Yang TJ. 2015. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. Sci. Rep. 5: 15655.

Kim NH, Choi, HI, Ahn, IO, Yang, TJ. 2012. EST-SSR marker sets for practical authentication of all nine registered ginseng cultivars in Korea. J. Ginseng Res. 36: 298-307.

Koressaar T, Remm M. 2007. Enhancements and modifications of primer design program Primer3. Bioinformatics 23: 1289-1291.

Liu K, Muse SV. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21: 2128-2129.

Ma KH, Kim NS, Lee GA, Lee SY, Lee JK, Yi JY, Park YJ, Kim TS, Gwag JG, Kwon SJ 2009. Development of SSR markers for studies of diversity in the genus *Fagopyrum*. Theor. Appl. Genet. 119: 1247-1254.

Meng W, Fei X, Peng Y, Duan XY, Zhou YL, Shen CY, Zhang GZ, Wang BT. 2014.

Development of SSR markers for a phytopathogenic fungus, *Blumeria graminis* f. sp. tritici, Using a FIASCO Protocol. J. Integr. Agr. 13: 100-104.

Metz S, Cabrera JM, Rueda E, Giri F, Amavet P. 2016. FullSSR: Microsatellite Finder and Primer Designer. Adv. bioinformatics 2016: 4.

Mittal N, Dubey AK 2009. Microsatellite markers-A new practice of DNA based markers in molecular genetics. Phcog. Rev. 3: 235.

Buwalda P. 1949. Umbelliferae In: van Steenis, C. G. G. J. (ed). Flora Malesiana ser. I 4: 112-140.

Park YJ, Lee JK, Kim NS. 2009. Simple sequence repeat polymorphisms (SSRPs) for evaluation of molecular diversity and germplasm classification of minor crops. Molecules 14: 4546-4569.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. 132: 365-386.

Seo A, Watanabe M, Hotta M, Murakami N. 2001. Allozyme variation of the three varieties of *Peucedanum japonicum* thunb. in Japan. APG: Acta Phytotaxonomica et Geobotanica 52: 135-148.

Seol YJ, Lee TH, Park DS, Kim CK. 2016. NABIC: A new access portal to search, visualize, and share agricultural genomics data. Evol. Bioinform. Online 12: 51-58.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3--new capabilities and interfaces. Nucleic Acids Res. 40: e115.

Varshney RK, Graner A, Sorrells ME. 2005. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 23: 48-55.

Varshney RK, Nayak SN, May GD, Jackson SA. 2009. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 27: 522-530.

CHAPTER III

Rapid amplification of four retrotransposon families promoted speciation and genome size expansion in the genus *Panax*

ABSTRACT

Genome duplication and repeat multiplication contribute to genome evolution in plants. Our previous work identified a recent allotetraploidization event and five high-copy LTR retrotransposon (LTR-RT) families *PgDel, PgTat, PgAthila, PgTork,* and *PgOryco* in *Panax ginseng*. Here, using whole-genome sequences, major repeats in five *Panax* species were quantified and were investigated their role in genome evolution. The diploids *P. japonicus, P. vietnamensis,* and *P. notoginseng* and the tetraploids *P. ginseng* and *P. quinquefolius* were analyzed alongside their relative *Aralia elata*. These species possess 0.8–4.9 Gb haploid genomes. The *PgDel, PgTat, PgAthila,* and *PgTork* LTR-RT superfamilies accounted for 39–52% of the *Panax* species genomes and 17% of the *A. elata* genome. *PgDel* included six subfamily members, each with a distinct genome distribution. In particular, the *PgDel1* subfamily occupied 23–35% of the *Panax* genomes and accounted for much of their genome size variation. *PgDel1* occupied 22.6% (0.8 Gb of 3.6 Gb) and 34.5% (1.7 Gb of 4.9 Gb) of the *P. ginseng* and *P. quinquefolius* genomes, respectively. Our findings indicate that the *P. quinquefolius* genome may have expanded due to rapid *PgDel1* amplification over the last million years as a result of environmental adaptation following migration from Asia to North America. GP of *PgDel2* is 2.5% in tetrapolids but approximately 4.8% in three diploids. fluorescence *in situ* hybridization analysis show that *PgDel1* spread all the chromosomes of *Panax* species and *PgDel2* occupy all chromosomes of three diploids but half of tetraploids that support our GP

calculation.

Keywords: *Panax* genus, long terminal repeat retrotransposons (LTR-RTs), genome expansion, allotetraplodization, genome evolution.

INTRODUCTION

Nuclear genome sizes in flowering plants are diverse, and can vary over 2,400-fold, ranging from 63 Mb in *Genlisea margaretae* to 149 Gb in *Paris japonica* (Pellicer 2010). This dramatic genome size variation is attributed to both whole-genome duplication and accumulation of repeated sequences, or repeats (SanMiguel*, et al.* 1998, SanMiguel*, et al.* 1996, Wendel 2000). During the diploidization process following genome duplication, euchromatic DNA is usually reduced by deletion of unnecessary paralogous regions(Leitch*, et al.* 2004, Yang*, et al.* 2006) while heterochromatic DNA is often expanded by species-specific multiplication of repeats (Lim*, et al.* 2007). Repeats are categorized into two major types: tandem repeats (TRs) and transposable elements (TEs) (Richard*, et al.* 2008). TRs exist in a head-to-tail arrangement in distinct chromosomal regions, generally found at centromeric, subtelomeric, and telomeric regions (Csink*, et al.* 1998, Lim*, et al.* 2007). By contrast, TEs are dispersed throughout the genome. TEs are classified based on their transposition mechanisms as class I (copy-and-paste) or class II (cut-and-paste). Class I TEs include the class I.1 LTR-retrotransposons (LTR-RTs) and the class I.2 non-LTR retrotransposons, whereas class II TEs include DNA transposons (Piégu*, et al.* 2015). Repeats play important roles in gene regulation, evolution, and adaptation (Feschotte*, et al.* 2007, Oliver 2013, Volff 2006).

The family Araliaceae is composed of approximately 55 genera and 1,500 species, which include many valuable medicinal and ornamental plants (Wen 2001). Within this family, the genus *Panax* contains economically important medicinal plants including the

diploids *P. japonicus*, *P. vietnamensis*, and *P. notoginseng* ($2n = 2x = 24$), and the tetraploids *P. quinquefolius* and *P. ginseng* ($2n = 4x = 48$). These five species are perennial and absolute shade plants that have been used for medicinal purposes in Asia and North America because of their beneficial effects on human health (Yun 2001). Although *Panax* species display relatively limited morphological diversity, particularly at aerial part, their genome sizes vary from 2.02 Gb (*P. vietnamensis*) to 4.9 Gb (*P. quinquefolius*) (Obae*, et al.* 2012, Pan*, et al.* 2014). Several genomic studies have been conducted to elucidate the genome structure, function, and evolution of genomes in the *Panax* genus (Bai*, et al.* 1997, Ho*, et al.* 2002, Hong*, et al.* 2004, Jang*, et al.* 2017, Jayakodi*, et al.* 2014, Jayakodi*, et al.* 2015, Kim*, et al.* 2015, Kim*, et al.* 2017, Kim 2012).

Recently, the evolution of five *Panax* species was described by comparative analysis of complete chloroplast genome sequences and ribosomal DNA (Kim*, et al.* 2017). The major repeats were also characterized that occupied more than 35% of the *P. ginseng* genome, namely five high-copy LTR-RT families (Choi*, et al.* 2014, Jang*, et al.* 2017). In this study, I aimed to explore the role of major repeats in the evolution of the *Panax* genus, which shows large genome size variation. Accordingly, a reliable quantification method was established for major repeats within a genome using low-coverage whole-genome sequences and quantified each of these LTR-RTs in the genomes of five *Panax* species. Our comparative analysis revealed dynamic impacts of these major repeats on genome size variation, speciation, and evolution in the *Panax* genus.

MATERIAL AND METHODS

Plant materials, genomic DNA isolation, and Illumina sequencing

Eleven *P. ginseng* cultivars as well as *P. quinquefolius*, *P. notoginseng*, *P. japonicus*, *P. vietnamensis* and *A. elata* were used for genomic DNA preparation and sequencing (Table 1-4 and 1-6). *P. ginseng* cv. Chunpoong was used as a representative for GP estimation in current analysis. Leaf tissue for the above species, apart for *P. notoginseng*, *P. japonicus*, *P. vietnamensis*, and *A. elata*, was obtained from the ginseng farms of Seoul National University and Korean Ginseng Corporation (http://www.kgc.or.kr). *A. elata* and *P. vietnamensis* leaf tissue was collected from Susinogapy Corporation (http://www.susinogapy.com), Korea, and Da Lat City, Tay Nguyen Institute of Scientific Research, Vietnam, respectively. *P. notoginseng* and *P. japonicus* leaf tissue was collected from Dafang Country, Guizhou province, and Enshi County, Hubei province, China, respectively.

Genomic DNA was extracted using a modified cetyltrimethylammonium bromide (CTAB) method (Allen 2006). All genomic libraries were prepared according to the recommended Illumina paired-end standard protocol (http://www.illumina.com). The whole genomes of those plants listed in Table 3-4 and 3-6 were sequenced using an Illumina genome analyzer at the National Instrumentation Center for Environmental Management (NICEM: http://nature.snu.ac.kr/kr.php) and LabGenomics

(www.labgenomics.co.kr/), South Korea. All sequence data were uploaded to the National Agricultural Biotechnology Information Center (http://nabic.rda.go.kr) (Table 3-4 and 3-6) (Seol, *et al.* 2016).

Major repeat sequences of *Panax ginseng.*

In our previous study, the major repeats of *P. ginseng* were described including 11 LTR-RTs and two tandem repeat sequences (Pg167TR and 45S rDNA), which occupy more than one third of the genome (Choi, *et al.* 2014, Jang, *et al.* 2017, Kim, *et al.* 2015). These reference sequences were used as queries to estimate their abundance in *Panax* and *Aralia* genomes. Most of LTR-RTs of major repeats analyzed in this work have a complete structure that includes both flanking LTRs and inter-LTR domains, with the exception of *PgAthila* that has one LTR (Choi, *et al.* 2014). The 45S rDNA of *P. ginseng* was used as a representative rDNA sequence for all *Panax* species (Table 3-1).

