A DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Genome structure and evolution of *Panax ginseng* C. A. Meyer

BY

HONG-IL CHOI

FEBRUARY, 2013

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Genome structure and evolution of

Panax ginseng C. A. Meyer

UNDER THE DIRECTION OF DR. TAE-JIN YANG

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

BY

HONG-IL CHOI

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

APPROVED AS A QUALIFIED DISSERTATION OF HONG-IL CHOI FOR THE DEGREE OF DOCTOR OF PHILOSOPHY BY THE COMMITTEE MEMBERS

FEBURUARY, 2013

CHAIRMAN _____

Suk-Ha Lee, Ph.D.

VICE CHAIRMAN _____

Tae-Jin Yang, Ph.D.

MEMBER _____

Hee-Jong Koh, Ph.D.

MEMBER _____

Doil Choi, Ph.D.

MEMBER _____

Hyun Hee Kim, Ph.D.

Genome structure and evolution of *Panax ginseng* C. A. Meyer

HONG-IL CHOI

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

GENERAL ABSTRACT

Ginseng (*Panax ginseng* C. A. Meyer) has been known as a valuable medicinal herb for thousands of years in East Asia. Although medicinal components and their functions of ginseng have been widely investigated, it could be regarded as an underdeveloped crop in genetics and genomics research areas. This study was conducted to elucidate genome structure and evolution in ginseng by analyses of expressed sequence tags (ESTs) and bacterial artificial chromosome (BAC) sequences. The EST analysis based on

the calculation of synonymous substitutions per synonymous site (Ks) in paralog and ortholog pairs revealed that two rounds of polyploidy events occurred in the common ancestor of ginseng and American ginseng (*P. quinquefolius* L.) and subsequent divergence of the two species. The sequence analysis of repeat-rich BAC clones characterized the major component of the ginseng genome, long terminal repeat retrotransposons (LTR-RTs). *Ty3/Gypsy*-like elements were more predominant than *Ty1/Copia*. High abundance of the LTR-RTs were revealed by whole genome shotgun (WGS) read mapping and fluorescence *in situ* hybridization (FISH) analysis. Particularly, the *PgDel1* elements played major roles in expanding heterochromatic regions as well as remodeling euchromatic regions. The *PgDel2* elements showed biased intensity of FISH signals on half the total chromosomes, which demonstrate the alloploidy-like nature of ginseng. Insertion time of the LTR-RTs indicated that LTR-RTs may proliferate after the recent polyploidy event in the ginseng genome. These results suggest that the ginseng genome of the present day has been expanded and evolved by two rounds of polyploidy events and accumulation of LTR-RTs.

Keywords: Bacterial artificial chromosome (BAC), ginseng (*Panax ginseng* C. A. Meyer) genome, expressed sequence tag (EST), fluorescence *in situ* hybridization (FISH), long terminal repeat retrotransposon (LTR-RT), polyploidy

Student Number: 2007-21304

CONTENTS

CHAPTER II. Repeat-rich bacterial artificial chromosome (BAC) sequences and fluorescence *in situ* hybridization (FISH) unveil major heterochromatic components and alloploid-like genome structure in *Panax ginseng*

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS USED

BAC	Bacterial artificial chromosome
BP	Biological process
CC	Cellular component
CDD	Conserved Domain Database
CD-Search	Conserved Domain Search
EST	Expressed sequence tag
FISH	Fluorescence *in situ* hybridization
GISH	Genomic in situ hybridization
GO	Gene ontology
IGS	Intergenic spacer
ITS	Internal transcribed spacer
Ks	Synonymous substitution rate
LARD	Large retrotransposon derivative
LTR-RT	Long terminal repeat retrotransposon
MF	Molecular function
MITE	Miniature inverted-repeat transposable elements
MYA	Million years ago
NR database	Non-redundant protein sequence database
ORF	Open reading frame
SRA	Sequence Read Archive
TRIM	Terminal-repeat retrotransposon in miniature
Ts	Transition
TSD	Target site duplications
Tv	Transversion
WGS	Whole genome shotgun

GENERAL INTRODUCTION

Ginseng (*Panax ginseng* C. A. Meyer) is a well-known medicinal herb belonging to the family Araliaceae and has been used as oriental medicine for thousands of years (Yun 2001). Cultivation of ginseng has been started since 15th century, before then, it had only gathered from mountainous areas (Park et al. 2012). The major components showing pharmacological effects are ginsenosides, which are known for their beneficial properties to the central nervous system, cardiovascular, endocrine and immune systems (Attele et al. 1999).

In ginseng research, medicinal components and their functions have been widely investigated. However, breeding, genetic and genomic studies had been rarely performed because of difficulty in maintaining plants and reproducing progenies. Minimum growth time of three to four years is necessary to produce a small number of seeds, approximately 40 seeds per plant (Choi et al. 1992), thus hindering systematic management of genetic materials. Despite study difficulties, eight elite cultivars, Chunpoong, Yunpoong, Gumpoong, Gopoong, Sunpoong, Sunun, Sunwon and Chungsun, have been bred by pure line selection and have been registered as commercial varieties since 1997 in South Korea (Lee et al. 2008). Not only the breeding field, genetic and genomic studies have been successively reported in recent ten years such as functional analyses of genes (Kim et al. 2008; Kim et al. 2011; Tansakul et al. 2006), marker development (Choi et al. 2011; Kim et al. 2012; Sun et al. 2011), construction of DNA library (Bang et al. 2010; Hong et al. 2004), and transcriptome sequencing (Chen et al. 2011; Wu et al. 2012).

The haploid (1C) genome size of the ginseng was estimated at 3.12 Gb (Hong et al. 2004). Enormous genome size diversity in the angiosperms reported with an extraordinary range of more than 1200-fold (Zonneveld et al.

2005). The major two factors affecting genome obesity have been reported as polyploidy and repetitive DNAs, which are driving forces of genome variation in size and evolution (Soltis et al. 2003; Lysak et al. 2009; Bennetzen 2005; Gaut et al. 2000).

Ginseng of which the chromosome number is $2n = 48$ has been regarded as a tetraploidy species because the basic chromosome number of the genus *Panax* is $x = 12$ (Wen and Zimmer 1996; Yi et al. 2004). The basic number of $x = 12$ was also reported in the study about the family Araliaceae, while the hypothesis of $x = 6$ is plausible based on the existence of taxa with $2n = 18$, 36, and 60 (Plunkett et al. 2004; Yi et al. 2004). Although the previous studies suggested the polyploidy level of ginseng simply based on the chromosomal data, constant multi-band patterns observed in the process of expressed sequence tag (EST) derived marker development could make a well-founded conjecture that the genomes of *Panax* species are highly duplicated (Choi et al. 2011; Kim et al. 2012). Lee and Wen (2004) suggested the likelihood of allopolyploidy in tetraploid *Panax* species because of their incongruent phylogenies between chloroplast intergenic region and ITS datasets. However, there has been no report with considerably large-scale molecular data to explain the evolution of ginseng and related species.

Transposable elements are the major components of repetitive DNAs, which are omnipresent in eukaryotic genomes (Kidwell and Lisch 1997). Transposable elements can be classified into two classes by their transposition mechanisms, which are class I retrotransposons with a copy-and-paste mode and class II DNA transposons with a cut-and-paste mode (Finnegan 1989; Wicker et al. 2007). Of them, long terminal repeat retrotransposons (LTR-RTs) belonging to class I have been known as the main factor in size variation of plant genomes (Bennetzen et al. 2005; Kumar and Bennetzen 1999; Vitte and Panaud 2005). A few studies reported that ginseng is likely to have a large proportion of LTR-RTs in its genome based on the analysis of repetitive DNA

sequences (Ho and Leung 2002) and bacterial artificial chromosome (BAC) -end sequences (Hong et al. 2004). However, no large-scale sequence analysis has been conducted for characterization of main repeat components in the ginseng genome.

The objective of this study is to elucidate genome structure and evolutionary process in ginseng. To uncover genome-level duplication events and speciation, EST data were generated from a cDNA library and comparative analysis was conducted between ginseng and American ginseng. To identify major repeat components in the ginseng genome three repeat-rich BAC clones were selected and sequenced. Whole genome shotgun (WGS) read mapping were performed to estimate genome proportion of the characterized repetitive elements. Fluorescence *in situ* hybridization (FISH) analysis was applied to the various stages of ginseng chromosomes with probes derived from repetitive and non-repetitive regions.

REFERENCES

Attele AS, Wu JA, Yuan CS (1999) Ginseng pharmacology: multiple constituents and multiple actions. Biochem Pharmacol 58:1685-1693.

Bang KH, Lee JW, Kim YC, Kim DH, Lee EH, Jeung JU (2010) Construction of genomic DNA library of Korean ginseng (*Panax ginseng* CA MEYER) and development of sequence-tagged sites. Biol Pharm Bull 33:1579-1588

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621-627.

Bennetzen JL, Ma JX, Devos K (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95:127-132.

Chen S, Luo H, Li Y, Sun Y, Wu Q, Niu Y, Song J, Lv A, Zhu Y, Sun C, Steinmetz A, Qian Z (2011) 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in *Panax ginseng*. Plant Cell Rep 30 (9):1593-1601.

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

Choi KT, Kim YT, Kwon WS (1992) Present status in development of new ginseng varieties. J Ginseng Res 16:164-168

Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. Trends Genet 5:103-107.

Gaut BS, d'Ennequin ML, Peek AS, Sawkins MC (2000) Maize as a model for the evolution of plant nuclear genomes. Proc Natl Acad Sci USA 97:7008-7015.

Ho ISH, 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.

Kidwell MG, Lisch D (1997) Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci USA 94:7704-7711

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.

Kim TD, Han JY, Huh GH, Choi YE (2011) Expression and functional characterization of three squalene synthase genes associated with saponin biosynthesis in *Panax ginseng*. Plant Cell Physiol 52:125-137.

Kim YJ, Shim JS, Krishna PR, Kim SY, In JG, Kim MK, Yang DC (2008) Isolation and characterization of a glutaredoxin gene from *Panax ginseng* C. A. Meyer. Plant Mol Biol Rep 26 (4):335-349.

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Ann Rev Genet 33:479-532

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

Lee JS, Lee SS, Lee JS, Ahn IO (2008) Effect of seed size and cultivars on the ratio of seed coat dehiscence and seedling performance in *Panax ginseng*. J Ginseng Res 32:257-263

Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ (2009) The Dynamic ups and downs of genome size evolution in Brassicaceae. Mol Biol Evol 26 (1):85-98.

Park HJ, Kim DH, Park SJ, Kim JM, Ryu JH (2012) Ginseng in traditional herbal prescriptions. J Ginseng Res 36:225-241.

Plunkett GM, Wen J, Lowry PP (2004) Infrafamilial classifications and characters in araliaceae: insights from the phylogenetic analysis of nuclear (ITS) and plastid (*trn*L-*trn*F) sequence data. Plant Syst Evol 245:1-39

Soltis DE, Soltis PS, Bennett MD, Leitch IJ (2003) Evolution of genome size in the angiosperms. Am J Bot 90:1596-1603.

Sun H, Wang HT, Kwon WS, Kim YJ, In JG, Yang DC (2011) A simple and rapid technique for the authentication of the ginseng cultivar, Yunpoong, using an SNP marker in a large sample of ginseng leaves. Gene 487:75-79.

Tansakul P, Shibuya M, Kushiro T, Ebizuka Y (2006) Dammarenediol-II synthase, the first dedicated enzyme for ginsenoside biosynthesis, in *Panax ginseng*. FEBS Lett 580:5143-5149.

Vitte C, Panaud O (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res 110:91-107.

Wen J, Zimmer EA (1996) Phylogeny and biogeography of *Panax* L (the ginseng genus, Araliaceae): Inferences from ITS sequences of nuclear ribosomal DNA. Mol Phylogenet Evol 6:167-177

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P,

Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973-982.

Wu B, Wang MZ, Ma YM, Yuan LC, Lu SF (2012) High-throughput sequencing and characterization of the small RNA transcriptome reveal features of novel and conserved microRNAs in *Panax ginseng*. Plos One 7

Yi TS, Lowry PP, Plunkett GM, Wen J (2004) Chromosomal evolution in Araliaceae and close relatives. Taxon 53:987-1005

Yun TK (2001) Brief introduction of *Panax ginseng* C.A. Meyer. J Korean med sci 16:S3-5

Zonneveld BJM, Leitch IJ, Bennett MD (2005) First nuclear DNA amounts in more than 300 angiosperms. Ann Bot 96:229-244.