Quantification of major repeats using WGS

The GP of each repeat was quantified by masking nucleotides of WGS reads into the representative repeat sequence using RepeatMasker (ver. 4.0.6) (Smit, *et al.*). WGS reads were trimmed based on their quality score (minimum quality score: ≥20) using the software Trimmomatic ver. 0.33 (Bolger 2014). WGS reads were directly surveyed for homology to each repeat using RepeatMasker, using the slow search parameters and option that does

not mask low complexity DNA or simple repeats (applying '-s –no low'). Homologous nucleotides were masked and the amounts of masked nucleotides were counted to calculate GP for each repeat. The RepeatMasker-based genomic proportion (R-GP) was calculated as the proportion that masked nucleotides make of total nucleotides in each data set: R-GP (%) = (masked read length / total read length) × 100. The actual amounts of each repeat in the genome was estimated based on the R-GP and the size of the genome: Repeat amount = (R-GP / 100) × genome size) (Fig. 4B). The mapping-based GP (M-GP) and copy number of *PgDel1_1* LTR-RTs were estimated using CLC mapper ver. 4.21.104315 (CLC Inc, Aarhus, Denmark) with the parameters of minimum 50% fraction of the read and 80% similarity (Fig. 1-4, 1-7 and 1-8).

Fluorescence in situ hybridization (FISH) analysis

Preparation of *P. notoginseng*, *P. ginseng*, and *P. quinquefolius* chromosome spreads and FISH procedures were performed according to dual-color FISH analysis protocols (Waminal*, et al.* 2012). Briefly, root tips were treated with 2 mM 8-hyroxyquinoline, fixed with Carnoy's solution, and were enzymatically digested with pectolytic enzyme solution (2% Cellulase R-10 (C224, Phytotechnology Laboratories) and 1% Pectolyase Y-23 (P8004.0001, Duchefa)) in 100 mM citrate buffer) for 1 h. Root tips were then squashed onto slides pre-cleaned with 70% ethanol. Air-dried slides were fixed in 2% formaldehyde for 5 min and dehydrated with a series of ethanol treatments (70%, 90%, and 100%) (Vrana 2012). *PgDel1* and *PgDel2* probes were obtained by PCR amplification using *P. ginseng*

genomic DNA and primers detailed in our previous study (Choi*, et al.* 2014). *PgDel1* was labeled with Cy5-dUTP (Jena Bioscence), whereas *PgDel2* was labeled with Diethyl amino coumarin-5-dUTP (NEL455001EA, Perkin Elmer) or Alexa Fluor 488-5-dUTP (C11397, Life Technologies). Images were captured using an Olympus BX53 epifluorescence microscope equipped with a Leica DFC365 FS CCD camera, and processed using Cytovision version 7.2 (Leica Microsystems, Germany). Further image enhancements were performed using Adobe Photoshop CS6.

RESULTS

Whole genome sequence (WGS)-based quantification of major repeats in *P. ginseng*

In *P. ginseng*, 11 LTR-RT subfamilies contained within five superfamilies, namely *PgDel1–6*, *PgTat1* and *2*, *PgAthila*, *PgTork*, and *PgOryco*, and two tandem repeat sequences, namely Pg167TR and 45S rDNA (Table 3-1) were recently reported. These 13 repeats are high-copy, major repeats and are estimated to occupy more than 41% of the *P. ginseng* genome (Table 3-2, 3-3, 3-4, 3-5, and Fig. 3-1) (Choi*, et al.* 2014, Jang*, et al.* 2017, Kim*, et al.* 2015). Here, to quantify these major repeats in the WGS datasets of five *Panax* species was aimed. The amount of each repeat was determined by calculating its genomic proportion (GP) in each WGS, via quantification of homologous nucleotides in each WGS based on repeat masking using RepeatMasker (Smit*, et al.* 2013-2015). RepeatMasker-based GP (R-GP) estimation and the quantification of each major repeat were validated using various WGS data sets. Repeat quantification in WGS data sets with different genome coverages (0.00005–10x), as well as in WGS data sets from different libraries using *P. ginseng* cv. 'Chunpoong' and in WGS data sets from different ginseng cultivars were then compared. The reproducibility of R-GP estimation for each of 13 repeats was evaluated in each WGS (Table 3-2, 3-3, 3-4, 3-5, and Fig. 3-1).

The R-GP of each repeat displayed little variation in datasets of the same WGS that represented nine different genome coverages, and low variation in datasets from four

different WGS libraries created using the same ginseng cultivar (Supplementary Table 3-2 and 3-3). Furthermore, low R-GP variation for repeats was observed across WGS datasets of 11 ginseng cultivars, with the 13 repeats displaying a R-GP of 41.0–46.3% (Supplementary Table 3-5 and Fig. 3-1). High-copy LTR-RTs showed little variation, while low-copy LTR-RTs occupying less than 1% GP and tandem repeat units such as Pg167TR and 45S rDNA showed relatively high variation (Supplementary Table 3-2 and 3-3). The R-GP of *PgDel1* was 23–26% among 11 cultivars (Supplementary Table 3-5 and Fig. 3-1). Overall, our results showed that the R-GP estimation of major repeats was reliable.

Table 3-1. Summary of major repeat elements analyzed in this study.

Type	Name	Length (bp)	BAC seq. (Acc. No.)	Position in BAC seq.
Ty3/Gypsy	*PgDel1_1*	10,039	KF357944	76821-80102, 82966-89722
	PgDel1_2	10,120	KF357944	15991-26110
	PgDel1_3	9,477	KF357943	23334-23362, 32933-42380
	PgDel1_4	8,004	KF357943	23363-25995, 27556-32926
	PgDel1_5	7,714	KF357943	46781-51181, 74068-75415, 82113-83407, 95702-96371
	PgDel2	12,515	KF357942	89398-101912
	PgDel3	11,809	KF357944	6065-7415, 36961-47418
	PgDel4	11,050	KY513615	75,443-76,688, 77,213-87,016
	PgDel5[a]	12,860	KY513617	1-10587
	PgDel6	12,252	KY513617	20,535-32,786
	PgTat1_1	22,881	KF357943	51182-74062
	PgTat1_2	12,289	KF357943	83408-95696
	PgTat2	10,965	KF357942	41381-52345
Ty1/Copia	*PgAthila*	9,893	KF357943	142151-152043
	PgTork	9,707	KF357944	9572-18462, 25797-26612
	PgOryco	7,772	KF357942	30843-32623, 35385- 41375
Tandem Repeat	Pg167TR[b]	1,577	KF357942	10074-11650
	45S rDNA[c]	5,877	KM036295	1-5877
Total	15ea	184,528		

[a]*PgDel5* has an incomplete right LTR, due to a shortage of BAC sequence; thus, complete structure of *PgDel5* was predicted using the *P. ginseng* draft genome sequences (data not shown) [b]*P. ginseng* tandem repeats (PgTR) were composed of 9.4 copy number units of 167 bp consensus sequences with 87 % similarity and 1 indels among the units. [c]The 45S rDNA sequence includes only transcriptional unit sequences (18S-ITS1-5.8S-ITS2-26S)

Table 3-2. Summary of GP calculation for major repeats in various genome coverage data sets of *P. ginseng* cv. Chunpoong

Amount of WGS (Mbp)	0.18	0.36	3.6	36	360	1,800	3,600	18,000	36,000	CV (%)
Genome coverage	0.00005x	0.0001x	0.001x	0.01x	0.1x	0.5x	1x	5x	10x	
PgDel1	21.82	23.79	24.1	23.75	24.11	21.9	24.23	24.09	24.06	3.23
PgDel2	2.09	2.27	2.51	2.71	2.62	2.45	2.64	2.65	2.65	5.66
PgDel3	3.17	2.84	2.51	2.52	2.61	2.53	2.59	2.6	2.6	4.04
PgTat1	9.93	6.36	5.92	6.05	5.89	6.56	6.05	6.04	6.03	3.75
PgTat2	0.54	0.64	0.64	0.7	0.7	0.9	0.72	0.72	0.72	11.29
PgAthila	0.54	1.34	1.43	1.32	1.44	1.47	1.45	1.43	1.43	3.8
PgTork	0.51	1.24	1.14	1.29	1.24	0.97	1.23	1.21	1.22	8.31
PgOryco	0	0.07	0.09	0.1	0.11	0.09	0.1	0.1	0.11	13.53
PgTR	1.57	1.18	1.07	1.11	1.2	2.06	1.16	1.19	1.21	25.29
45S rDNA	0.54	0.73	0.7	0.7	0.76	1.99	0.66	0.75	0.75	51.11
Total	40.71	40.47	40.09	40.24	40.68	40.92	40.85	40.79	40.77	0.75

0.1x means 0.1 x genome coverage data set. [a]SD: standard deviation. [b]CV: coefficient of variation. CV value were calculated by 0.0001x to 10x, except for 0.00005x.