CHAPTER I

Evolutionary relationship of *Panax ginseng* and *P. quinquefolius* inferred from sequencing and comparative analysis of expressed sequence tags

ABSTRACT

Gene and genome duplication events have long been accepted as driving forces in the evolution of angiosperms. *Panax ginseng* C. A. Meyer and *P. quinquefolius* L., which inhabit eastern Asia and eastern North America, respectively, are famous medicinal herbs and are similar in growth condition, morphological and genetic characteristics. However, no molecular data regarding their evolution has been available. In this study, expressed sequence tags (ESTs) were generated from a cDNA library of *P. ginseng* and comparative analyses were conducted to reveal genome-level duplication and speciation of *P. ginseng* and *P. quinquefolius* by in-depth comparison of paralog and orthlog ESTs. Sequencing and assembly of 5,760 clones from the cDNA library resulted in 4,552 uniESTs of *P. ginseng* and these were subjected to initial annotation steps. Comparative analysis was conducted with the uniESTs and transcriptome data of *P. quinquefolius* retrieved from the public database. Paralog pairing and analysis of the distribution of synonymous substitutions per synonymous site (Ks) showed two coincident peaks in both *Panax* species, implying two rounds of genome duplication in their common ancestor. Comparison of orthologs revealed one Ks peak that is younger than the two peaks identified from the analysis of paralogs. However, absolute dating of genome duplication and speciation events is needed to address caveats related to their long generation times, speculated to be more than eight to ten years in the wild. This is the first report regarding the evolutionary relationship of *Panax* species at the genome-wide level, and will provide a foundation to unravel the genome structure of the enigmatic genus *Panax* and the family Araliaceae.

INTRODUCTION

Iterative explosive genome duplication and long-term diploidization processes have been recognized as driving forces in biological evolution that provide gene gain and loss or gain and diversification such as pseudogenization, subfunctionalization and neofunctionalization (Ohno 1970; Stebbins 1966). A few rounds of whole genome duplication are commonly found in the evolution of most angiosperms (Soltis et al. 2009).

Expressed sequence tags (ESTs) are short nucleotide sequences obtained randomly from the 5' or 3' end of cDNA clones (Pandey and Lewitter 1999). Cost-effective high-throughput sequencing technology has enhanced the generation and accumulation of ESTs. Furthermore, large numbers of ESTs are available in public databases for numerous organisms (Nagaraj et al. 2007). ESTs are useful not only for obtaining information regarding gene expression, but also for genome annotation, gene structure prediction, comparative gene analysis between species and understanding genome evolution.

Sequence-level analyses have clarified and reinforced evidence for recent and ancient polyploidy. The number of synonymous substitutions per synonymous site (Ks) can serve as a divergence scale for two homologous gene sequences because synonymous substitution does not change amino acids and thus is subject to less selection pressure than non-synonymous substitution is. Therefore, Ks values have been used as a molecular clock for dating evolutionary events such as gene duplication and sequence divergence (Kimura 1980; Li 1997; Sterck et al. 2005). Among sequence-level analyses, EST analyses have contributed to elucidating the evolutionary histories of model plant species (Blanc and Wolfe 2004), major crop species (Schlueter et al. 2004), major angiosperm lineages (Cui et al. 2006), poplar (Sterck et al.

2005), Compositae (Barker et al. 2008), and apple (Sanzol 2010).

Korean ginseng (*Panax ginseng* C. A. Meyer) and American ginseng (*P. quinquefolius* L.) belong to the Araliaceae family and are herbaceous medicinal plants possessing pharmacologically active constituents known as ginsenosides (Ha et al. 2002; Sticher 1998). *P. ginseng* is well known as a panacea and has been used in eastern Asia for over 2000 years (Attele et al. 1999). Although *P. quinquefolius* has been used for medicinal purposes by native Americans for a long time, the disjunct distribution of *P. ginseng* and *P. quinquefolius* was revealed by the discovery of *P. quinquefolius* in North America in 1716 based on the knowledge of morphology and habitat of *P. ginseng* (Persons 1994; Wen 1999).The chromosome number of *P. ginseng* has been reported as $2n = 44$ (Sugiura 1936; Yang 1981) or $2n = 48$ (Harn and Whang 1963). The latter value ($2n = 48$) is more reliable and is supported by recent studies (Choi et al. 2009; Ko et al. 1993; Li et al. 1985; Yi et al. 2004). *P. quinquefolius* also possesses 2n = 48 chromosomes (Blair 1975; Choi et al. 2009; Ren et al. 1994; Yi et al. 2004). Meanwhile, some *Panax* species, such as *P. notoginseng* (Burk.) F. H. Chen, *P. sikkimensis* Ban., *P. sokpayensis* S. K. Sharma & M. K. Pandit and *P. bipinnatifidus* Seem., have half as many chromosomes ($2n = 24$) (Kondo et al. 1992; Sharma et al. 2010; Yi et al. 2004). Simple estimation based on chromosome numbers suggests that *P. ginseng* and *P. quinquefolius* probably evolved by tetraploidization between species having $2n = 24$ chromosomes (Wen and Zimmer 1996). In development of EST-based markers, more than two bands were always observed in hundreds of trials using different ESTs, indicating that the genomes of *Panax* species are highly duplicated (Choi et al. 2011; Kim et al. 2012b).

Phylogenetic relationships between *Panax* species have been reported using internal transcribed spacer (ITS) regions (Artyukova et al. 2005; Choi and Wen 2000; Wen et al. 2001; Wen and Zimmer 1996) and chloroplast DNA

(Choi and Wen 2000; Lee and Wen 2004; Zhu et al. 2003) data. Wen and Zimmer (1996) also described biogeographical distributions on the basis of phylogenetic and chromosomal data. Most of these studies have supported a close genetic relationship among *P. ginseng*, *P. japonicus* C. A. Meyer, and *P. quinquefolius*, with those species forming a monophyletic group. However, the polyploidy and speciation process of *Panax* species inhabiting eastern Asia and eastern North America have remained unclear due to insufficient data.

The objective of this study is the generation of ESTs from *P. ginseng* to provide transcriptional information and to investigate the evolutionary history and genetic relationship between *P. ginseng* and *P. quinquefolius* by comparative analysis of homologous genes in the species.

MATERIALS AND METHODS

Plant material, construction of a cDNA library and normalization

An entire one-year-old plant of *P. ginseng* cv. Chunpoong was collected from an experimental field of the Korea Ginseng Corporation Central Research Institute in Daejeon, South Korea, and immediately frozen in liquid nitrogen. Total RNA was extracted with TRIzol Reagent (Invitrogen Corp., USA). cDNA synthesis and construction of the cDNA library were performed according to the manufacturer's instructions (CoreBioSystem Co., Korea). Briefly, total RNA was first ligated with RNA oligomers, followed by reverse transcription with oligo-dT primers to prepare the first strand cDNA. Then, the second strand of cDNA was synthesized by PCR using RNA oligomers. Subsequently, double stranded cDNA was inserted into the unidirectional cloning vector pCNS-D2 (Oh et al. 2004). The library was constructed by transformation of the vector into *Escherichia coli* and the insert size of cDNA was confirmed by colony PCR with T7 and SP6 primers and digestion of isolated total plasmids with *Eco*RI and *Not*I. Normalization procedures were followed after the construction of the cDNA library. The library was infected with helper phage to induce single-stranded DNA (tracer) and tracer DNA was used to synthesize driver DNA by polymerase chain reaction. Tracer DNA was then hybridized with driver DNA, and single-stranded DNA was isolated from hybridized DNA duplexes. Isolated single-stranded DNA was converted to doubled-strand DNA and cloned into *Escherichia coli* Top10F'. Insert sizes were confirmed by digestion of isolated total plasmids with *Eco*RI and *Not*I.

Generation, collection and assembly of ESTs

A total of 5,760 clones were randomly selected from the *P. ginseng* normalized full-length cDNA library and sequenced from the 5' end using an ABI 3730xl capillary DNA sequencer. The generated sequences were assembled using the CAP3 program with default parameter values (Huang and Madan 1999) after vector trimming. ESTs shorter than 100 bp were discarded in the assembly process.

Root transcriptome data for *P. quinquefolius* (Sun et al. 2010) were collected from the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra/; accession no. SRX012184). The collected sequencing data were processed and assembled using the program Newbler version 2.3 (Roche, USA) with default parameter values. ESTs shorter than 50 bp were discarded in the assembly process and contigs shorter than 100 bp were discarded after the assembly process. Single reads not assigned to any contigs were not used for further analysis due to their short lengths and uncertainty.

Functional annotation of *P. ginseng* uniESTs

BLASTX (Altschul et al. 1997) searches of the *P. ginseng* uniESTs against the NCBI non-redundant database were conducted with a cut off value of 10^{-5}. The results were imported into the BLAST2GO program for functional annotation and GO analysis (Conesa and Gotz 2008). Briefly, mapping and annotation steps were carried out for retrieving and assigning GO terms to the uniESTs, and annotation results were summarized using plant GOslim mapping. Directed acyclic graphs for three categories, biological process (BP), molecular function (MF) and cellular component (CC), were drawn for visualization of the hierarchical structures of the GO terms and summary statistics.

Identification of paralogs and orthologs

All-against-all BLASTN (Altschul et al. 1997) searches were performed for each uniEST set to identify paralog and ortholog pairs. Two sequences were defined as a paralog pair if they aligned over more than 300 bp with more than 40 % identity (Blanc and Wolfe 2004). For ortholog pairs, reciprocal BLASTN searches were conducted and two sequences were defined as an ortholog pair if they showed best hits to each other and aligned over more than 300 bp (Blanc and Wolfe 2004).

Estimation of Ks values and elimination of redundant Ks values

Open reading frames (ORFs) were predicted based on the BLASTX results against the Arabidopsis protein database (TAIR10) (Swarbreck et al. 2008) using the program getorf in the EMBOSS package. If there was no available BLASTX result for a sequence, the longest of the predicted ORFs was selected. DNA alignments of the ORFs of paralog and ortholog pairs were conducted using CLUSTALW (Larkin et al. 2007), and PAL2NAL (Suyama et al. 2006) was used for aligning codons and removing gaps. The number of synonymous substitutions per synonymous site between the pairs was estimated using the program codeml in the PAML package (Yang 1997). For pairs of paralogs showing no synonymous substitution (Ks = 0), one of the two sequences was discarded from the dataset because they were likely to be derived from multiple entries of the same gene (Blanc and Wolfe 2004; Sanzol 2010). A gene family composed of n members can originate from n - 1 gene duplication events, but the number of possible pairwise comparisons within a gene family is n (n - 1)/2, which leads to multiple estimates of the gene duplications in paralogs (Blanc and Wolfe 2004; Maere et al. 2005;

Sterck et al. 2005). To eliminate the redundant Ks values, single linkage clustering and hierarchical clustering methods were adopted, as reported by Blanc and Wolfe (2004). Ks values over 2.0 were discarded in the final step because they are related to uncertainty about saturation of substitutions (Blanc and Wolfe 2004; Li 1997; Sanzol 2010).

RESULTS

Construction of cDNA library

Total RNA was extracted from the whole plant of one-year-old *P. ginseng* and a cDNA library was constructed after normalization and full-length cDNA enrichment. Insert sizes were estimated from 192 randomly picked clones by colony PCR and agarose gel electrophoresis. Of those, 190 had inserts, and the average insert size was ca. 1.12 kb, which is slightly longer than that of poplar (~1.09 kb) (Ralph et al. 2008), but shorter than those of bermuda grass (~1.2 kb) (Kim et al. 2008), *Arabidopsis* (~1.28 kb) (Asamizu et al. 2000), wheat (~1.3 kb) (Yang et al. 2009), onion (~1.6 kb) (Kuhl et al. 2004) and jatropha (~2.1 kb) (Natarajan et al. 2010).

Generation, collection, and assembly of ESTs

ESTs were generated by sequencing 5' ends of 5,760 cDNA clones. Among them, 5,677 gave high quality sequences covering a total length of 4.50 Mb with an average of 793 bp of trimmed reads. The high quality sequences were deposited in the NCBI dbEST (accession no. JK983480 - JK989156). Sequence assembly was conducted using the program CAP3, resulting in 4,552 uniESTs consisting of 878 contigs and 3,674 singlets. GS-FLX transcriptome data of *P. quinquefolius* (Sun et al. 2010) were retrieved from the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra/; accession no. SRX012184). A total of 209,747 pyrosequencing reads were assembled into 13,615 contigs using the program Newbler version 2.3 (Roche, USA). For further comparative analyses between the two species, the 4,552 uniESTs from *P. ginseng* and 13,615 contigs from *P. quinquefolius* were used (Table 1-1).

Table 1-1. Summary of assembled datasets used in this study.

	P. ginseng	*P. quinquefolius* [a]
Sequencing platform	ABI3730XL	GS-FLX Titanium
Number of EST reads used for assembly	5,677	209,747
Number of contigs	878	13,615
Number of singlets	3,674	-
Total length of uniESTs (bp)	3,833,501	7,416,933

[a]Sequences were obtained from the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra/; accession no. SRX012184)

Functional annotation of the uniEST set of *P. ginseng*

Of the 4,552 uniESTs, 4,302 showed BLASTX hits against the NCBI non-redundant database (Fig. 1-1). The most hits were with sequences from *Vitis vinifera* L., followed by sequences from *Arabidopsis thaliana* (L.) Heynh., and *Populus trichocarpa* Torr. & A. Gray. Almost 90% (223 out of 250) of the uniESTs that had no hits were over 300 bp in length and represent putative ginseng-specific novel transcripts. Of the 4,302 uniESTs that did give rise to hits, 4,035 were mapped and 3,366 sequences were annotated in the BLAST2GO (Conesa and Gotz 2008) analysis pipeline. Distributions of gene ontology (GO) terms at graph level 2 are shown in Fig. 1-2. The major categories from the three key biological domains MF, BP and CC in the level 2 GO distributions were metabolic process and cellular process, binding and catalytic activity, and cell and cell part, respectively (Fig. 1-2).

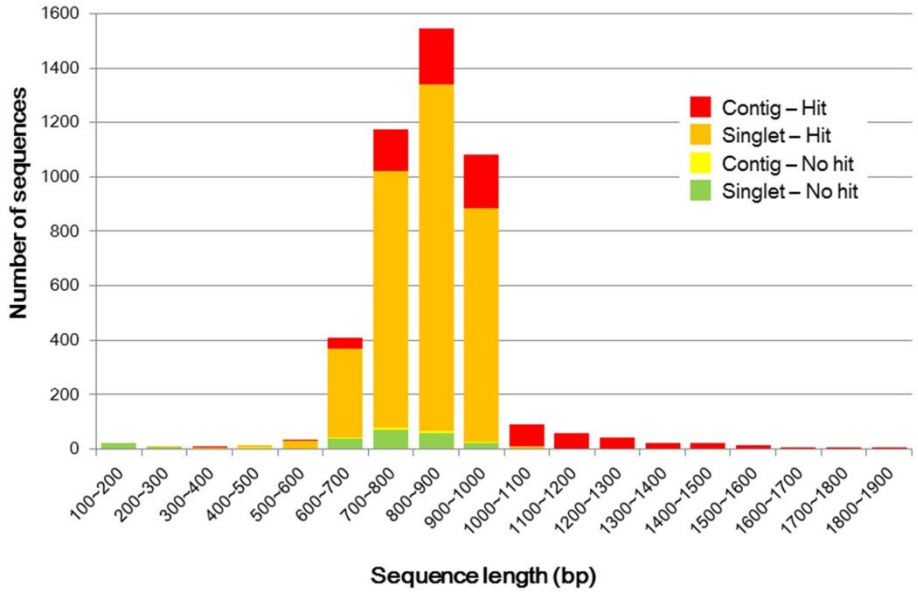

Figure 1-1. Length and BLASTX hit distribution of the *P. ginseng* uniEST set. Assembly of 5,677 *P. ginseng* ESTs from a normalized full-length cDNA library resulted in 878 contigs and 3,674 singlets. BLASTX searches against the NCBI non-redundant database showed significant hits for 4,302 uniESTs. Most of the uniESTs shorter than 300 bp showed no hits. 26 contigs and 197 singlets longer than 300 bp can be considered potential novel transcripts specific to *P. ginseng*.

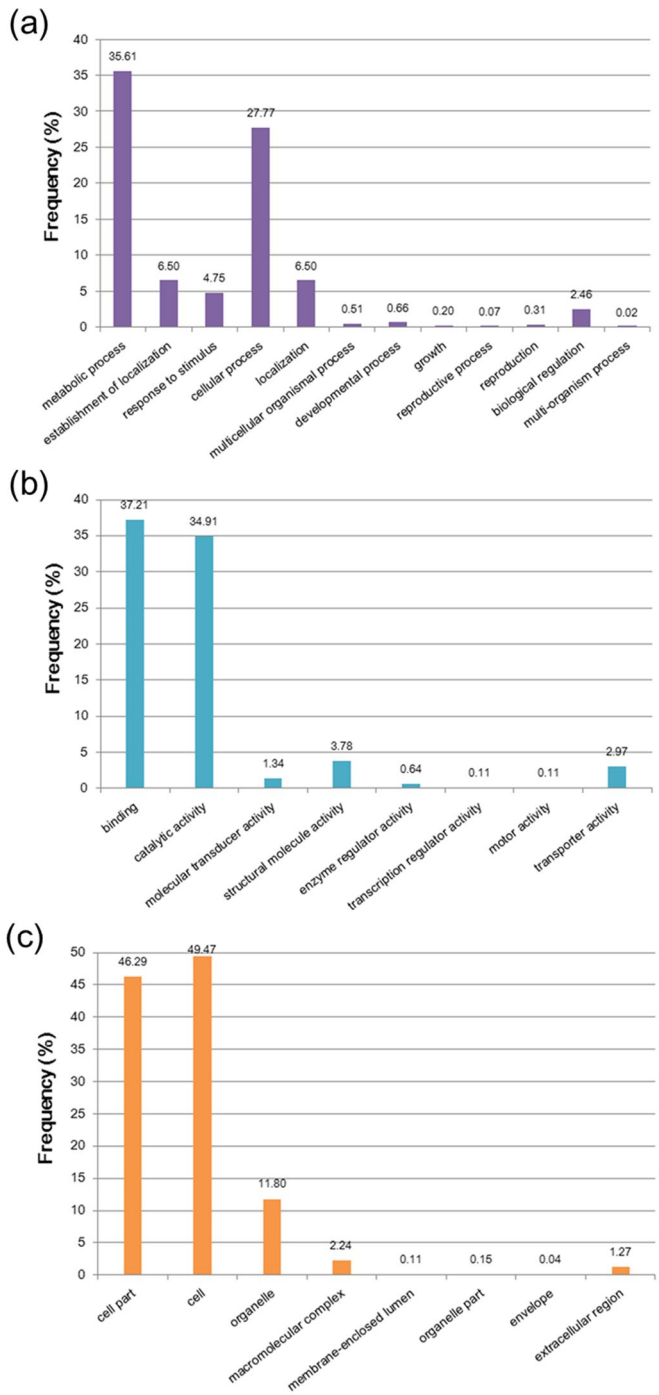

Figure 1-2. Functional classification of *P. ginseng* uniESTs based on gene ontology (GO) annotation. The program BLAST2GO was used to obtain GO terms for the *P. ginseng* uniESTs. Level 2 GO distribution of (a) biological process, (b) molecular function, and (c) cellular component categories.

Identification of paralogs and orthologs

Paralog and ortholog pairs were defined based on BLASTN results (Table 1-2). Gene families were identified using a single linkage clustering method in the paralogs. A total of 570 gene families were found, which consisted of 1,277 uniESTs from *P. ginseng* and 1,000 gene families consisting of 2,409 uniESTs among the *P. quinquefolius* paralogs (Table 1-2). The average gene family size was slightly smaller in *P. ginseng* than in *P. quinquefolius* (Table 1-2), which was presumed to be due to the difference in size of the datasets. For orthologs, e-values and bit scores of the reciprocal BLASTN results were used to identify the most significantly similar pairs. A total of 1,813 ortholog pairs were identified, accounting for 39.8 % and 13.3 % of the *P. ginseng* and *P. quinquefolius* uniEST sets, respectively.

Table 1-2. Summary statistics for identification of paralog pairs in *P. ginseng* and *P. quinquefolius* uniEST sets

	Number of paralog sequences	% in dataset	Number of gene families	Gene family Size	Number of Ks values
P. ginseng	1277	28.05	570	2.24	707
P. quinquefolius	2409	17.69	1000	2.41	1409

Peaks in the Ks distribution

Ks values were calculated for the predicted coding regions of each pair of paralogs and orthologs. Initial peaks at Ks = 0 to 0.1 and bell-shaped secondary peaks in the range of Ks = 0.25 to 0.45 were observed between the paralogs of the two species (Fig. 1-3a and Table 1-3). In the area of the secondary peaks, median Ks values were similar for both species, 0.3556 and 0.3394 in *P. ginseng* and *P. quinquefolius*, respectively. The similar pattern of secondary peaks found in the two species may indicates that they share a common whole genome duplication that occurred in their progenitor at the mode value of Ks = 0.3 to 0.4 (1R in Fig. 1-3a).

The Ks distribution between orthologs showed only the initial peak, and 71.17 % of the Ks values were between 0 and 0.1. Cumulative ratios of the paralogs at Ks = 0 to 0.1 were 49.93 % and 37.26 % in *P. ginseng* and *P. quinquefolius*, respectively. Very recent genome duplications called neopolyploidy can be obscured because small-scale gene or segmental duplications can be intermingled to form an initial peak (Barker et al. 2008; Cui et al. 2006). Therefore, the Ks values were re-plotted in the range of 0 to 0.2 using a smaller interval size (0.005) to discover whether there was a hidden peak (Fig. 1-3b and Table 1-4). Ks distributions between paralogs with the smaller interval revealed new peaks in the range of Ks = 0.02 to 0.04 in both species. These peaks might represent recent genome duplication in both species (2R in Fig. 1-3b). Ks distribution between the orthologs using the smaller interval revealed a small ancillary peak in the range of Ks = 0.010 to 0.015 (Fig. 1-3b and Table 1-4). This peak might represent a divergence event between *P. ginseng* and *P. quinquefolius*.

Taken together, the Ks distribution analysis suggests that the common ancestor of *P. ginseng* and *P. quinquefolius* may have had undergone two large-scale genome level duplication events (one in ancient times and the

other recently) and that later the two species have diverged.

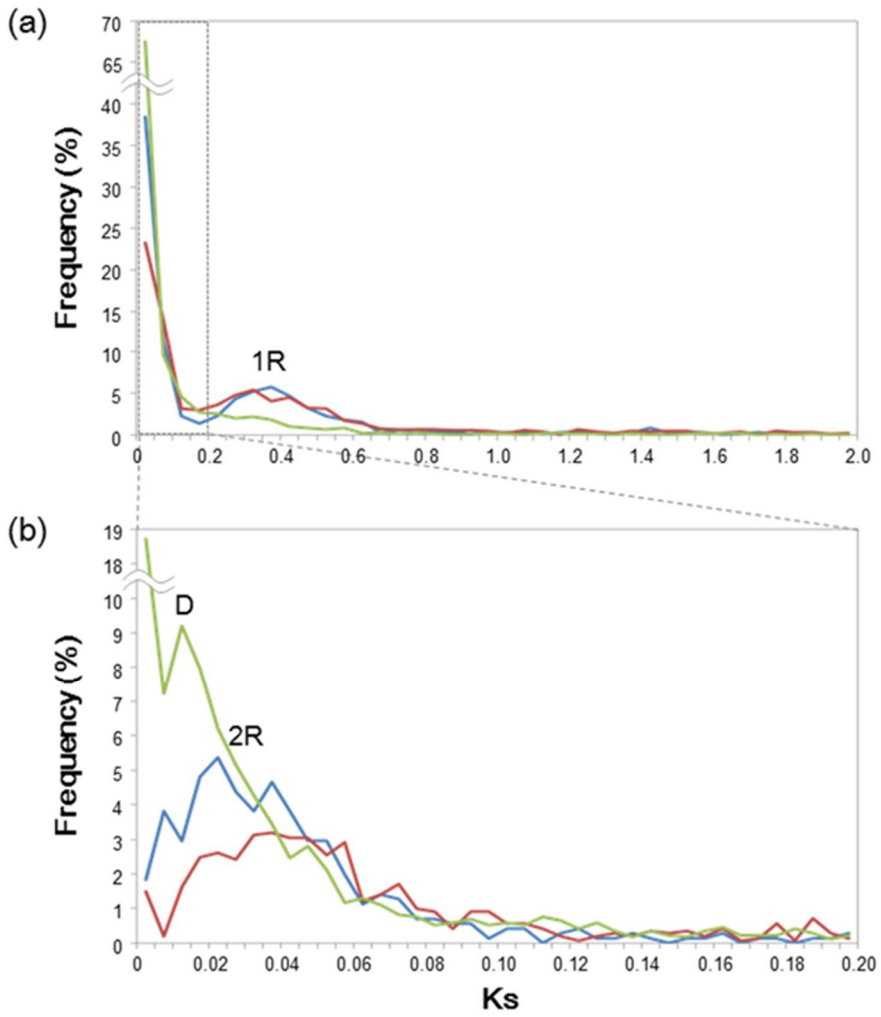

Figure 1-3. Ks distributions of paralogs and orthologs in *P. ginseng* and *P. quinquefolius* uniEST sets. Ks frequencies of *P. ginseng* paralogs, *P. quinquefolius* paralogs, and orthologs between *P. ginseng* and *P. quinquefolius* were plotted with blue, red, and green lines, respectively. Ks distributions were plotted using two interval scales: (a) plot in the range of 0 to 2 with 0.05 intervals, (b) 0 to 0.2 with 0.005 intervals. 1R and 2R indicate the peaks for the first and second rounds of genome duplication, respectively. D indicates the peak for divergence of the two species.