Table 3-3. Summary of GP calculation for major repeats using various WGS libraries of *P. ginseng* cv. Chunpoong

GP	Lib. #1	Lib. #2	Lib. #3	Lib. #4	Average	SD	CV (%)
PgDel1	24.06	25.02	25.03	26.6	25.18	1.05	4.18
PgDel2	2.65	2.97	2.63	2.61	2.72	0.17	6.29
PgDel3	2.6	2.97	2.55	2.38	2.62	0.25	9.48
PgTat1	6.03	8.12	5.62	5.6	6.34	1.20	18.93
PgTat2	0.72	1.04	0.56	0.56	0.72	0.23	31.37
PgAthila	1.43	1.69	1.22	1.21	1.39	0.22	16.20
PgTork	1.22	1.09	1.38	1.58	1.32	0.21	16.03
PgOryco	0.11	0.08	0.13	0.15	0.12	0.03	25.48
Pg167TR	1.21	2.56	0.75	0.77	1.32	0.85	64.47
45S rDNA	0.75	1.11	0.29	0.43	0.64	0.37	56.78
Total	40.77	46.65	40.16	41.89	42.37	2.94	6.95

Table 3-4. Summary of WGS data of eleven cultivars of *Panax ginseng* used for major repeats survey

Speices name	Raw data		Trimmed data		Genome Coverage (x)	NABIC accession number
	Reads	Total bases	Reads	Total bases		
Panax ginseng cv. CS	21,757,942	2,197,552,142	19,589,258	1,950,766,361	0.54	NN-0141-000001
Panax ginseng cv. GO	17,477,870	1,765,264,870	15,600,912	1,557,424,038	0.43	NN-0140-000001
Panax ginseng cv. GU	14,469,452	1,461,414,652	12,423,896	1,234,138,042	0.34	NN-0143-000001
Panax ginseng cv. HS	19,840,654	2,003,906,054	18,238,910	1,821,128,340	0.51	NN-0142-000001
Panax ginseng cv. JK	15,461,684	1,561,630,084	13,697,724	1,361,616,353	0.38	NN-2381-000001
Panax ginseng cv. SH	18,255,214	1,843,776,614	16,320,609	1,623,507,629	0.45	NN-0190-000001
Panax ginseng cv. SO	16,863,306	1,703,193,906	15,311,929	1,526,296,690	0.42	NN-0192-000001
Panax ginseng cv. SP	17,406,574	1,758,063,974	15,891,182	1,584,152,283	0.44	NN-0191-000001
Panax ginseng cv. SU	19,081,012	1,927,182,212	17,394,151	1,734,358,202	0.48	NN-0194-000001
Panax ginseng cv. YP	19,000,000	1,919,000,000	17,348,966	1,739,503,772	0.48	NN-0135-000001
10 cultivars of *P. ginseng*	179,613,708	18,140,984,508	161,817,537	16,132,891,710	0.45	

Table 3-5. Summary of GP calculation for major repeats using 11 *Panax ginseng* cultivars

GP	CP	CS	GO	GU	HS	JK	SH	SO	SP	SU	YP	Average	SD	CV(%)
PgDel1	22.62	24.44	24.86	25.74	22.85	24.02	23.84	25.38	24.73	25.23	23.41	24.28	1.03	4.25
PgDel2	1.45	1.65	1.57	1.68	1.56	1.61	1.67	1.65	1.74	1.68	1.58	1.62	0.08	4.88
PgDel3	2.07	2.59	2.5	2.64	2.36	2.56	2.59	2.53	2.72	2.63	2.14	2.48	0.21	8.41
PgDel4	0.7	0.82	0.82	0.82	0.77	0.81	0.85	0.83	0.87	0.83	0.72	0.80	0.05	6.54
PgDel5	0.93	0.93	0.9	0.94	0.85	0.91	0.92	0.93	0.93	0.92	0.99	0.92	0.03	3.60
PgDel6	1.77	2	1.92	2.03	1.85	1.96	2.04	1.98	2.05	2.02	1.86	1.95	0.09	4.70
PgTat1	6.03	7.05	6.48	6.96	6.49	6.77	7.42	6.91	7.65	6.96	5.89	6.78	0.53	7.87
PgTat2	0.72	0.8	0.7	0.74	0.74	0.74	0.85	0.78	0.9	0.79	0.63	0.76	0.07	9.62
PgAthila	1.43	1.69	1.57	1.65	1.59	1.66	1.73	1.66	1.8	1.68	1.35	1.62	0.13	8.04
PgTork	1.22	1.17	1.18	1.18	1.08	1.12	1.12	1.21	1.16	1.22	1.43	1.19	0.09	7.67
PgOryco	0.11	0.1	0.1	0.1	0.09	0.1	0.09	0.1	0.1	0.1	0.14	0.10	0.01	13.13
Pg167TR	1.21	1.46	1.27	1.45	1.16	1.4	1.74	1.44	1.43	1.28	0.81	1.33	0.23	17.55
45S rDNA	0.75	0.37	0.24	0.21	0.28	0.5	1.2	0.54	0.2	0.55	0.15	0.45	0.31	68.39
Total	41.01	45.07	44.11	46.14	41.67	44.16	46.06	45.94	46.28	45.89	41.1	44.31	2.10	4.75

GP, genome proportion; CP: cv. Chunpoong, CS: cv. Chungsun, GO: cv. Gopoong, GU: cv. Gumpoong, HS: cv. Hwangsook, JK: cv. Jakyung, SH: cv. Sunhyang, SP: cv. Sunpoong, SU: cv. Sunwun, SW: cv. Sunwon and YP: cv. Yunpoong, WGS data of Supplementary Table S4 were used.

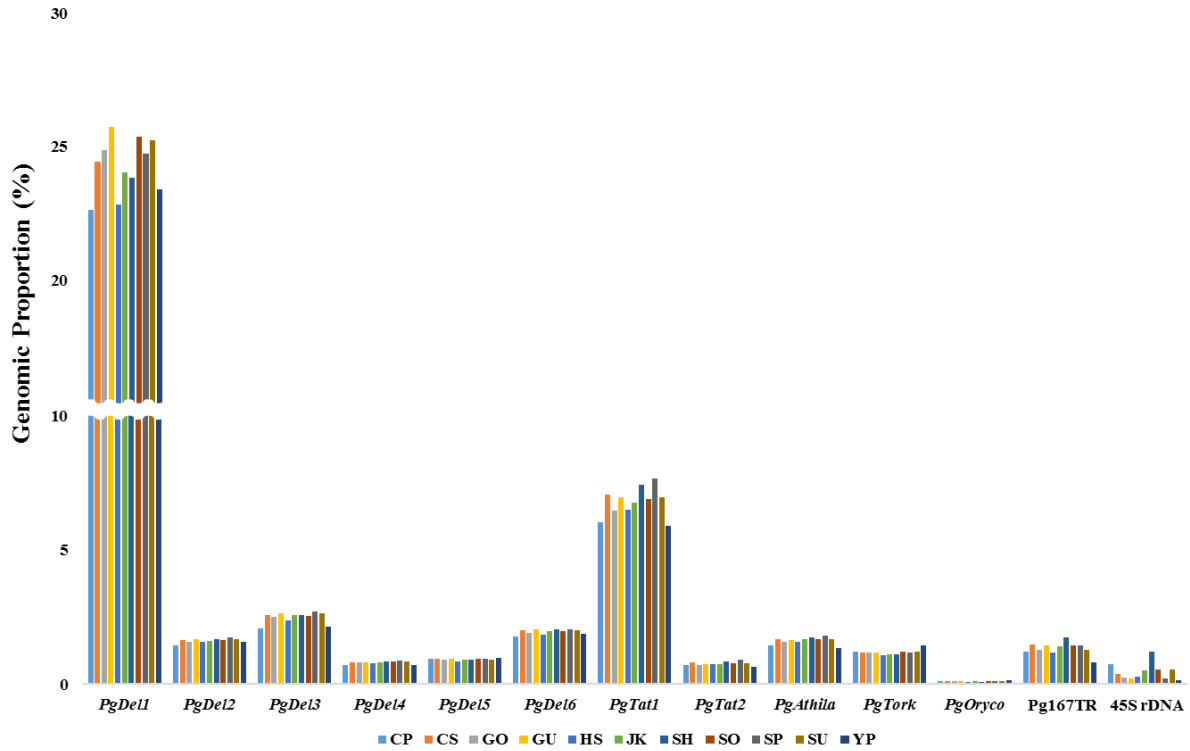

Figure 3-1. Genomic proportion (GP) of the major repeats in 11 cultivars of *P. ginseng*. GP of 13 repeats in the 11 cultivars of *Panax ginseng*. (CP: cv. Chunpoong, CS: cv. Chungsun, GO: cv. Gopoong, GU: cv. Gumpoong, HS: cv. Hwangsook, JK: cv. Jakyung, SH: cv. Sunhyang, SP: cv. Sunpoong, SU: cv. Sunwun, SW: cv. Sunwon and YP: cv. Yunpoong).

Genomic quantification of major repeats in five *Panax* species

The above WGS-based R-GP estimation to quantify the major repeats was used in the genomes of five *Panax* species alongside a species from a related genus (Table 3-6). The quantification of repeats using PE reads corresponding to 0.3–1.5x haploid genome equivalents for each species revealed a R-GP of 46%, 45%, 50%, 41%, 53%, and 17% in *P. japonicus*, *P. vietnamensis*, *P. notoginseng*, *P. ginseng*, *P. quinquefolius*, and *Aralia elata,* respectively (Fig. 3-3). Each individual major repeat possessed a similar R-GP in the five *Panax* species, whereas in *A. elata*, the R-GP was comparatively low. The *Ty3/Gypsy*-type LTR-RT families, such as *PgDel1-6*, *PgTat1-2*, and *PgAthila*, covered approximately 37.7–47.5% of the genomes. Among these, *PgDel1* had 22.6–34.5% R-GP in five *Panax* species but approximately 1% R-GP in *A. elata*. In particular, the larger genome of the *Panax* species in *P. quinquefolius* had a high amount of *PgDel1* elements, with a R-GP of 34.5% (Fig. 3-3).

The R-GP of *PgDel2*, *PgDel5*, *PgDel6*, and *PgTork* displayed large variation among the five *Panax* species (Fig. 3-3). *PgDel2* had 2.6–3.0% R-GP in the three diploid *Panax* species, and 1.5% and 1.4% R-GP in the two tetraploids *P. ginseng* and *P. quinquefolius*, respectively, which was approximately half of that measured in the diploids. *PgDel5* was more abundant in *P. notoginseng* and *A. elata* compared to that in other species. *PgDel6* had 4.3% and 5% R-GP in the two diploids *P. japonicus* and *P. vietnamensis*, respectively, whereas it had 1.5–2.4% R-GP in the remaining three *Panax* species. The R-GP of *PgTork* varied dynamically between *Panax* species (Fig. 3-3).