Table 1-3. Distribution of Ks values in the range of 0 to 2 with 0.05 intervals

Ks class	*P. ginseng* paralogs		*P. quinquefolius* paralogs		Orthologs	
	Frequency	%	Frequency	%	Frequency	%
0 - 0.05	272	38.47	328	23.28	1151	67.55
0.05 - 0.1	81	11.46	197	13.98	164	9.62
0.1 - 0.15	16	2.26	45	3.19	79	4.64
0.15 - 0.2	10	1.41	42	2.98	47	2.76
0.2 - 0.25	16	2.26	51	3.62	44	2.58
0.25 - 0.3	31	4.38	67	4.76	35	2.05
0.3 - 0.35	37	5.23	76	5.39	37	2.17
0.35 - 0.4	41	5.80	58	4.12	31	1.82
0.4 - 0.45	33	4.67	64	4.54	17	1.00
0.45 - 0.5	23	3.25	46	3.26	14	0.82
0.5 - 0.55	16	2.26	45	3.19	12	0.70
0.55 - 0.6	13	1.84	25	1.77	14	0.82
0.6 - 0.65	11	1.56	20	1.42	4	0.23
0.65 - 0.7	2	0.28	11	0.78	6	0.35
0.7 - 0.75	4	0.57	10	0.71	2	0.12
0.75 - 0.8	4	0.57	10	0.71	5	0.29
0.8 - 0.85	2	0.28	9	0.64	2	0.12
0.85 - 0.9	3	0.42	8	0.57	1	0.06
0.9 - 0.95	1	0.14	8	0.57	0	0.00
0.95 - 1	2	0.28	7	0.50	3	0.18
1 - 1.05	2	0.28	3	0.21	2	0.12
1.05 - 1.1	2	0.28	8	0.57	0	0.00
1.1 - 1.15	1	0.14	6	0.43	3	0.18
1.15 - 1.2	2	0.28	0	0.00	4	0.23

Table 1-3. (Continued)

Ks class	*P. ginseng* paralogs		*P. quinquefolius* paralogs		Orthologs	
	Frequency	%	Frequency	%	Frequency	%
1.2 - 1.25	3	0.42	9	0.64	2	0.12
1.25 - 1.3	1	0.14	5	0.35	2	0.12
1.3 - 1.35	0	0.00	3	0.21	1	0.06
1.35 - 1.4	2	0.28	7	0.50	4	0.23
1.4 - 1.45	6	0.85	5	0.35	1	0.06
1.45 - 1.5	2	0.28	7	0.50	1	0.06
1.5 - 1.55	3	0.42	7	0.50	2	0.12
1.55 - 1.6	2	0.28	2	0.14	3	0.18
1.6 - 1.65	0	0.00	3	0.21	2	0.12
1.65 - 1.7	1	0.14	5	0.35	3	0.18
1.7 - 1.75	2	0.28	1	0.07	0	0.00
1.75 - 1.8	1	0.14	7	0.50	2	0.12
1.8 - 1.85	2	0.28	4	0.28	0	0.00
1.85 - 1.9	2	0.28	4	0.28	2	0.12
1.9 - 1.95	0	0.00	2	0.14	1	0.06
1.95 - 2	1	0.14	3	0.21	1	0.06
>2	54	7.64	191	13.56	-	-
Total	707	100	1409	100	1704	100

Table 1-4. Distribution of Ks values in the range of 0 to 0.2 with 0.005 intervals

Ks class	*P. ginseng* paralogs		*P. quinquefolius* paralogs		Orthologs	
	Frequency	%	Frequency	%	Frequency	%
0 - 0.005	13	1.84	21	1.49	319	18.72
0.005 - 0.01	27	3.82	3	0.21	123	7.22
0.01 - 0.015	21	2.97	23	1.63	157	9.21
0.015 - 0.02	34	4.81	35	2.48	136	7.98
0.02 - 0.025	38	5.37	37	2.63	106	6.22
0.025 - 0.03	31	4.38	34	2.41	88	5.16
0.03 - 0.035	27	3.82	44	3.12	73	4.28
0.035 - 0.04	33	4.67	45	3.19	59	3.46
0.04 - 0.045	27	3.82	43	3.05	42	2.46
0.045 - 0.05	21	2.97	43	3.05	48	2.82
0.05 - 0.055	21	2.97	36	2.56	36	2.11
0.055 - 0.06	14	1.98	41	2.91	20	1.17
0.06 - 0.065	8	1.13	17	1.21	22	1.29
0.065 - 0.07	10	1.41	20	1.42	19	1.12
0.07 - 0.075	9	1.27	24	1.70	14	0.82
0.075 - 0.08	5	0.71	14	0.99	13	0.76
0.08 - 0.085	5	0.71	13	0.92	9	0.53
0.085 - 0.09	4	0.57	6	0.43	10	0.59
0.09 - 0.095	4	0.57	13	0.92	12	0.70
0.095 - 0.1	1	0.14	13	0.92	9	0.53
0.1 - 0.105	3	0.42	8	0.57	10	0.59
0.105 - 0.11	3	0.42	8	0.57	9	0.53
0.11 - 0.115	0	0.00	6	0.43	13	0.76

Table 1-4. (Continued)

Ks class	*P. ginseng* paralogs		*P. quinquefolius* paralogs		Orthologs	
	Frequency	%[a]	Frequency	%[a]	Frequency	%[a]
0.115 - 0.12	2	0.28	3	0.21	11	0.65
0.12 - 0.125	3	0.42	1	0.07	7	0.41
0.125 - 0.13	1	0.14	3	0.21	10	0.59
0.13 - 0.135	1	0.14	4	0.28	6	0.35
0.135 - 0.14	2	0.28	3	0.21	3	0.18
0.14 - 0.145	1	0.14	5	0.35	6	0.35
0.145 - 0.15	0	0.00	4	0.28	4	0.23
0.15 - 0.155	1	0.14	5	0.35	3	0.18
0.155 - 0.16	1	0.14	3	0.21	6	0.35
0.16 - 0.165	2	0.28	6	0.43	8	0.47
0.165 - 0.17	0	0.00	1	0.07	4	0.23
0.17 - 0.175	1	0.14	2	0.14	4	0.23
0.175 - 0.18	1	0.14	8	0.57	4	0.23
0.18 - 0.185	0	0.00	1	0.07	7	0.41
0.185 - 0.19	1	0.14	10	0.71	5	0.29
0.19 - 0.195	1	0.14	4	0.28	2	0.12
0.195 - 0.2	2	0.28	2	0.14	4	0.23
Total	379	53.61	612	43.44	1441	84.57

[a]Percentage of frequency in total Ks values

DISCUSSION

Two rounds of genome duplications are found in both *P. ginseng* and *P. quinquefolius*

Panax ginseng and *P. quinquefolius* have been assumed to be tetraploid, but there had been no reports on their genome-level duplication history. In this study, the Ks distributions were plotted for paralogs to identify duplication events. Although the uniEST datasets of the two species were generated by different sequencing methods, different in data size, and assembled with different assembler programs, coincident accumulation patterns which may be derived from common evolutionary relationship between the two species were observed for the Ks values between paralogs in each species.

Even though initial peaks contain younger gene duplications such as tandem and segmental duplications because these occur continuously and frequently (Blanc and Wolfe 2004; Cui et al. 2006), the peaks observed in Fig. 1-3b may be considered reliable indicators of genome-level duplication events based on three lines of evidence. First, large fractions, 19.8 % and 13.1 % of the total Ks values from *P. ginseng* and *P. quinquefolius*, respectively, were found in the range of Ks = 0.02 to 0.04. Second, the apparent peaks were identified at a narrow interval in both species even though the peaks were not observed in other species without recent genome duplication (Blanc and Wolfe 2004). Third, the noticeable peak for the Ks value between orthologs was also identified only at the shortest interval. Likelihood of recent whole genome duplication can be also supported by my previous attempts at EST-derived marker development. PCR of several hundreds of EST-derived microsatellite markers almost always gave rise to multiple bands, which included the expected bands along with unexpected additional bands (Choi et

al. 2011, Kim et al. 2012b). Genomic *in situ* hybridization (GISH) results for *P. ginseng* and *P. quinquefolius* also supported that the origin of polyploidy is identical between the two species (Choi et al. 2009). The irregular pattern of the paralog Ks graphs shown in Fig. 1-3b may be caused by too short interval and intermingling of the recent genome duplication with tandem or segmental duplications. Further genome-level analysis using genome sequences or large scale transcriptome data will provide more specific details regarding the recent duplication event.

Panax quinquefolius recently diverged from *P. ginseng* by migration

Close floristic affinity and disjunct distribution of congeneric plants between eastern Asia and eastern North America have long been recognized and discussed with regard to clarification of their evolution (Boufford and Spongberg 1983; Guo 1999; Li 1952; Raven 1972; Thorne 1972; Wen 1999). More than 65 genera of seed plants were confirmed to be disjunctly distributed in different continents (Wen 1999), and *P. ginseng* and *P. quinquefolius* are representative disjunct species between eastern Asia and eastern North America. The close genetic relationship between the two species has been demonstrated by not only similarities in physiological, morphological, chromosomal characteristics and their medicinal components, but also compatibility in interspecific crosses. Wen and Zimmer (1996) suggested evolutionary relationships among the *Panax* species based on the chromosomal data and ITS phylogeny. They inferred that *P. quinquefolius* in eastern North America and *Panax* relatives in eastern Asia may have diverged recently based on their low ITS sequence divergence, highly similar morphological characteristics and identical chromosome numbers.

In the distribution of Ks values between orthologs, the initial peak

persisted when the interval sizes were gradually shortened. Moreover, 77.19 % of ortholog pairs showed Ks values in the range of 0 to 0.1. This result indicates an extraordinarily close genetic relationship between the two disjunctly distributed species that is also supported by their similar morphological traits and the particular growth conditions shared between the two species, such as growing in shade, moist soil and a cool climate. This corresponds with Wen and Zimmer's report (1996) showing extremely high ITS sequence similarity (99.35 %) between the two species. The previous study based on comparison of 44,563 bp of chloroplast intergenic spacer (IGS) sequence showed that *P. quinquefolius* is very closely related to *P. ginseng* with a nucleotide substitution rate of 0.0009, whereas *P. notoginseng* is diverged from the other two species, with a nucleotide substitution rate of 0.0039 even though it inhabits China (Kim et al. 2012a). Collectively, it is likely that *P. quinquefolius* diverged from *P. ginseng*, because eastern Asia is the center of diversity for *Panax* species (Wen and Zimmer 1996), and the two species have the same number of chromosomes and share genome duplication history.

Although it is difficult to elucidate the migration route because *P. ginseng* and *P. quinquefolius* are disjunctly distributed only in eastern Asia and eastern North America without intermediate species in any other region, the most feasible route is through the Bering land bridge (Hopkins 1967; Wen 1999). The Bering land bridge has connected eastern Asia and western North America many times since the Mesozoic era (Wen 1999). During the last 2 million years, climatic changes between cold glacial and interglacial periods have made the bridge appear and disappear repeatedly, and the glacial periods created frozen regions, whereas refuges preserved genetic variation for organisms (see details in DeChaine 2008). Thus, the Bering land bridge has been proposed as a migration route for flora between eastern Asia and eastern North America for a long time (DeChaine 2008; Hopkins 1967; Qian and

Ricklefs 2000; Wen 1999).

Estimation of molecular clocks for evolutionary events

Molecular clocks of critical genome evolution events can be inferred from Ks values by adopting a proper synonymous substitution rate (λ). When the λ of 6.1×10^{-9} substitutions/synonymous site/year was adopted, which was calculated by analyses of multiple genes in vascular plants (Lynch and Conery 2000), the ancient duplication event (mode Ks = 0.3 to 0.4, 1R in Fig. 1-3a) was estimated to have occurred at ca. 24.6 to 32.8 MYA. The recent genome duplication event revealed by the young paralog peaks at Ks = 0.02 to 0.04 (2R in Fig. 1-3b) was estimated at ca. 1.6 to 3.3 MYA in their common ancestor. On the basis of the secondary peak that appeared in the ortholog Ks distribution (mode Ks = 0.010 to 0.015, D in Fig. 1-3b), *P. ginseng* and *P. quinquefolius* diverged ca. 0.8 to 1.2 MYA. The previous estimation of molecular clocks based on chloroplast IGS sequences showed that *P. quinquefolius* diverged from *P. ginseng* at ca. 0.29 MYA which is slightly more recent than the estimate from this study (Kim et al. 2012a).

Molecular clock estimates can be adjusted by application of proper λ values because the values vary significantly in different plants, which mean λ is lineage-specific (Fawcett et al. 2009). Gaut et al. (Gaut et al. 2011) summarized and defined the three main components that affect variation of plant nucleotide substitution rates and their interaction: sites (sequence context), genes (function related) and lineages (annual vs. perennial). In the case of *Populus*, which needs four to six years of growth time for reproduction under appropriate conditions, λ was estimated to be about six times slower than that of *Arabidopsis* (Sterck et al. 2005; Tuskan et al. 2006). There has been no clear report for generation time in *P. ginseng* and *P. quinquefolius* under wild conditions, although it takes a minimum of three

years under cultivation conditions. The growth rate of ginseng under wild conditions is extremely slow compared to that in cultivation and dormancy breaking presumably takes about two years in the wild. *Panax* species are herbal plants with a long lifespan, growing up to a hundred years under shaded conditions. Growth of the plant can be stopped by damage to young shoots in the growing season and the plant can remain dormant for one to two years. Considering that, their generation time in the wild could be regarded as more than eight to ten years. Therefore, dating of the evolutionary events should be adjusted based on more clues, such as fossil records. It is possible that the evolutionary events described here actually took place earlier than the estimated dates.

High-quality ESTs from *P. ginseng* were produced, which can be utilized as genetic resources and for gene discovery, functional study, and improvement of breeding systems in ginseng and related species. Furthermore, the most important large evolutionary events in *P. ginseng* and *P. quinquefolius* were characterized by comparative analysis of EST homologs. Two rounds of genome duplication events are common to both *P. ginseng* and *P. quinquefolius,* indicating the two events occurred in their common ancestor and then the two species diverged very recently via migration. Although absolute dating still remains obscure due to different rates of synonymous substitution in different species, this is the first study to establish the evolutionary process in *Panax* species. The data contributes to understanding the genome structure in ginseng as well as the evolutionary relationship of the genus *Panax* and the family Araliaceae. Unraveling the entire genome sequence of *P. ginseng* is on the way using next generation sequencing technology (Choi et al. 2010). This study provides a good base for understanding the entire genome even though it points out the difficulties in sequencing the *P. ginseng* genome. Upcoming genome-level sequence analyses will provide further information on the genome structure and

evolutionary process in *P. ginseng* and related species.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang JH, 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

Artyukova EV, Gontcharov AA, Kozyrenko MM, Reunova GD, Zhuravlev YN (2005) Phylogenetic relationships of the far eastern Araliaceae inferred from ITS sequences of nuclear rDNA. Genetika 41:649-658

Asamizu E, Nakamura Y, Sato S, Tabata S (2000) A large scale analysis of cDNA in *Arabidopsis thaliana*: generation of 12,028 non-redundant expressed sequence tags from normalized and size-selected cDNA libraries. DNA Res 7:175-180

Attele AS, Wu JA, Yuan CS (1999) Ginseng pharmacology: multiple constituents and multiple actions. Biochem Pharmacol 58:1685-1693

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore W, Knapp SJ, Rieseberg LH (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol Biol Evol 25:2445-2455

Blair A (1975) Karyotype of five plant species with disjunct distributions in Virginia and the Carolinas. Am J Bot 62:833-837

Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667-1678

Boufford DE, Spongberg SA (1983) Eastern Asian-eastern North American phytogeographical relationships-A history from the time of Linnaeus to the twentieth century. Ann Mo Bot Gard 70:423-439

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

Choi HW, Koo DH, Bang KH, Paek KY, Seong NS, Bang JW (2009) FISH and GISH analysis of the genomic relationships among *Panax* species. Genes Genomics 31:99-105

Choi HI, Kim NH, Jung JY, Park HM, Park HS, Park JY, Lee J, Park J, Lee J, Choi BS, Ahn IO, Lee JS, Choi D,Yang TJ (2010) Current status of Korean ginseng (*Panax ginseng*) genome mapping and sequencing. In Advances in

Ginseng Research - Proceedings of 10th International Symposium on Ginseng: 13-16 Sep 2010; The Korean Society of Ginseng, Seoul, Korea. Edited by Yang DC, Kim SK, 762-778

Choi HK, Wen J (2000) A phylogenetic analysis of *Panax* (Araliaceae): integrating cpDNA restriction site and nuclear rDNA ITS sequence data. Plant Syst Evol 224:109-120

Conesa A, Gotz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008:619832

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16:738-749

DeChaine EG (2008) A bridge or a barrier? Beringia's influence on the distribution and diversity of tundra plants. Plant Ecol Div 1:197-207

Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci USA 106:5737-5742

Gaut B, Yang L, Takuno S, Eguiarte LE (2011) The patterns and causes of variation in plant nucleotide substitution rates. Ann Rev Ecol Evol Syst 42:245-266

Guo Q (1999) Ecological comparisons between Eastern Asia and North America: historical and geographical perspectives. J Biogeography 26:199-206

Ha WY, Shaw PC, Liu J, Yau FCF, Wang J (2002) Authentication of *Panax ginseng* and *Panax quinquefolius* using amplified fragment length polymorphism (AFLP) and directed amplification of minisatellite region DNA (DAMD). J Agric Food Chem 50:1871-1875

Harn C, Whang J (1963) Development of female gametophyte of *Panax ginseng*. Korean J Bot 6:3-6

Hopkins DM (1967) The Bering Land Bridge. Stanford University Press, Stanford, CA

Huang XQ, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9 (9):868-877

Kim C, Jang CS, Kamps TL, Robertson JS, Feltus FA, Paterson AH (2008) Transcriptome analysis of leaf tissue from Bermudagrass (*Cynodon dactylon*) using a normalised cDNA library. Funct Plant Biol 35:585-594

Kim JH, Jung JY, Choi HI, Kim NH, Park JY, Lee Y, Yang TJ (2012a) Diversity and evolution of major *Panax* species revealed by scanning the entire chloroplast

intergenic spacer sequences. Gen Res Crop Evol doi:10.1007/s10722-012-9844-4

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

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120

Ko KM, Song JJ, Hwang B, Kang YH (1993) Cytogenetic and histological characteristics of ginseng hairy root transformed by *Agrobacterium rhizogenes*. Korean J Bot 36:75-81

Kondo K, Taniguchi K, Tanaka R, Gu ZJ (1992) Karyomorphological studies in Chinese plant-species involving the Japanese floristic element, I. American Camellia Yearbook 1992:131-156

Kuhl JC, Cheung F, Yuan Q, Martin W, Zewdie Y, McCallum J, Catanach A, Rutherford P, Sink KC, Jenderek M, Prince JP, Town CD, Havey MJ (2004) A unique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders Asparagales and Poales. Plant Cell 16:114-125

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and clustal X version 2.0. Bioinformatics 23:2947-2948

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

Li FC, Sun X, Gong XC (1985) The analysis of the chromosomal morphology and Giemsa C-banding pattern in Ginseng. Sci Agric Sin (5):31-35

Li HL (1952) Floristic relationships between eastern Asia and eastern North America. Trans Am Philos Soc 42:371-429.

Li WH (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA 102:5454-5459

Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinformatics 8:6-21

Natarajan P, Kanagasabapathy D, Gunadayalan G, Panchalingam J, Shree N, Sugantham PA, Singh KK, Madasamy P (2010) Gene discovery from *Jatropha curcas* by sequencing of ESTs from normalized and full-length enriched cDNA library from developing seeds. BMC Genomics 11:606

Oh JH, Sohn HY, Kim JM, Kim YS, Kim NS (2004) Construction of multi-purpose vectors, pCNS and pCNS-D2, are suitable for collection and functional study of large-scale cDNAs. Plasmid 51:217-226.

Ohno S (1970) Evolution by gene duplication. Springer, New York, NY

Pandey A, Lewitter F (1999) Nucleotide sequence databases: a gold mine for biologists. Trends Biochem Sci 24:276-280

Persons WS (1994) American ginseng: green gold. Bright Mountain Books, Asheville, NC

Qian H, Ricklefs RE (2000) Large-scale processes and the Asian bias in species diversity of temperate plants. Nature 407:180-182

Ralph SG, Chun HJ, Cooper D, Kirkpatrick R, Kolosova N, Gunter L, Tuskan GA, Douglas CJ, Holt RA, Jones SJ, Marra MA, Bohlmann J (2008) Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding. BMC Genomics 9:57

Raven PH (1972) Plant species disjunctions: A summary. Ann Mo Bot Gard 59:234-246

Ren Y, Xu Q, Piao T, Sun B (1994) Karyotype analysis of American ginseng. J Jilin Agric Univ 16:43-46

Sanzol J (2010) Dating and functional characterization of duplicated genes in the apple (*Malus domestica* Borkh.) by analyzing EST data. BMC Plant Biol 10:87

Sharma SK, Bisht MS, Pandit MK (2010) Synaptic mutation-driven male sterility in *Panax sikkimensis* Ban. (Araliaceae) from Eastern Himalaya, India. Plant Syst Evol 287:29-36

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome 47:868-876

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng CF, Sankoff D, dePamphilis CW, Wall PK, Soltis PS (2009) Polyploidy and angiosperm

diversification. Am J Bot 96:336-348

Stebbins GL (1966) Chromosomal variation and evolution. Science 152:1463-1469

Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, Van de Peer Y (2005) EST data suggest that poplar is an ancient polyploid. New Phytol 167:165-170

Sticher O (1998) Getting to the root of ginseng. Chemtech 28:26-32

Sugiura T (1936) A list of chromosome numbers in angiospermous plants. II. In: Proceedings of the Imperial Academy 12:144-146

Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EM, Chen S (2010) *De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11:262

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609-W612

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res 36:D1009-D1014

Thorne RF (1972) Major disjunctions in the geographic ranges of seed plants. Q Rev Biol 47:365-411

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen G, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596-1604

Wen J (1999) Evolution of eastern Asian and eastern North American disjunct distributions in flowering plants. Annu Rev Ecol Syst 30:421-455

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 Bot 26:144-167

Wen J, Zimmer EA (1996) Phylogeny and biogeography of *Panax* L. (the ginseng genus, Araliaceae): Inferences from ITS sequences of nuclear ribosomal DNA. Mol Phylogenet Evol 6:167-177

Yang DQ (1981) The cyto-taxonomic studies on some species of *Panax* L. Acta Phytotax Sin 19:298-303

Yang D, Tang ZH, Zhang LP, Zhao CP, Zheng YL (2009) Construction, characterization, and expressed sequence tag (EST) analysis of normalized cDNA library of thermo-photoperiod-sensitive genic male sterile (TPGMS) wheat from spike developmental stages. Plant Mol Biol Rep 27:117-125

Yang ZH (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556

Yi T, Lowry PP, Plunkett GM, Wen J (2004) Chromosomal evolution in Araliaceae and close relatives. Taxon 53:987-1005

Zhu S, Fushimi H, Cai S, Komatsu K (2003) Phylogenetic relationship in the genus *Panax*: inferred from chloroplast *trn*K gene and nuclear 18S rRNA gene sequences. Planta Med 69:647-653

CHAPTER II

Repeat-rich bacterial artificial chromosome (BAC) sequences and fluorescence *in situ* hybridization (FISH) unveil major heterochromatic components and alloploid-like genome structure in *Panax ginseng*

ABSTRACT

Ginseng (*Panax ginseng* C. A. Meyer) has a large genome size estimated to be 3,120 Mb that experienced two rounds of polyploidy events. However, the genome composition of ginseng has been veiled due to insufficient data. This is the first report to characterize major component of the ginseng genome by analysis of repeat-rich bacterial artificial chromosome (BAC) sequences. In-depth sequence analysis revealed that long terminal repeat retrotransposons (LTR-RT) and their derivatives were major constituents occupying more than 80% in the selected three BAC clones. *Ty3/Gypsy*-like elements (*PgDel*, *PgTat*, and *PgAthila*) were more predominant than *Ty1/Copia* (*PgTork* and *PgOryco*). Mapping of 30 Gb whole genome shotgun (WGS) reads showed that 38% of the reads was mapped on the BAC sequences and 34% of the reads were mapped on the LTR-RTs characterized in this study, which indicates high abundance of the LTR-RTs in the ginseng genome. Non-repetitive regions were apparently distinguished by low mapping frequency. Complex insertion patterns, high genome proportion, and widespread distribution of fluorescence *in situ* hybridization (FISH) signals of the LTR-RTs indicate that the LTR-RTs have been contributed to the expansion of the ginseng genome. Particularly, the *PgDel1* elements played major roles in expanding heterochromatic regions as well as remodeling euchromatic regions. The *PgDel2* elements showed biased intensity of FISH signals on half the total chromosomes, which demonstrate the alloploidy-like nature of ginseng. Insertion time of the LTR-RTs showed that LTR-RTs may proliferate after the recent polyploidy event in the ginseng genome. Using a gene structure in the BAC sequence, four paralogous gene copies were identified by survey of the ginseng whole genome draft sequence that they are associated with the two rounds of polyploidy in ginseng. While the FISH signals of the gene structure

revealed as four signals that they were derived from two of the four gene copies. This data suggest that LTR-RTs have been closely related to evolution of the ginseng genome, particularly, genome expansion and the polyploidy events.

INTRODUCTION

Ginseng (*Panax ginseng* C. A. Meyer) has been widely used as valuable medicine in East Asia for millennia, which is a perennial herbal plant and a representative species in the genus *Panax* and family Araliaceae (Park et al. 2012a; Yun 2001). The family Araliaceae has approximately 55 genera and 1,500 species including a lot of medicinal and ornamental plants (Wen et al. 2001).

Ginseng is most well-researched species in Araliaceae. Studies of ginseng have been focused mainly on its medicinal components and their efficacy (Kim and Park 2011; Park et al. 2012a; Qi et al. 2011). However, genetic and genomic studies such as functional analyses of genes (Kim et al. 2008; Kim et al. 2011; Tansakul et al. 2006), marker development (Choi et al. 2011; Jo et al. 2011; Kim et al. 2012; Sun et al. 2011), construction of DNA library (Bang et al. 2010; Hong et al. 2004), and transcriptome sequencing (Chen et al. 2011; Wu et al. 2012) have been recently initiated.

Ginseng has $2n = 48$ chromosomes (Waminal et al. 2012) that has been regarded as tetraploid ($2n = 4x$), based on the chromosome numbers of $2n = 24$ and 48 in the genus *Panax* and generally in the family Araliaceae (Wen 1999; Yi et al. 2004). On the other hand, $x = 6$ was suggested as an ancient basal chromosome number of Araliaceae that was originally postulated for the order Apiales (Plunkett et al. 2004; Yi et al. 2004). This hypothesis was proved by the recent study that showed two rounds of polyploidy events co-occurring in both ginseng and American ginseng, which were inferred by calculation of the number of synonymous substitutions per synonymous site in coding regions (Ks) of paralogous gene pairs from expressed sequence tags (Choi et al. 2012). Although two rounds of genome duplications have been suggested from molecular data, further details of the ginseng genome are still

hidden due to the lack of genomic sequence data.