Table 3-6. Summary of WGS data of five *Panax* species and the related *A. elata* used for a survey of major repeats

Species	Chromosome number	Genome size (Gb)	NGS sequencing platform	Average Read length (bp)	Reads (M)[d]	Total bases (Mb)[e]	Genome Coverage (x)	NABIC accession number
P. ginseng	2n=48	3.6	HiSeq	101	36.2	3,605	1.00	NN-0076-000001
P. quinquefolius	2n=48	4.9	HiSeq	101	12.4	1,236	0.25	NN-0189-000001
P. notoginseng	2n=24	2.5	MiSeq	300	8.2	2,247	0.90	NN-1913-000001
P. japonicus	2n=24	~2.0[a]	MiSeq	300	8.3	2,271	1.14	NN-1914-000001
P. vietnamensis	2n=24	2.0	NextSeq	150	35.2	5,126	2.56	NN-1915-000001
A. elata	2n=24[b]	0.8[c]	HiSeq	101	40.4	4,052	2.50	NN-0919-000001
5 *Panax* species and 1 *Panax*-related species					140.7	18,536		

[a]Genome size was estimated in the present study. [b]Chromosome number was determined by DAPI (4′,6-diamidino-2-phenylindole) staining (Supplementary Fig. 1-2). [c]The genome size of *A. elata* was considered to be approximately 0.8 Gb in this study, based on the genome sizes of related species(Bai*, et al.* 2012). [d, e]Quality-controlled WGS reads were used in the current study.

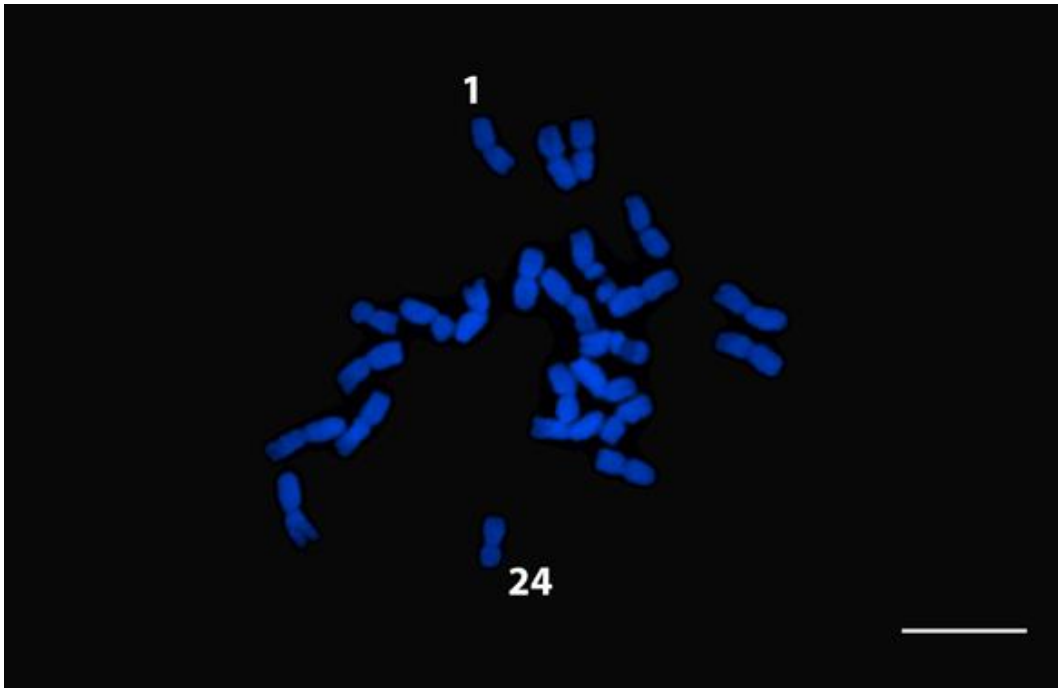

Figure 3-2. DAPI staining for confirmation of chromosome number in *Aralia elata*. Bar = 10 μm.

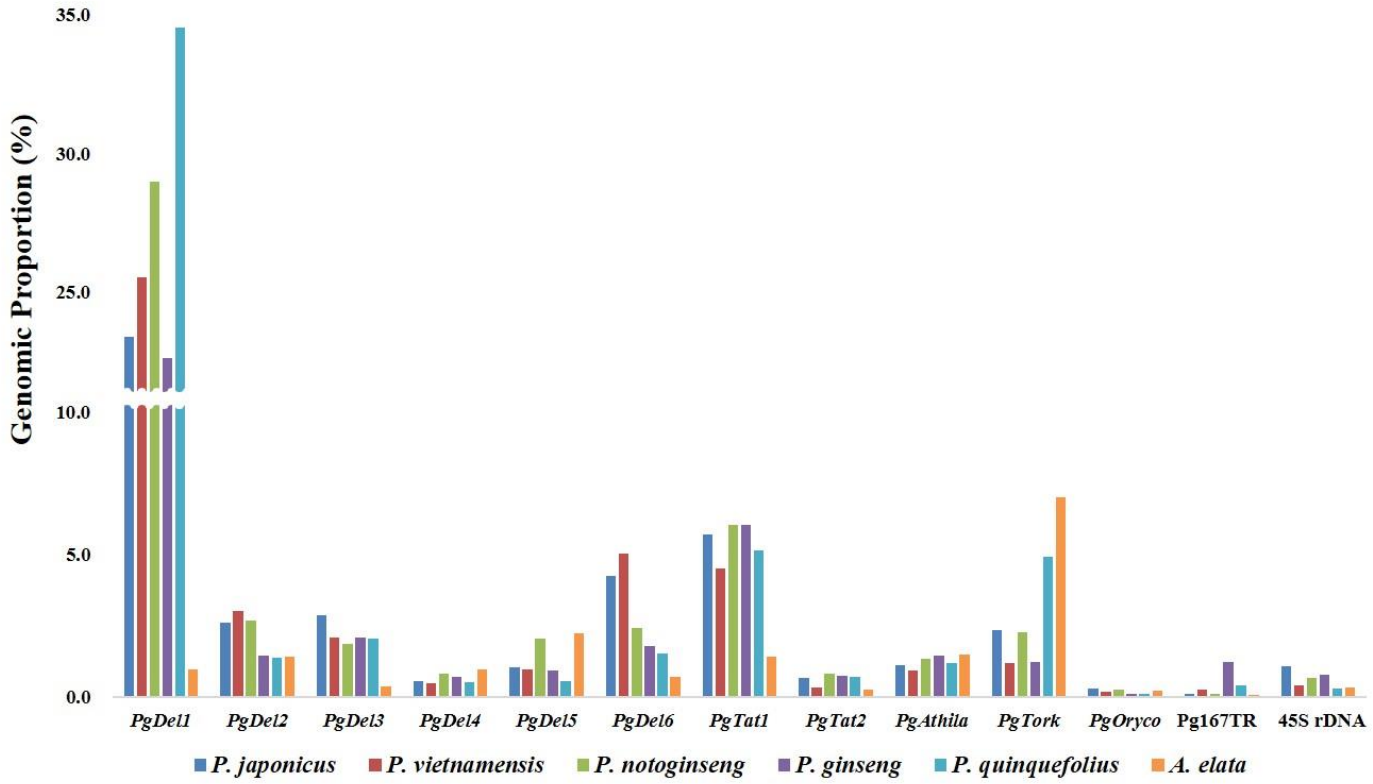

Figure 3-3. Genomic proportion of the major repeats in *Panax* species and a related species. Genomic proportion (GP) of 13 repeats in 6 *Panax* species and the related species *A. elata*.

Dynamics of the *PgDel1* subfamily members in *Panax* species

The structural dynamics of *PgDel1* subfamily members were analyzed in the *Panax* species. Five *PgDel1* subfamily members (*PgDel1_1–5*) were identified from three complete BAC clone sequences (GenBank accession nos. KF357943, KF357944, and KF357942) (Choi*, et al.* 2014). These five members displayed relatively complete structures including both LTRs and an inner sequence, although there were nested insertions caused by other repeats or subsequent deletion events. Inspection of the complete unit of these repeats, which was 7.7–10.1 kb, revealed an overall similarity in the large structural variations in the LTR regions. To estimate the distribution of *PgDel1* members in the *P. ginseng* genome, the 1x genome coverage Chunpoong WGS data was mapped onto the representative *PgDel1_1* element because of the well-conserved LTR domains of *PgDel1*. Mapping depth had a range of 111–157,407 with an average of 50,952 (mode and median values were 48,399 and 47,503, respectively) (Fig. 3-4).

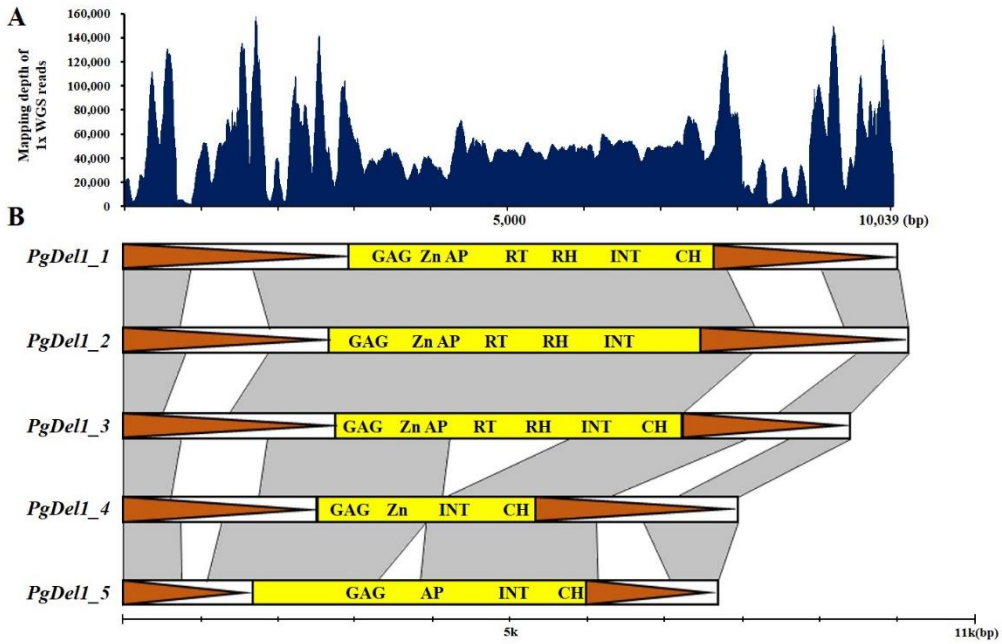

Figure 3-4. Structural characteristic of five *PgDel1* subfamily members. (A) Representation of the distribution of 1x WGS data of *P. ginseng* cv. CP. (B) Horizontal schematic diagrams of *PgDel1* subfamily members 1–5. Boxed orange triangles indicate LTR regions of *PgDel1*. Yellow boxes indicate the internal LTR-RTs domains of *PgDel1* detected in each subfamily member. (AP: aspartic protease, CH: chromodomain, GAG: capsid protein, INT: integrase, RH: RNase H, RT: reverse-transcriptase, and Zn: zinc knuckle). Homologous sequence were indicated as grey panels.

Cytogenomic mapping of *PgDel1* and *PgDel2* in three *Panax* species

To validate the R-GP variation identified via *in silico* analysis, the distribution patterns of *PgDel1* and *PgDel2* were analyzed by fluorescence *in situ* hybridization (FISH) using somatic metaphase chromosomes of three *Panax* species: *P. notoginseng*, as a representative of the three diploid *Panax* species, and the two tetraploids *P. ginseng* and *P. quinquefolius*. The *PgDel1* elements displayed high-density FISH signals throughout the chromosomes in all three *Panax* species (Fig. 3-5A, B, and C). The intensive FISH signal of *PgDel1* throughout the chromosome regardless of the ploidy level of the species it originated from supported our *in silico* analysis results, which estimated 23–35% R-GP for *PgDel1* in the *Panax* species (Fig. 3-3 and 3-5A-C).

PgDel2 had nearly two-fold greater R-GP values in the three diploid *Panax* species compared to the two tetraploids (Fig. 3-3). Consistent with this result, FISH analysis revealed different distribution patterns of *PgDel2* in diploid and tetraploid *Panax* species. *PgDel2* signal was localized to pericentromeric regions in all 24 chromosomes of the diploid *P. notoginseng*, whereas strong *PgDel2* signal was detected in half of the 48 chromosomes of both tetraploid *Panax* species (Fig. 3-5D, E, and F). In these tetraploids, *PgDel2* distribution was concentrated to the pericentromeric regions in *P. ginseng* chromosomes but was more broadly located in *P. quinquefolius* chromosomes (Fig. 3-5E and F).

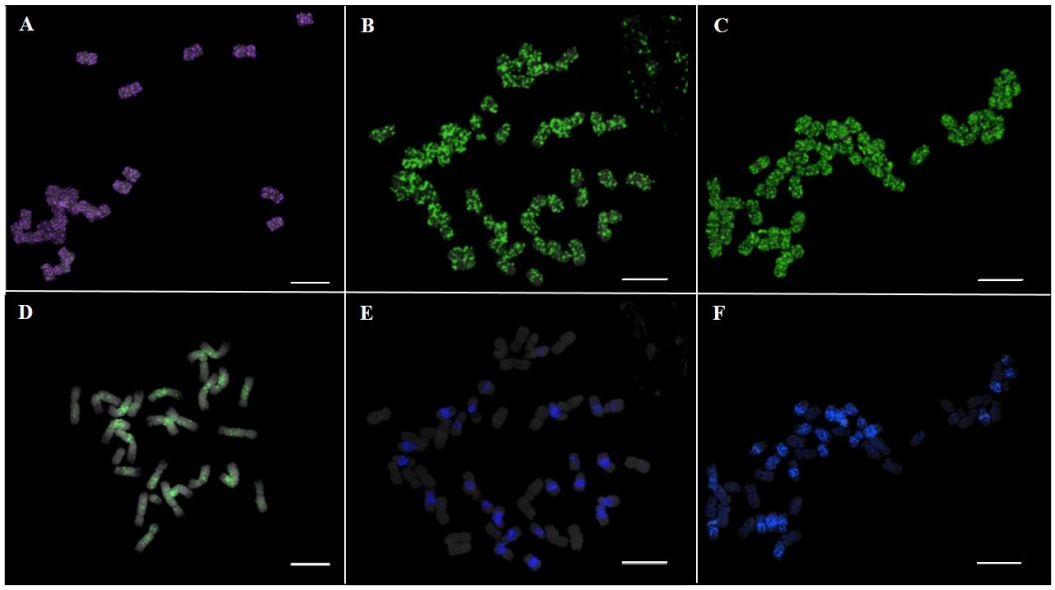

Figure 3-5. Fluorescence *in situ* hybridization (FISH) analysis of *PgDel1* and *PgDel2* distribution in *P. ginseng*, *P. quinquefolius*, and *P. notoginseng* chromosomes. The *PgDel1* FISH signals in somatic metaphase chromosomes of (A) *P. notoginseng* (purple), (B) *P. ginseng*, and (C) *P. quinquefolius*. The *PgDel2* FISH signals in somatic metaphase chromosomes of (D) *P. notoginseng*, (E) *P. ginseng* (blue), and (F) *P. quinquefolius* (blue). Bar = 10 μm.

Contribution of major repeats to genome size variation

The contribution of the four most abundant LTR-RT families, *PgDel*, *PgAthila*, *PgTat*, and *PgTork*, were investigated to the overall genome contents. Each family was present in varied proportions in the six analyzed species (Fig. 3-3 and 3-6). Combined, the four LTR-RTs had a 39–52% GP in each of five species, corresponding to 0.9–2.6 Gb. Of these repeats, *PgDel* occupied 30–41% of R-GP, accounting for 0.7–2.0 Gb. *PgTat* had a 5-7% R-GP, corresponding to 97–285 Mb. The estimated quantity of *PgTork* was 241 Mb in *P. quinquefolius*, whereas it was 24–57 Mb in the other four *Panax* species. Interestingly, *PgTork* was the most abundant LTR-RT in the *A. elata* genome, occupying 7% R-GP (56 Mb) (Fig. 3-6).

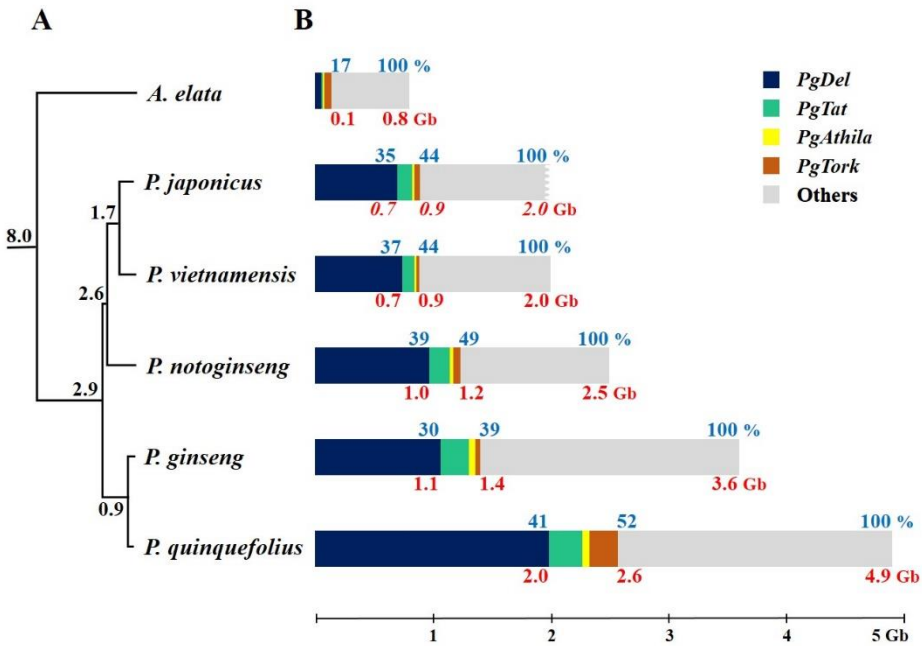

Figure 3-6. Comparison between proportions of four major repeats in five *Panax* species and *A. elata*. **(A)** Phylogenetic relationships based on chloroplast sequences modified from *Kim* et al. (Kim*, et al.* 2017). Estimated time since divergence (MYA) is indicated at the root of branch divisions. **(B)** The predicted genome size of *Panax* species and *A. elata* are depicted as bar charts with the estimated amounts of *PgDel*, *PgTat*, *PgAthila*, and *PgTork* families contained in each genome represented by blue, green, yellow, and brown regions, respectively. Genome contents not containing these repeats are represented by the grey region. Blue letters above bars indicate GP of *PgDel* alone (left), GP of four LTR-RT families (middle), and total GP of the genome (right). Red letters below bars indicate estimated amount in Gb of *PgDel* contents alone (left), contents of four LTR-RT families (middle), and total genome size (right). For *A. elata*, blue letters above bars indicate GP of four LTR-RT families (left) and total GP of the genome (right), whereas red letters below bars indicate estimated amount in Gb of four LTR-RT families (left) and total genome size (right). Total genome size of *P. japonicus* was estimated in the present study.

DISCUSSION

In this work, low-coverage WGS sequences were used to calculate the GP of major repeats. The prevalence of each repeat were estimated by determining GP using various WGS data sets, based on the calculation of masked homologous sequence in raw WGS reads by RepeatMasker (Smit*, et al.* 2013-2015). GP can also be calculated using clustered WGS reads or mapped WGS reads (Macas*, et al.* 2007, Perumal*, et al.* 2017). Mapping-based GP (M-GP) and clustering-based GP (C-GP) calculations are based on numbers of homologous WGS reads, whereas R-GP calculation is based on real amounts of homologous sequences in WGS reads. The ability of R-GP and M-GP methods were compared to estimate *PgDel1* GP using different WGS sets, which resulted in a consistent pattern whereby R-GP calculations estimated 3–4% more GP than M-GP calculations (Fig. 3-7 and 3-8). This variation may be attributed to the difference in how homologous sequences are counted in both methods, namely the number of homologous reads and the number of homologous nucleotides for M-GP and R-GP, respectively.