The estimated genome size of ginseng was reported as 3.12 Gb on a haploid chromosomes equivalent (Hong et al. 2004). My own data have suggested the ginseng genome size of approximately 3.5 Gb based on k-mer coverage in the genome assembly and flow cytometry analysis (unpublished). American ginseng (*P. quinquefolius* L.), which is closely related to ginseng genetically as well as morphologically (Choi et al. 2012; Wen 1999), has an estimated genome size of 4.91 Gb (Obae & West, 2012). Repetitive DNAs have been accepted as an important factor for genome size variation (Bennetzen 2005; Gaut et al. 2000; Lysak et al. 2009), which can be classified into two categories: tandem repeats and transposable elements (Kubis et al. 1998). Tandem repeats are constituted with tens or thousands of repeat units and usually found in specific genome regions, such as centromere or telomere (Heslop-Harrison 2000; Kubis et al. 1998; Lim et al. 2007; Yang et al. 2005). Transposable elements are ubiquitous in genomes of most living organisms, which are mobile DNA fragments that can move and replicate inside the genome (Kidwell and Lisch 1997). They can be divided into two classes by their transposition mechanisms, class I: retrotransposons with a copy-and-paste mode and class II: DNA transposons with a cut-and-paste mode (Finnegan 1989; Wicker et al. 2007). Of them, long terminal repeat retrotransposons (LTR-RTs) belonging to class I are known as the most abundant in plant genomes (Feschotte et al. 2002; Kumar and Bennetzen 1999; Wicker et al. 2007). Previous reports about analyses of repetitive DNA sequences using the low $C_o t$ method (Ho and Leung 2002) and bacterial artificial chromosome (BAC) -end sequences (Hong et al. 2004) provided partial evidence for the high abundance of retroelements in the ginseng genome. However, the genome composition and characteristics of ginseng have been veiled due to insufficient data. Identification of major repeat elements via analysis of long sequence information is essential to understand

organization and structural characteristics of the huge ginseng genome, and further to strategize ongoing ginseng genome project (Choi et al. 2012).

In this study, the major repeat elements of the ginseng genome, LTR-RTs, are presented by analysis of repeat-rich BAC sequences with long size. Whole genome shotgun (WGS) read mapping was used to estimate the abundance of the analyzed LTR-RTs in the ginseng genome. Fluorescence *in situ* hybridization (FISH) analysis was used to see the distribution of the major LTR-RTs. An evolutionary scenario of the ginseng genome is suggested through the information of polyploidy and this analysis. This is the first study to investigate major components of the ginseng genome and will assist to study underdeveloped genomes of Araliaceae.

MATERIALS AND METHODS

Selection of BAC clones, sequencing and assembly

A total of 2,492 BAC-end sequences (BESs) of the BAC library of *P. ginseng* var. Chunpoong (Hong et al. 2004) (accession numbers BZ956677.1 - BZ959168) and the sequence of *Del* retrotransposon (accession number X13886.1) were retrieved from NCBI GenBank. The BESs showing high redundancy were selected by all-against-all BLASTN search (expectation value < 10^{-20} and more than 10 hits) (Altschul et al. 1997). An internal coding region of the *Del* retrotransposon was used as a query in TBLASTX search against the redundant BESs. BAC clones that have significant TBLASTX hits (expectation value < 10^{-30} and identity > 70%) in their BESs were selected. The ginseng BAC clones (Hong et al. 2004) was kindly provided by Yong Pyo Lim (Chungnam National University, Daejeon, Korea). Among the selected BAC clones in the BLAST search, three BAC clones with large insert size of more than 100 kb were finally chosen.

Each BAC clone was sequenced on ABI3730xl (Applied Biosystems) and GS-FLX (Roche) platforms in NICEM, Seoul National University, Korea. In GS-FLX sequencing, each BAC clone was sequenced in a single lane of a pico-titer plate which was divided into 16 lanes. The Phred/Phrap/Consed software package (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998) was used for sequence assembly and manual editing. Gaps in the assembled sequences were manually filled by PCR-based sequencing.

Sequence analysis and annotation

Conserved domains in the BAC sequences were detected by NCBI Conserved Domain Search (CD-Search) (Marchler-Bauer et al. 2009) with default

parameters. Detailed positions of LTR-RTs and their derivatives in the BAC sequences were detected by CENSOR (http://www.girinst.org/censor/) (Kohany et al. 2006) and following manual inspection based on dot-plot analysis using PipMaker (Schwartz et al. 2000) and BLAST2 search. Tandem repeats were found using Tandem Repeat Finder (Benson 1999) with default parameters. To predict genic regions, FGENESH (Salamov and Solovyev 2000) program was preliminary used with Arabidopsis model and default parameters. The gene structure was finally identified on the basis of transcriptome data (unpublished) and BLASTP search against the NCBI non-redundant protein sequence database (NR). Miniature inverted-repeat transposable elements (MITEs) were detected based on the result of MITE-Hunter (Han and Wessler 2010) analysis using the whole-genome draft sequences. With BLASTN search and information of terminal inverted repeats and target site duplications (TSDs) from the MITE-Hunter analysis, the positions of MITEs in the BAC sequences were identified. Unknown repeat regions were determined when sequence regions showed no similarity to any known repeat element or gene, but exhibited high copy numbers in BLASTN search against the whole-genome draft sequences with an E-value threshold of 10^{-6}.

Phylogenetic analysis of LTR-RTs

Protein coding domains in the LTR-RT were extracted from the CD-search results. Families of the LTR-RTs in the BAC sequences were sorted by BLASTP search against the Core database downloaded from the GyDB. Phylogenetic analysis was conducted using reverse-transcriptase domains by MEGA version 5.05 (Tamura et al. 2011). The protein sequences were aligned using CLUSTALW (Thompson et al. 1994) with default parameters, and the phylogenetic tree was generated by the Neighbor-joining method (Saitou and

Nei 1987) under the Poisson correction model with 1,000 times of bootstrap replications.

Insertion time estimation of LTR-RTs

Insertion times of LTR-RTs were estimated in a manner similar to that described by (SanMiguel et al. 1996). Both of LTR sequences in a LTR-RT were aligned by using CLUSTALW with default parameters. The numbers of transition (Ts) and transversion (Tv) mutations were calculated by using MEGA version 5.05 (Tamura et al. 2011). A pairwise distance (K) between the LTRs was calculated using the Kimura 2-parameter model (Kimura 1980). Based on the report that the rate of nucleotide substitution in intergenic regions was almost double of the synonymous substitution rate in coding regions (Ma and Bennetzen 2004), a substitution rate of 1.22×10^{-8} was adopted. This value was 2-fold of the synonymous substitution rate calculated from the analyses of multiple genes in vascular plants (Lynch and Conery 2000), which was used in the previous EST analysis in ginseng (Choi et al. 2012).

Utilization of whole genome shotgun sequences

Whole genome shotgun (WGS) sequences have been generated from *P. ginseng* var. Chunpoong using Solexa sequencing technology on HiSeq2000 platform (Illumina, USA) in Macrogen Inc., Seoul, Korea. Thirty Gb of 500 bp paired-end data at read length of 101 bp were randomly sampled from the WGS sequence data by using an in-house Perl script. The reads were mapped to the BAC sequences using CLC Assembly Cell version 4.06beta software with the parameters of minimum 50 % alignment of read sequences and 80 % similarity. The number of read depth for each nucleotide position on the sequences was substituted by multiplying 3.12 / 30 (the estimated genome

size of ginseng / total read length used for mapping) to calculate proportion of the LTR-RTs characterized in the BAC sequences in the ginseng genome. Genome proportion of the LTR-RTs was estimated by adding up all depths of the nucleotide positions belonging to the LTR-RTs and their derivatives in each family.

Ginseng whole genome draft sequences (a database version 0.1 of *Panax ginseng* (2011/11/21) deposited in http://im-crop.snu.ac.kr) were used for additional sequence analysis. Paralogous sequences of a genic region were investigated in the whole-genome draft sequences by BLASTN search using the coding sequences. Sequences with a paralogous relationship were aligned using the BLASTZ algorithm (Schwartz et al. 2003).

Fluorescence *in situ* hybridization analysis

Primer pairs for FISH probes were designed from the internal regions of the LTR-RTs by amplification of reverse-transcriptase and adjacent domain regions. To design single or low copy probes in the BAC sequences, non-repetitive regions were found by BLASTN search against the whole-genome draft sequences. Probes and related primer sequence information are available in Table 2-1.

Somatic metaphase chromosomes were prepared from stratified seeds according to the methods of Waminal et al. (2012). The seeds were provided by Korea Ginseng Corporation (KGC) Natural Resources Research Institute (Daejeon, Korea). Pachytene chromosomes were obtained from 1.2 μm length flower buds according to the methods reported by Park et al. (2012b). The probes were labeled with either biotin 16-dUTP or digoxygenin 11-dUTP by nick translation following the manufacturer's protocol (Roche, Germany). The hybridization solution contained 50% formamide, 10% dextran sulfate, 2× SSC, 5 ng μl^{-1} salmon sperm DNA, and 25 ng μl^{-1} of each probe DNA,

adjusted with DNase- and RNase-free water (Sigma, USA, #W4502) to a total volume of 40 µl/slide. FISH experiments were done as reported by Waminal et al. (2012). Biotin-labeled probes were detected with streptavidin-Cy3 (Zymed, USA) while dig-labeled probes were detected with anti-dig-FITC (Sigma, USA). Signals were amplified once or twice using secondary and tertiary antibodies (for biotin-labeled probes, biotinylated anti-streptavidin (Vector Laboraties, USA) and streptavidin-Cy3 (Zymed, USA), respectively; for dig-labeled probes: anti-mouse-FITC (Sigma, USA) and anti-rabbit-FITC (Sigma, USA), respectively) in 1:100 or 1:500 concentration in TNB buffer [0.1 M Tris-HCl, 0.15 M NaCl, 1% (w/v) blocking reagent]. Chromosomes were counter-stained with DAPI (4',6-diamidino-2-phenylindole) (1 µg/ml) in Vectashield (Vector Laboratories, USA). Images were captured using Olympus BX51 fluorescence microscope equipped with a CCD camera (CoolSNAP™ cf) and analyzed using Genus 3.1 software (Applied Imaging, USA). Final images were enhanced using Adobe Photoshop CS6.

Table 2-1. Information of probe types and primer sequences used for FISH analysis

Probe type	Sequence ID	Amplified region		Size (bp)	Primer	
PgDel1	6J17_1	83068	85913	2846	Forward	CAAGATATTTGATGTTTTGGGATG
					Reverse	TAAAGTCGGTTCTACTCGTATGTTTG
PgDel2	8D23	154141	157308	3168	Forward	GTTAACTTACTGACTGGAGCAATGAC
					Reverse	AAATGTCTTCCTTAATCTCTGTGACTG
PgTat1	8D23	57177	60449	3273	Forward	TTAAAGGTTTTAGAAGAAGCCTATCTACC
					Reverse	ATTACCAACTCTTAATCTCCAAAGACAGT
PgTork	6J17_1	11463	13960	2498	Forward	CAAAATGACTAAATCTCCTTTTAGTGG
					Reverse	TGGTATCTCATACCCATCTTCTCTAAG
Non-repetitive	8D23	132089	135684	3596	Forward	ACTTCTTGAGAAATCTAAATGAAACACA
					Reverse	ATCATGTAACAATTTCAAGCAATAAAAC
Gene region 1	5J07	60597	64339	3743	Forward	TATCTCTAATTTTGGACCCTAAGGAATA
					Reverse	CAGATACAATAATTAATGGGAGTCGTTA
Gene region 2	5J07	65192	68411	3220	Forward	GCCAACTATATATGGAGATAAAACTGTTC
					Reverse	GTTTACTACCCATTTACCCTTAATAATCC

RESULTS

Selection and sequencing of three repeat-rich BAC clones

I tried to select BAC clones harboring highly redundant repeats in the ginseng genome based on the BES information. A total of 672 redundant BESs were screened by all-against all BLASTN search. Based on the previous reports about analyses of repetitive DNA sequences (Ho and Leung 2002) and the BESs (Hong et al. 2004), the sequence of *Del* retrotransposon (accession number X13886.1) characterized from *Lilium henryi* (Smyth et al. 1989) was retrieved from NCBI GenBank. At the given threshold for filtering of TBLASTX hits, 40 BAC clones were selected for sequencing candidates. Finally, two BAC clones, PgH008D23 (8D23) and PgH006J17 (6J17) were selected by the BES similarity to *Del*. And the other, PgH005J07 (5J07) was selected by the high redundancy in the BAC-end sequence data.

Each BAC clone was individually sequenced using both Sanger and pyrosequencing technology. (Table 2-2). Complete assembled sequences of 167 kb and 107 kb were obtained for two BACs, 8D23 and 5J07, respectively (Table 2-2). Two contigs resulted in the assembly for 6J17, but the longer was 98,975 bp which was enough to do further analysis (Table 2-2).