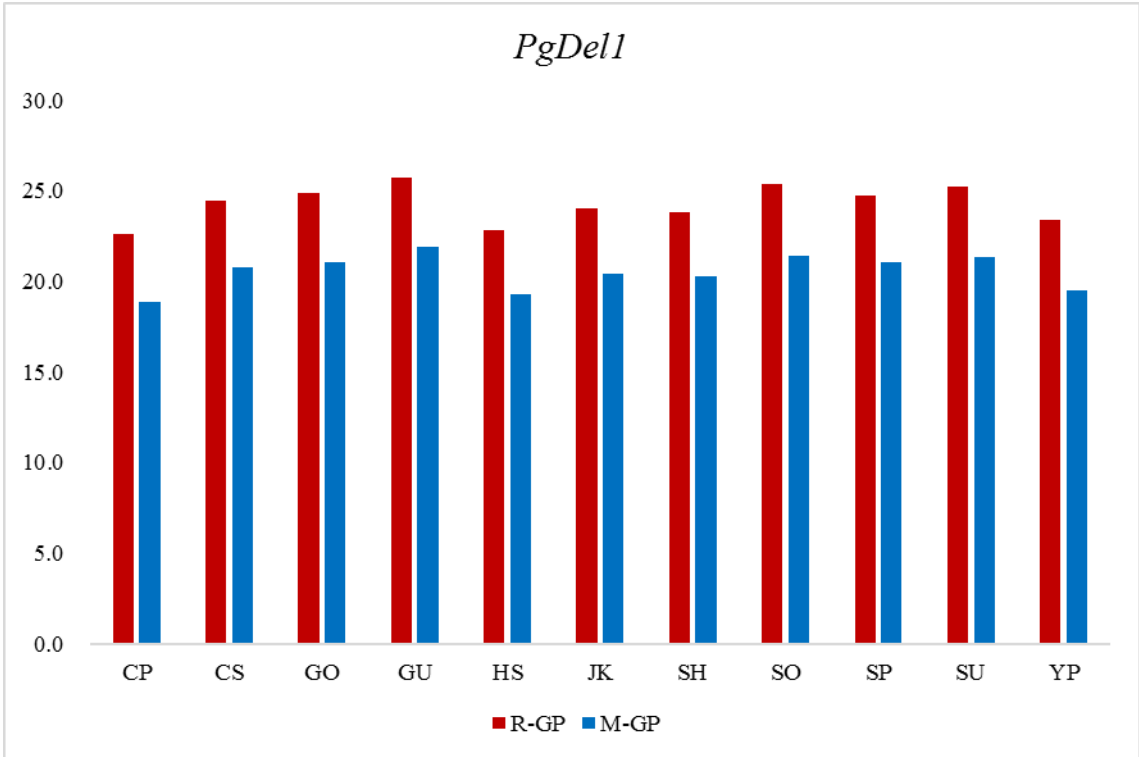

Figure 3-7. Comparison of R-GP and M-GP for *PgDel1* GP estimation using WGS data of 11 ginseng cultivars. The GPs were calculated based on RepeatMasker (R-GP) and CLC Mapper (M-GP) and are indicated with red and blue bars, respectively. The 11 cultivars are listed in Table 1-4.

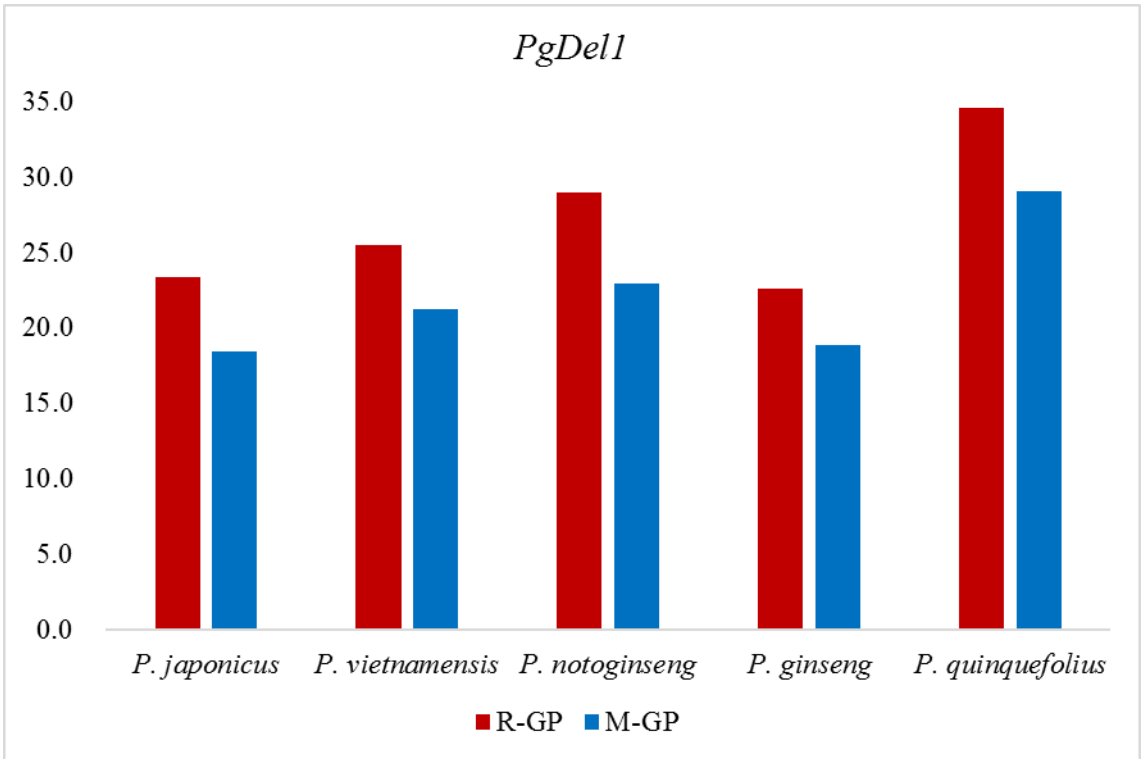

Figure 3-8. Comparison of R-GP and M-GP for *PgDel1* GP estimation using WGS data of five *Panax* species. The GPs were calculated based on RepeatMasker (R-GP) and CLC Mapper (M-GP) and are indicated with red and blue bars, respectively.

The estimated GPs were highly reproducible for the major high-copy LTR-RTs, although relatively high CV values (25-65%) were observed for GP estimation of tandem repeats using different genome coverage and different WGS libraries (Supplementary Table 3-2 and 3-3). The number of tandem repeat reads might be uneven because of biased fragmentation during WGS library construction (Supplementary Table 3-2 and 3-3). Overall, though, the coefficients of variation (CVs) of the high copy *PgDel1* and the low copy *PgOryco* were 3.23% and 13.53%, respectively, when GP using various levels of genome coverage was estimated in the data sets, i.e., 0.0001−10x genome coverage for WGS reads of *P. ginseng.* Slightly increased variation were observed when the genome coverage below 0.00001x was reduced, but all the data showed similarly low CVs when over 0.0001x coverage WGS reads was utilized. As WGS data can be produced at low cost by high-throughput NGS processes and over 1 Mbp of WGS reads produced reproducible GP estimation, genome coverage in WGS data may be not a critical constraint limiting the application of this approach for analysis.

Although there is some variation, the GPs calculated here for the major repeats were reproducible and are thus representative of the abundance of each repeat in the genomes of the different species. However, it is possible that the true GP for each major repeat is higher than the GP estimate presented here because, in our analysis, only a single representative structure was used for each repeat and other structural variations were not considered (Devos 2002). For example, five *PgDel1* elements displayed large structural variation in the LTR region and bias in WGS read mapping (111-157,407 copies) for the

representative *PgDel1* family member in the *P. ginseng* genome (Fig. 3-4).

Our results point to tetraploidization and four LTR-RTs as the primary reasons for genome size variation in the genus *Panax*. Divergence of a common ancestor into the genera *Panax* and *Aralia* is predicted to have occurred approximately eight MYA (Kim *et al.* 2017). *A. elata* was estimated approximate haploid genome equivalent size of 0.8 Gb on 12 chromosome pairs (Supplementary Fig. 3-2). However, the genome sizes of *Panax* species (2.0–4.9 Gb) are much larger than that of *A. elata*. The multiplication of some major repeats may influenced the genome size in the *Panax* lineage. In particular, a large proportion of the increased genome size can be explained by multiplication of the four LTR-RTs investigated here, which occupied 0.9–2.6 Gb in *Panax* lineage (Fig. 3-6). The GP of the four LTR-RTs was 39% (1.4 Gb) and 52% (2.6 Gb) in two tetraploids, *P. ginseng* and *P. quinquefolius*, respectively, and 44–49 % (0.9–1.2 Gb) in the three diploid *Panax* species. Among them, *PgDel* was the predominant repeat with a GP of 30–41%, which corresponds to 0.7–2.0 Gb in the five *Panax* species.

LTR-RTs make up a large proportion of the genomes of many higher plants (Gill *et al.* 2010, Hawkins 2009, Neumann 2006). The repeats can play an important role as promoters of genomic diversification and speciation (Levy 2013). It is possible that, even in the same genus, a rapid burst of retrotransposition can induce genome size variance with different evolutionary effects, as observed for *Oryza*, *Nicotiana*, and *Genlisea* (Piegu *et al.* 2006, Renny-Byfield *et al.* 2013, Vu *et al.* 2015). Here abundant, high-copy LTR-RTs were investigated and a comparative analysis of these repeats in *Panax* species was performed

to understand their influence on genome evolution. The presence of these repeats in five *Panax* species and a further related species suggests that they likely existed in the genome of a common ancestor (Fry *et al.* 1977). However, extensive multiplication of LTR-RTs occurred only in the *Panax* genus and appears to have a decisive effect on the expansion of the genome sizes in *Panax* species (Fig. 3-3 and 3-6). This finding suggests that the repeat amplification occurred concomitantly with or following divergence in the five *Panax* species during the last eight million years.

Six *PgDel* subfamilies were identified based on LTR sequences from *P. ginseng* BAC clone sequences (Choi *et al.* 2014, Jang *et al.* 2017). Among them, *PgDel1* was highly abundant in each *Panax* species. The abundance, sequence diversity, and cytogenetic distribution of *PgDel1* LTR-RTs indicated that considerable multiplication and transposition may have occurred across the five *Panax* species genomes (Fig. 3-3, 3-4, 3-5, and 3-6). A positive correlation (coefficient of 0.6 with p-value of 0.40) was found between the R-GP values for *PgDel1* and the genome size of each *Panax* species (Table 3-1 and Fig. 3-6). This correlation indicates that the accumulation of *PgDel1* elements has greatly contributed to the increased genome sizes in the genus *Panax*. In this regard, the genome size of *P. japonicus* might be below 2.0 Gb, based on the relatively small *PgDel1* GP in the diploid *Panax* species (Fig. 3-3 and 3-6).