Table 2-2. Summary of sequence generation and assembly of three repeat-rich ginseng BAC clones

BAC name	ABI3730XL		GS-FLX		Contig	
	Number of reads	Total length (bp)	Number of reads	Total length (bp)	Number of contigs	Total length (bp)
8D23	942	704,044	12,917	3,119,868	1	167,296
6J17	564	449,990	13,531	3,241,934	2	106,850 (98,975 + 7875)
5J07	383	246,319	12,466	2,948,207	1	107,467

Sequence annotation of repeat-rich BAC sequences

Careful sequence analysis revealed that more than 90 % of the three BAC clones consisted of repetitive elements with complex insertion patterns (Fig. 2-1). The two BACs, 8D23 and 6J17 showed that repetitive elements spread more than 95% of the sequences (Fig. 2-1a and b). Nested insertion patterns of the repetitive elements were common throughout all the BAC sequences but more complex patterns were observed in the two BACs, which have nested insertion of maximum eight to nine elements (Fig. 2-1a and b). On the other hand, 507 exhibited non-repetitive regions in the middle of the sequence spanning 17.8 kb (Fig. 2-1c). A gene structure was predicted in the non-repetitive region that contained four exons. The first intron was markedly long, which was 4779 bp. The coding region showed 98% query coverage and 78% sequence similarity with haloacid dehalogenase-like hydrolase domain-containing protein of *Arabidopsis thaliana* (NCBI Reference Sequence: NP_179027.2) in BLASTX search against NCBI non-redundant database (Fig. 1).

LTR-RTs and their derivatives were most abundant that their proportion was 80 % in the total BAC sequences (Fig. 2-1). The BES regions, a right end of 8D23 and a left end 6J17, respectively, were identified as internal sequences of *Del*-like retrotransposons, which corresponded to the BES analysis (Fig. 2-1a and b). The redundant BES of 5J07 was identified as a border region of the 3' LTR of the *Del*-like retrotransposon (right end of Fig. 2-1c). Two degenerated DNA transposons and five MITE-like elements were found (Fig. 2-1a and c). A distinctive 167-bp tandem repeat and several

unknown repetitive sequences were characterized (Fig. 2-1). Position information including ten LTR-RTs, ten solo LTRs, five MITEs and the other elements characterized with full-structures and TSDs in the BAC sequences is shown in Table 2-3.

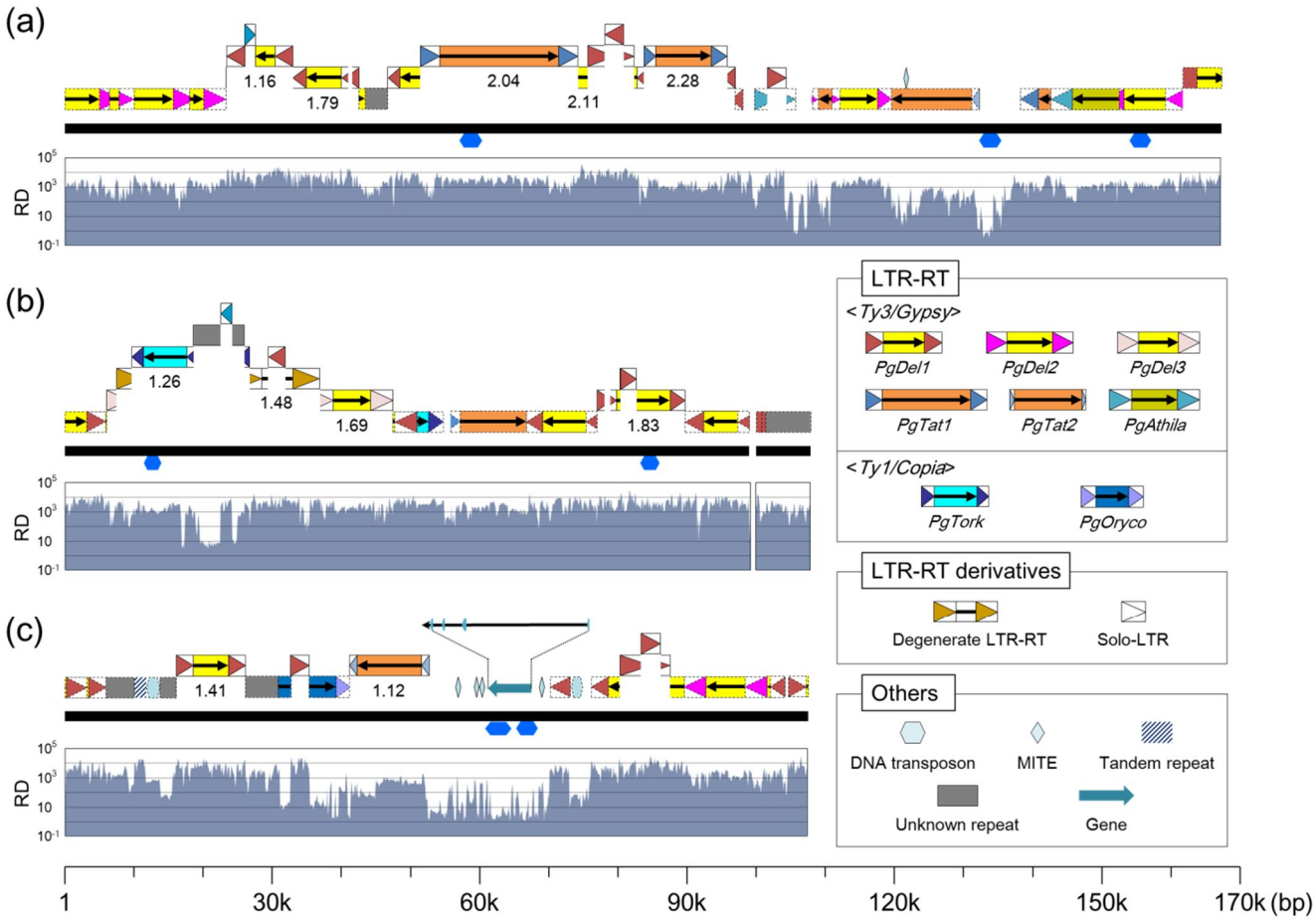
(a)
1.16
1.79
2.04
2.11
2.28
RD
(b)
1.26
1.48
1.69
1.83
RD
(c)
1.41
1.12
RD
LTR-RT
<Ty3/Gypsy>
PgDel1
PgDel2
PgDel3
PgTat1
PgTat2
PgAthila
<Ty1/Copia>
PgTork
PgOryco
LTR-RT derivatives
Degenerate LTR-RT
Solo-LTR
Others
DNA transposon
MITE
Tandem repeat
Unknown repeat
Gene
1
30k
60k
90k
120k
150k
170k (bp)

Figure 2-1. Sequence analysis of the three repeat-rich BAC clones. (a) 8D23, (b) 6J17, and (c) 5J07. Horizontal black bars represent the BAC sequences. Physical maps of the BAC sequences are shown on the black bars. The types of elements depicted in the physical maps are sorted in the right box. The triangle of LTR region is colored by the sequence similarity. Insertion time of each LTR-RT calculated from its LTR pair is shown under the elements. The repeat elements having full-structure with TSDs are depicted with straight lines. The other repeat elements having partial or truncated structure and putative repeat regions are depicted with dotted lines. The exon structure of the gene is depicted with green pentagons on the black arrows in the enlarged section of (c). Graphs under the black bars represent the depth distribution of whole genome shotgun read mapping. The numbers of read depth (RD) for each nucleotide position are plotted after substitution (see details in Materials and Methods). The vertical-axis representing RD is converted to a logarithmic scale. The regions from which FISH probes were designed are represented as blue hexagons underneath the black bars.

Table 2-3. Detailed position information of repetitive elements in the BAC sequences.

Type	BAC sequence	TSD-Left	Position on the BAC sequence	TSD-Right	Length
PgDel1	8D23	AAGTA	23334-23362, 32933-40836, 41409-42380	AAGTA	8905
PgDel1	8D23	ACATCA	23363-25995, 27556-32926	ACATCA	8004
PgDel1	8D23	GTCAC	46781-51181, 74068-75415, 82113-83407, 95702-96371	GTAAC	7714
PgDel1	6J17_1	ACTTA	76821-78052, 78592-80102, 82966-89722	ACTTA	9500
PgDel1	5J07	AAATC	15991-26110	AAATC	10120
PgDel3	6J17_1	TGGGA	6065-7415, 36961-47418	TGGGA	11809
PgTat1	8D23	GCGAC	51182-74062	GCGAC	22881
PgTat1	8D23	AACTT	83408-93507, 94416-95696	AACTT	11381
PgTat2	5J07	AGTAG	41381-52345	AGTAG	10965
PgTork	J17_1	GAAAC	9572-18462, 25797-26612	GAAAC	9707
Degenerated[a]	6J17_1	TATTT	7416-9571, 26618-29285, 31892-36955	TATTT	9888
Solo-LTR	8D23	TGAGG	25996-27550	TGAGG	1555
Solo-LTR	8D23	CATAC	75416-77855, 80606-82107	CATGC	3942
Solo-LTR	8D23	ATGAT	77856-80600	ATGAT	2745
Solo-LTR	8D23	TTAGT	101353-104081	TTAGT	2729
Solo-LTR	6J17_1	ATAGC	22433-24030	ATAGC	1598
Solo-LTR	6J17_1	CCAAA	29286-31886	CCAAA	2601
Solo-LTR	6J17_1	GGAAA	80103-82960	GGAAA	2858
Solo-LTR	5J07	GTGTG	32639-35379	GTGTG	2741
Solo-LTR	5J07	ACGTG	80102-83215, 86143-87603	ACGTG	4575
Solo-LTR	5J07	ATCAT	83216-86137	ATCAT	2922
MITE	8D23	GTCACTGT	121277-122470	GTCACTGT	1194
MITE	5J07	CCTGACAT	68525-69228	CCTGACAC	704
MITE	5J07	AATGTTTTA	56701-56851	AATGTTTTA	151
MITE	5J07	TAA	59196-59616	TAA	421
MITE	5J07	TA	60054-60499	TA	446
Tandem repeat	5J07	-	10068-11670	-	1603
Unknown repeat	6J17_1	CACCA	18463-22432, 24036-25791	CACCA	5726

[a]Degenerated LTR-RT

Characterization and classification of LTR-RTs

The major components of the BAC sequences were LTR-RTs and their derivatives (Fig. 2-1). A total of 34 LTR-RTs were found, which includes 10 of full and 24 partial or truncated structures (Fig. 2-1 and Table 2-3). Among the 34 LTR-RTs, 31 were *Ty3/gypsy*-like elements, and the other three were *Ty1/copia*-like elements. The 31 *Ty3/gypsy*-like elements were divided into three families that 23 belonged to the *Del* family (named *PgDel*, Fig. 2-1 and 2) another seven belonged to *Tat* (named *PgTat*, Fig. 2-1 and 2), and the other one belonged to *Athila* (named *PgAthila*, Fig. 2-1 and 2) by the classification system of GyDB. Two additional domains were detected in the *PgDel* elements, which were the zinc-knuckle (accession number of cl15298 in NCBI CDD) and chromodomain (accession number of cl15261 in NCBI CDD). Both the two domains were also observed in the *Del* element, while the locations of zinc-knuckle domains were different between *PgDel* and *Del*. The zinc-knuckle domain was placed between capsid protein and aspartic protease domains in *PgDel*; between RNaseH and integrase domains in *Del*. In the case of three *Ty1/copia*-like elements, two belonged to *Oryco* (named *PgOryco*, Fig. 2-1 and 2), and the other belonged to *Tork* (named *PgTork*, Fig. 2-1 and 2). All of the *Ty1/copia*-like elements identified here were lacking in the domain of aspartic protease (Fig. 2-2).

Phylogenetic analysis among the *Ty3/gypsy*-like elements using reverse-transcriptase domain presented that clustering patterns of the elements were in accord with the type of LTR sequences (Fig. 2-1, 2 and 3). The *PgDel* family which was most abundant in the analyzed sequences showed three types of

subfamilies (Fig. 2-1, 2 and 3). The three subfamilies presented sequence similarity in their internal regions each other, which was 60% or fewer in the nucleotide level (Fig. 2-2). The *PgTat* family was the second abundant and two subfamilies were identified that no significant nucleotide sequence similarity was found in either LTR regions or internal regions (Fig. 2-1, 2 and 3).

Sequence divergence between each LTR was calculated for the 11 elements which contained full LTRs with flanking TSDs were identified (Table 2-4). The ratios of transition to transversion (Ts/Tv) of the 11 elements were ranged from 1.4 to 4.3 and the average value was 2.2. The pairwise distances were ranged from 0.027 to 0.056, which were converted into 1.1 – 2.3 million years ago (MYA) (Table 2-4).