Correlation between *PgDel1* abundance and genome size in *Panax* species could explain the expansive genome of *P. quinquefolius*, which is the largest within the *Panax* genus. The two tetraploids *P. ginseng* and *P. quinquefolius* were reported to exhibit a

difference of 1.3 Gb. Based on divergence of orthologous gene pairs, these species might be diverged approximately one MYA, following the recent allotetraploidization two MYA (Choi *et al.* 2013). The considerable disparity in genome size that has evolved between *P. ginseng* and *P. quinquefolius* is largely explained by the different amount of *PgDel1* in each genome, which is 0.8 and 1.7 Gb, respectively, indicating that *PgDel1* was exclusively amplified in *P. quinquefolius* during last one MY.

The difference between *PgDel1* GP in *P. ginseng* and *P. quinquefolius* can be explained by two hypotheses concerning TE dynamics. The first hypothesis is that there was a considerable loss of *PgDel1* GP in *P. ginseng* after speciation. Polyploidization often results in genome downsizing via expulsion of genomic DNA, mostly repetitive DNA sequence, for stable meiotic rebuilding in nascent polyploids (Fedoroff 2012, Renny-Byfield *et al.* 2011, Renny-Byfield *et al.* 2013). The second hypothesis is that there was a sizeable expansion of *PgDel1* GP in *P. quinquefolius* after speciation. I believe that the second hypothesis holds more merit than the first. Drastic environmental change could have triggered epigenetic restructuring (Fedoroff 2012, Kalendar *et al.* 2000, Tank *et al.* 2015), resulting in the unusual accumulation of LTR-RTs in *P. quinquefolius* (Alzohairy *et al.* 2014). *P. quinquefolius* is said to have migrated from Asia to America through the Bering land bridge during glacial and interglacial cycles one MYA (Choi *et al.* 2013). Consequently, *P. quinquefolius* would have been exposed to extreme abiotic stress during the process of migration and adaptation to new habitats. The influence of *PgDel1* amplification in genome organization and gene function accordingly might play an

important role in the interspecific genomic barriers between species.

PgDel1 made up a large proportion of the genome in all five *Panax* species analyzed in this study. In addition, other *PgDel* subfamily members also had notable genome distributions in the *Panax* species. *PgDel2* occupied approximately 1.4% GP in the two tetraploids and 2.8% GP in the three diploid *Panax* species (Fig. 3-3). This variation in *PgDel2* GP between diploids and tetraploids was confirmed by cytogenetic analysis using FISH. In the tetraploids *P. ginseng* and *P. quinquefolius*, *PgDel2* signals were observed in half of the chromosomes whereas all chromosomes of the diploid *P. notoginseng* displayed *PgDel2* signals (Fig. 3-5 D, E, and F). *PgDel5* and *PgDel6* showed big difference among three *Panax* species. *P. notoginseng* had approximately twice the amount of *PgDel5* than the other *Panax* species and *P. japonicus* and *P. vietnamensis* had more abundant *PgDel6* than other species. These findings highlight the likely importance of *PgDel* subfamilies contribute in diversification of *Panax* species.

REFERENCES

Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF. 2006. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat. Protoc. 1: 2320-2325.

Alzohairy AM, Jamal SM, Sabir B, Gábor Gyulai C, Rania AA, Younis D, Robert K, Jansen BE, Ahmed B. 2014. Environmental stress activation of plant long-terminal repeat retrotransposons. Funct. Plant Biol. 41: 557-567.

Bai C, Alverson WS, Follansbee A, Waller DM. 2012. New reports of nuclear DNA content for 407 vascular plant taxa from the United States. Ann. Bot. 110: 1623-1629.

Bai D, Brandle J, Reeleder R. 1997. Genetic diversity in North American ginseng (*Panax quinquefolius* L.) grown in Ontario detected by RAPD analysis. Genome 40: 111-115.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Choi HI, Kim NH, Lee J, Choi BS, Kim KD, Park JY, Lee SC, Yang TJ. 2013. Evolutionary relationship of *Panax ginseng* and *P. quinquefolius* inferred from sequencing and comparative analysis of expressed sequence tags. Genet. Resour. Crop Evol. 60: 1377-1387.

Choi HI, Waminal NE, Park HM, Kim NH, Choi BS, Park M, Choi D, Lim YP, Kwon SJ, Park BS, Kim HH, Yang TJ. 2014. Major repeat components covering one-third of

the ginseng (*Panax ginseng* C.A. Meyer) genome and evidence for allotetraploidy. Plant J. 77: 906-916.

Csink AK, Henikoff S. 1998. Something from nothing: the evolution and utility of satellite repeats. Trends in Genetics 14: 200-204.

Devos KM, Brown, JK, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. 12: 1075-1079.

Fedoroff NV. 2012. Transposable elements, epigenetics, and genome evolution. Science 338: 758-767.

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41: 331-368.

Fry K, Salser W. 1977. Nucleotide sequences of HS-alpha satellite DNA from kangaroo rat Dipodomys ordii and characterization of similar sequences in other rodents. Cell 12: 1069-1084.

Gill N, SanMiguel P, Dhillon BDS, Abernathy B, Kim HR, Stein L, Ware D, Wing R, Jackson SA. 2010. Dynamic *Oryza* genomes: repetitive DNA sequences as genome modeling agents. Rice 3: 251-269.

Hawkins JS, Proulx SR, Rapp RA, Wendel JF. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc. Natl. Acad. Sci. USA 106: 17811-17816.

Ho IS, Leung FC. 2002. Isolation and characterization of repetitive DNA sequences from *Panax ginseng*. Mol. Genet. Genomics 266: 951-961.

Hong CP, Lee SJ, Park JY, Plaha P, Park YS, Lee YK, Choi JE, Kim KY, Lee JH, Lee J, Jin H, Choi SR, Lim YP. 2004. Construction of a BAC library of Korean ginseng and initial analysis of BAC-end sequences. Mol. Genet. Genomics 271: 709-716.

Jang W, Kim NH, Lee J, Waminal NE, Lee SC, Jayakodi M, Park JY, Choi HI, Yang TJ. 2017. Complex genome structure of *Panax ginseng* revealed by ten BAC clone sequences obtained by 3rd generation SMRT sequencing platform using pooled DNA. Plant Breed. Biotech. 5: 25-35.

Jayakodi M, Lee SC, Park HS, Jang W, Lee YS, Choi BS, Nah GJ, Kim DS, Natesan S, Sun C, Yang TJ. 2014. Transcriptome profiling and comparative analysis of *Panax ginseng* adventitious roots. J. Ginseng Res. 38: 278-288.

Jayakodi M, Lee SC, Lee YS, Park HS, Kim NH, Jang W, Lee HO, Joh HJ, Yang TJ. 2015. Comprehensive analysis of *Panax ginseng* root transcriptomes. BMC Plant Biol. 15: 138.

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE-1* retrotransposon dynamics in response to sharp microclimatic divergence. Proc. Natl. Acad. Sci. USA 97: 6603-6007.

Kim K, Lee SC, Lee J, Lee HO, Joh HJ, Kim NH, Park HS, Yang TJ. 2015. Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. PloS ONE 10: e0117159.

Kim K, Nguyen VB, Dong J, Wang Y, Park JY, Lee SC, Yang TJ. 2017. Evolution of the Araliaceae family inferred from complete chloroplast genomes and 45S nrDNAs of

10 *Panax*-related species. Sci. Rep. In press.

Kim NH, Choi HI, Ahn IO, Yang TJ. 2012. EST-SSR marker sets for practical authentication of all nine registered ginseng cultivars in Korea. J. Ginseng Res. 36: 298-307.

Lee C, Wen J. 2004. Phylogeny of *Panax using* chloroplast trnC–trnD intergenic region and the utility of trnC–trnD in interspecific studies of plants. Mol. Phylogenet. Evol. 31: 894-903.

Leitch IJ, Bennett MD. 2004. Genome downsizing in polyploid plants. Biol. J. Linnean Soc. 82: 651-663.

Levy AA. 2013. Transposons in Plant Speciation (ed. Fedoroff, N. V.) ch9 (John Wiley & Sons, Inc.).

Lim KB, Yang TJ, Hwang YJ, Kim JS, Park JY, Kwon SJ, Kim J, Choi BS, Lim MH, Jin M. 2007. Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related *Brassica* species. Plant J. 49: 173-183.

Macas J, Neumann P, Navrátilová A. 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. BMC Genomics 8: 427.

Neumann P, Koblizkova A, Navratilova A, Macas J. 2006. Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics 173: 1047-1056.

Obae SG, West TP. 2012. Nuclear DNA content and genome size of American ginseng. J.

Med. Plants Res. 6: 4719-4723.

Oliver KR, McComb JA, Greene WK. 2013. Transposable elements: powerful contributors to angiosperm evolution and diversity. Genome Biol. Evol. 5: 1886-1901.

Pan YZ, Zhang YC, Gong X, Li FS. 2014. Estimation of genome size of four *Panax* species by flow cytometry. Plant Diversity and Resour. 233-236.

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? Bot. J. Linean Soc. 164: 10-15.

Perumal S, Waminal N, Lee J, Lee J, Choi BS, Kim HH, Grandbastien MA, Yang TJ. 2017. Elucidating the major hidden genomic components of the A, C, and AC genomes and their influence on Brassica evolution. BMC Biol. Submitted.

Piégu B, Bire S, Arensburger P, Bigot Y. 2015. A survey of transposable element classification systems–a call for a fundamental update to meet the challenge of their diversity and complexity. Mol. Phylogenet. Evol. 86: 90-109.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O. 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. 16: 1262-1269.

Renny-Byfield S, Chester M, Kovařík A, Le Comber SC, Grandbastien MA, Deloger M, Nichols RA, Macas J, Novák P, Chase MW, Leitch AR. 2011. Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. Mol. Biol. Evol. 28: 2843-2854.

Renny-Byfield S, Kovarik A, Kelly LJ, Macas J, Novak P, Chase MW, Nichols RA, Pancholi MR, Grandbastien MA, Leitch AR. 2013. Diploidization and genome size change in allopolyploids is associated with differential dynamics of low-and high-copy sequences. Plant J. 74: 829-839.

Richard GF, Kerrest A, Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol. Biol. Rev. 72: 686-727.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. Nat. Genet. 20: 43-45.

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.

Seol YJ, Lee TH, Park DS, Kim CK. 2016. NABIC: A New Access Portal to Search, Visualize, and Share Agricultural Genomics Data. Evol. Bioinform. Online 12: 51-58.

Smit A, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Tank DC, Eastman JM. Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, Brown JW, Sessa EB, Harmon LJ. 2015. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. New Phytol. 207: 454-467.

Volff J. N. 2006. Turning junk into gold: domestication of transposable elements and the

creation of new genes in eukaryotes. Bioessays 28: 913-922.

Vrana J, Simkova H, Kubalakova M, Cihalikova J, Dolezel J. 2012. Flow cytometric chromosome sorting in plants: the next generation. Methods 57: 331-337.

Vu GTH, Schmutzera T, Bulla F, Caoa HX, Fuchsa J, Trana TD, Jovtchevag G, Pistricka K, Steina N, Pecinkab A, Neumannc P, Novakc P, Macasc J, Deard PH, Blattnerae FR, Scholza U, Schubert I. 2015. Comparative genome analysis reveals divergent genome size evolution in a carnivorous plant genus. The Plant Genome 8: 3.

Waminal NE, Kim HH. 2012. Dual-color FISH karyotype and rDNA distribution analyses on four Cucurbitaceae species. Hort. Environ. Biotechnol. 53: 49-56.

Wen J, Plunkett GM, Mitchell AD, Wagstaff SJ. 2001. The evolution of Araliaceae: a phylogenetic analysis based on ITS sequences of nuclear ribosomal DNA. Syst. Botany. 26: 144-167.

Wendel JF. 2000. Genome evolution in polyploids. Plant Mol. Biol. 42: 225-249.

Yang TJ, Kim JS, Kwon SJ, Lim KB, Choi BS, Kim JA, Jin M, Park JY, Lim MH, Kim H I, Lim YP, Kang JJ, Hong JH, Kim CB, Bhak J, Bancroft I, Park BS. 2006. Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of *Brassica rapa*. Plant Cell 18: 1339-1347.

Yun TK. 2001. Brief introduction of *Panax ginsen g* CA Meyer. J. Korean Med. Sci. 16: S3-5.

GENERAL DISCUSSION

In the last few decades, whole genome sequencing (WGS) technology has provided a rapid advance in the plant genomic era (Metzker 2010, Michael *et al.* 2015, Shendure *et al.* 2008). However, genomic analysis for understanding many non-model plants or resource plants is still lacking due to technical and financial challenges. In this current study, it is shown that fundamental genomic study can be done with only a small amount of WGS data. This is a more advanced study about the application of small WGS data, which was previously done by the *de novo* assembly using low coverage WGS (dnaLCW), providing successful analysis of the chloroplast (cp) genome and nuclear ribosomal DNA (nrDNA). Using this, complete cp genome and 45S nrDNA sequences of *E. hamilotinanus* were assembled and phylogenetically analyzed with 11 related species (chapter I).

The key concept of this study is that repetitive sequences in plant genomes are sufficient enough for small scale WGS data due to their abundance, and this has been experimentally verified (Table 3-2 in chapter III). Based on this point, comparative analysis of polymorphic simple sequence repeats (pSSR) by detection of polymorphic SSR using low coverage WGS (dpsLCW) and major repeat analysis by RepeatMasker-based genomic proportion (R-GP) were conducted (chapter I, II, and III).

Simple sequence repeats (SSR) are one of the most useful elements in molecular marker systems for the genetic study due to their advantages: 1) high-reproducibility, 2) hyper-

variable nature, and 3) co-dominance nature (Park *et al.* 2009). The devised dpsLCW could easily distinguish pSSRs between two WGS data sets based on *in silico* prediction (chapter I). The dpsLCW was an economical, reliable, and scalable protocol for investigating pSSRs. The dpsLCW protocol was simulated and evaluated using *E. hamiltonianus* and *P. japonicum* (chapter I and II).

Efficient and reliable conditions for repeat quantification using RepeatMasker were described in chapter III (Smit *et al.* 2013-2015). The estimated genomic proportions (GPs) of major repeats indicates that four long terminal repeats retrotransposons (LTR-RTs), *PgDel, PgTat, PgAthila,* and *PgTork* makes up 39-52% of the five *Panax* species genomes. The observation of only four LTR-RTs could explain for the genome size variation due to the expansion of one *PgDel* elements in the *Panax* genus and phylogenetic relationships between diploids and tetraploids (chapter III).

These chapters showed new stages for plant genomics using low coverage of WGS data. Furthermore, *de novo* and basic genomic study in non-model plants or minor crops could be conducted using low coverage WGS.

REFFERENCES

Metzker ML. 2010. Sequencing technologies the next generation. Nat. Rev. Genet. 11: 31-46.

Michael TP, VanBuren R. 2015. Progress, challenges and the future of crop genomes. Curr. Opin. Plant Biol. 24: 71-81.

Park YJ, Lee JK, Kim NS. 2009. Simple sequence repeat polymorphisms (SSRPs) for evaluation of molecular diversity and germplasm classification of minor crops. Molecules 14: 4546-4569.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26: 1135-1145.

Smit A, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

국문 초록

전장 유전체 연구는 차세대 시퀀싱(NGS: next generation sequencing)기술이 개발된 이래로 급속도로 발전하고 있다. NGS로 인해 많은 양의 데이터를 생성할 수 있게 되었지만 완전한 전장 유전체 서열을 얻는 것은 여전히 지난하며, 이러한 연구는 모델 식물이나 주요 작물에 편중 되어 진행 되는 경우가 많았다. 반면에 자원 식물과 같은 비(非)모델 식물의 경우 연구가 미비한 실정이기 때문에, 본 연구에서는 대규모 데이터가 아닌 소규모의 유전체 데이터를 이용하여 수행할 수 있는 식물 유전체 연구에 대해 초점을 맞추었으며, 3종의 자원 식물을 대상으로 다양한 연구를 진행하였다. 이에 본 연구에서는 유전체 연구에 있어 유용하게 사용되는 바코딩 정보를 제공하는 엽록체와 핵내 리보솜 DNA서열을 완성하였고, 개체간 차이를 보기 위해 단순 서열 반복의 다형성(pSSR: polymorphic simple sequence repeat)에 대한 연구를 수행하였으며, 식물에서의 주요 반복 서열을 분석하는 연구를 진행하였다.

첫째, 본 연구에서는 약용 및 정원수로 사용되는 참빗살나무(*Euonymus hamiltonianus*)의 전장 유전체 서열을 사용하여 분자 마커(marker)의 재료로 널리 사용되는 엽록체(chloroplast)의 유전체 서열과 핵 리보솜 DNA(nuclear ribosomal DNA) 서열을 완성하였으며, 더 나아가 pSSR을 탐

지할 수 있는 파이프라인을 개발하였다. 본 챕터에서는 잎에서의 표현형이 다른 두 참빛살나무를 이용하여 157,360 bp의 엽록체서열과 5,824 bp의 45S 핵 리보솜 DNA 서열을 완성하였다. 161개 pSSR 후보 서열들을 발굴하였으며 이를 이용하여 20개의 프라이머를 디자인하고 그 중 7개의 마커에서 다형성을 확인하였다. 두번째 챕터에서는 약용 및 식용 작물로 널리 재배되고 있지만 유전학적 연구가 미비한 식방풍(*Peucedanum japonicum*)을 대상으로 이전 연구에서 개발한 pSSR 발굴 파이프라인을 이용하여 우리나라에 자생하고 있는 7개의 식방풍 개체들을 구분할 수 있는 pSSR마커를 개발하고자 하였고, 이를 통해 452개의 후보 서열들을 탐색하였으며, 제작한 10개의 프라이머 중 9개에서 다형성을 확인하였다. 세번째 챕터에서는 유전체 내의 반복서열을 정량할 수 있는 효율적인 조건 및 방법을 확립하였으며 이를 통해 수천년전부터 약용식물로 이용된 인삼 속(*Panax* genus)내의 식물들의 소규모 전장 유전체 데이터를 이용하여 주요 반복 서열에 대한 연구를 진행하였다. 인삼 속내 이배체(diploid)인 죽절삼(*P. japonicus*), 베트남삼(*P. vietnamensis*), 전칠삼(*P. notoginseng*)과 사배체(tetraploid)인 인삼(*P. ginseng*), 미국삼(*P. quinquefolius*)과 비교 연구를 위해 인삼속과 가장 가까운 근연종인 두릅(*Aralia elata*)을 이용하였다. 이들은 총 0.8-4.9 Gb로 다양한 유전체 크기를 가지고있다. 본 연구에서는 인삼속내 식물의 유전체 중 39-52%가 오직 4개의 long terminal repeat retrotransposons(LTR-RTs)

인 *PgDel, PgTat, PgAthila, PgTork*로 이루어져 있음을 발견하였다. 그 중에서도 *PgDel* LTR-RT superfamily는 인삼 속 유전체에서 23-35%를 차지하고 있을 정도로 5종의 식물의 유전체 크기가 증가함에 있어 많은 기여를 하고 있는 것으로 조사되었다. 특히 같은 사배체 식물임에도 인삼(3.6 Gb)과 미국삼(4.9 Gb)의 유전체 크기 차이가 매우 큰 것을 알 수 있었는데, 이러한 차이는 대부분은 0.9 Gb의 *PgDel*의 확장으로 인한 것이었으며 이는 약 백만년 전에 아시아 대륙에서 북미대륙으로 미국삼이 이주하면서 겪었을 환경적 변화에 적응하는 과정에서의 일어난 유전체내의 변화 때문일 것으로 추측된다. 또한 사배체에서와 이배체에서 *PgDel2* LTR-RT의 관찰을 통해 사배체 인삼속 식물의 한쪽 조상과 이배체 인삼속 식물의 공동조상이 매우 밀접한 관련이 있었을 것으로 추정할 수 있었다. 본 연구에서는 소규모의 데이터로도 기초적이고 또 기본적인 유전체연구가 가능하다는 것을 보여주었으며, 이는 추후 연구가 미비한 식물에서의 유전체 연구에도 도움을 줄 수 있을 것이다.

학번 : 2012-30301