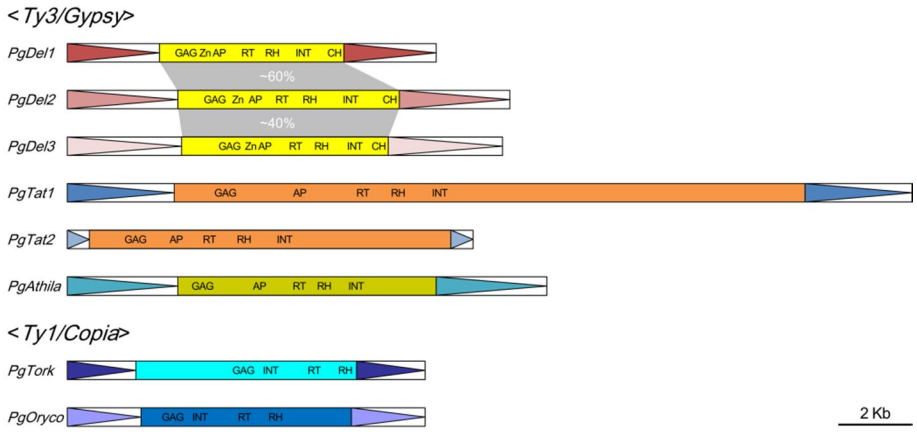

Figure 2-2. Schematic representation of ginseng LTR-RTs identified from the BAC sequences. Boxed triangles indicate LTR regions. Nucleotide-level sequence similarity observed between *PgDel* subfamilies is denoted in grey boxes. Detected domains, AP (aspartic protease), CH (chromodomain), GAG (capsid protein), INT (integrase), RH (RNase H), RT (reverse-transcriptase), and Zn (zinc knuckle) are denoted according to their locations in the internal regions of elements. The structures of *PgAthila* and *PgOryco* were deduced from the partial copies in the BAC sequences.

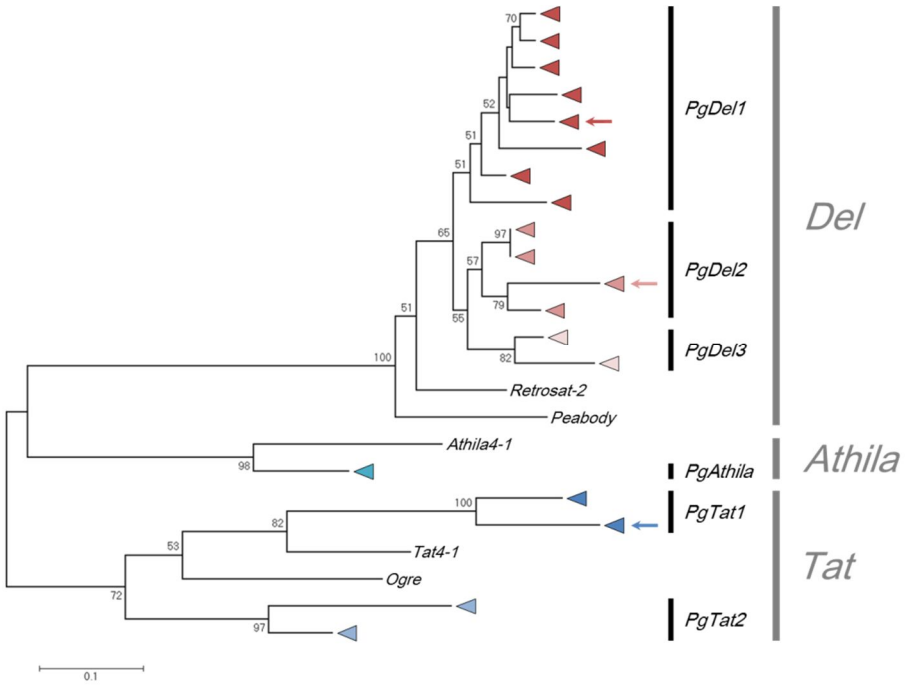

Figure 2-3. Phylogeny of *Ty3/Gypsy*-like retrotransposons analyzed in the ginseng BAC sequences. The Neighbor-joining tree was constructed using twenty reverse-transcriptase domains from the BAC sequences and five RT domains from GyDB showing the highest BLASTP hits against the twenty domains. Each element is depicted by a color triangle according to its LTR type. The elements used as FISH probes are denoted by the arrows. Bootstrap values were calculated by 1,000 replications and the values greater than 50% are shown next to the branches.

Table 2-4. Pairwise distances and insertion time estimation of 11 full-structured LTR-RTs.

Type	Sequence ID	LTR Position (bp)		LTR length (bp)	Ts[a]	Tv[b]	Ts/Tv	K[c]	Insertion time[d] (MYA)
PgDel1	8D23	L	23334-23362, 32933-34935	2032	49	36	1.36	0.0436	1.79
		R	39730-40836, 41409-42380	2079					
PgDel1	8D23	L	23363-25927	2565	48	21	2.29	0.0284	1.16
		R	30400-32926	2527					
PgDel1	8D23	L	46781-48440	1660	59	22	2.68	0.0516	2.11
		R	82421-83407, 95702-96371	1657					
PgDel1	6J17_1	L	76821-78052, 78592-79695	2336	66	35	1.89	0.0447	1.83
		R	87389-89722	2334					
PgDel1	5J07	L	15991-18624	2634	61	27	2.26	0.0344	1.41
		R	23447-26110	2664					
PgDel3	6J17_1	L	6065-7415, 36961-38719	3110	80	44	1.82	0.0411	1.69
		R	44312-47418	3107					
PgTat1	8D23	L	51182-54016	2835	100	35	2.86	0.0498	2.04
		R	71161-74062	2902					
PgTat1	8D23	L	83408-85101	1694	57	31	1.84	0.0557	2.28
		R	93136-93507, 94416-95696	1653					
PgTat2	5J07	L	41381-41980	600	13	3	4.33	0.0273	1.12
		R	51746-52345	600					
PgTork	6J17_1	L	9572-11432	1861	41	15	2.73	0.0308	1.26
		R	17418-18462, 25797-26612	1861					
Degenerated[e]	6J17_1	L	7416-9571, 26618-28530	4069	101	41	2.46	0.0362	1.48
		R	32905-36955	4051					

[a] Number of transition mutations

[b] Number of transverion mutations

[c] Kimura distance

[d] Insertion times were estimated by adopting substitution rate of 1.22×10^{-8} (see details in Materials and method section)

[e] Degenerated LTR-RT observed in 6J17

LTR-RT derivatives

The element of which LTR pair is colored in goldenrod spanning 7.4 to 37.7 kb in Fig. 2-1b showed a degenerated structure of LTR-RTs, which was an intermediate structure between LARDs (large retrotransposon derivatives) (Kalendar et al. 2004) and TRIMs (terminal-repeat retrotransposons in miniature) (Witte et al. 2001). Although it had long LTRs more than 4 kb with TSDs, the internal region was shrunken and degenerated that no domain was detected, and miniaturized as 1.8 kb in length (Fig. 2-1b and Table 2-3).

A total of ten solo-LTR structures were found with TSDs. Eight of them were derived from *PgDel1* elements (Fig. 2-1 and Table 2-3) and the other two were derived from a different type that were not detected in the analyzed BAC sequences. All of the ten solo-LTRs were nested in other elements that six were found in internal regions of the LTR-RTs, another three were on LTR regions, and the other one was on the unknown repeat (Fig. 2-1 and Table 2-3).

Other repetitive elements

Two of transposase motifs of DNA transposons were detected on 5J07, but either terminal inverted repeats or TSDs were not found possibly due to the degeneration (Fig. 2-1). A total of five MITE elements were detected in the sequences (Fig. 2-1 and Table 2-3). Four of the five were found in the non-repeat region, and the other was in the internal region of LTR-RT (Fig. 2-1). No significant sequence similarity showed among the MITE-like elements found in the BAC sequences. Some of regions in the sequences were regarded as unknown repeats based on the BLASTN search against the whole genome draft sequences (Fig. 2-1 and Table 2-3). Only one unknown element was found in the left LTR region of *PgTork* with TSDs (Fig. 2-1 and Table 2-3).

Genome proportion of LTR-RTs in ginseng

Total 41.38% of 30 Gb WGS reads were mapped to the BAC sequences (Table 2-5). The total length of mapped nucleotides to the BAC sequences was 11.49 Gb, which was 38.29% of the WGS reads used. The average coverage of mapping against the BAC sequences was 30107.58 (Table 2-5).

Mapping frequency was significantly different between repetitive and non-repetitive regions (Fig. 2-1). *PgDel1*, which was the most predominant repeat in the BAC sequences, exhibited the highest mapping frequency among the all families and subfamilies in not only internal regions but also LTR regions including the eight solo-LTRs that the substituted depth of more than 30,000 was observed (Fig. 2-1). In the LTR regions of LTR-RTs, the both terminals of LTR showed relatively higher depth than the middle part of LTR (Fig. 2-1). In the internal regions of LTR-RTs, no conspicuous depth difference was observed, except for the elements of *PgTat1* that exhibited higher depth in 5' coding regions than in 3' non-coding regions (Fig. 2-1). Contrary to the repetitive regions, the gene region of 5J07 and some areas between LTR-RTs in 8D23 exhibited low mapping frequency. Low depth was also observed in the junctions and adjacent regions of repetitive elements, which was sharper at the place where different types of elements met each other (Fig. 2-1).

The genome proportion of the LTR-RTs found in the BAC sequences was calculated as 1,073 Mb that was 34.4% of the ginseng genome size (Table 2-6). Position information of the LTR-RT families in the BAC sequences used for calculation of genome proportion is available in Table 2-7. Among *Ty3/Gypsy*, *PgDel* was the major family, revealing the genome proportion of 899 Mb (Table 2-6 and Table 2-7). *PgDel1* was the most abundant subfamily in *PgDel*, which took up 86.7% of the total proportion of *PgDel* elements. Although *PgDel2* and *PgDel3* were less frequent than *PgDel1*, they also

showed several tens of Mb (Table 2-6 and Table 2-7). In the *PgTat* family, *PgTat1* was estimated to be 13.2 times larger proportion than that of *PgTat2* (Table 2-6 and Table 2-7). Although the *PgAthila* element was found as a partial single copy in the BAC sequences (Fig. 2-1), its proportion was measured at 16.1 Mb in the ginseng genome (Table 2-6 and Table 2-7). The *Ty1/copia* group showed much smaller proportion than that of *Ty3/Gypsy* (Table 2-6 and Table 2-7). The *PgTork* family took up almost all part of the *Ty1/Copia* group (Table 2-6 and Table 2-7).

Table 2-5. Summary result of whole genome shotgun read mapping to the BAC sequences.

Category	Count	%
Total number of reads used	297,029,704	100
Total length of reads used (bp)	30,000,000,104	100
Number of reads mapped	122,906,713	41.38
Number of nucleotides mapped[a]	11,488,293,860	38.29
Average coverage[b]	30107.58	-

[a] Total length of mapped nucleotides in the mapped reads

[b] Average coverage against the total BAC sequences

Table 2-6. Proportion of LTR-RTs in the ginseng genome estimated from WGS read mapping.

Type	Length[a] (kb)	Genome proportion[b] (Mb)	%[c]
Ty3/Gypsy		1,037.51	33.25
PgDel		899.22	28.82
PgDel1	128,837	779.73	24.99
PgDel2	44,950	72.30	2.32
PgDel3	23,461	47.19	1.51
PgTat		122.15	3.92
PgTat1	50,641	113.56	3.64
PgTat2	22,272	8.59	0.28
PgAthila	13,014	16.13	0.52
Ty1/copia		35.84	1.15
PgTork	13,567	34.29	1.10
PgOryco	7,772	1.55	0.05
Total		1,073.35	34.40

[a]Length of the regions belonging to the LTR-RT family in the BAC sequences

[b]Estimated proportions in the ginseng genome from the mapping depth of WGS reads (See details in MATERIALS AND METHODS)

[c]Percentage based on the ginseng genome size of 3.12 Gb

Table 2-7. Nucleotide positions of the regions belonging to the LTR-RT families in the BAC sequences used for calculation of genome proportion in ginseng.

Type	Regions in the BAC sequences		
	5J07	6J17_1	8D23
Ty3/Gypsy			
PgDel			
PgDel1	1-6033 15991-26110 32639-35379 70144-72956 76079-89417 101933-107467	1-6064 29286-31886 47419-50836 66621-98975	23334-42380 46781-51181 74063-83407 95697-97841 101353-104081 161146-167296
PgDel2	89418-101932	-	1-23333 152044-152597
PgDel3	-	6065-7415 36961-47418	107892-119543
PgTat			
PgTat1	41381-52345	-	119544-121276 122471-132044
PgTat2	-	55704-66620	51182-74062 83408-95696 137597-142150
PgAthila	-	-	99505-101352 104082-105354 142151-152043
Ty1/Copia			
PgTork	-	9572-18462 25792-26612 50837-54691	-
PgOryco	30863-32638 35385-41380	-	-

Distribution of LTR-RTs in ginseng genome

FISH analysis using pachytene chromosomes enables to distinguish between heterochromatic and euchromatic regions by differential staining on the chromosomes (Stack 1984). The ginseng pachytene chromosomes showed that euchromatic regions (blue arrowheads) intermingled with heterochromatic regions (red arrowheads) (Fig. 2-4a). The FISH result of *PgDel1*, the biggest subfamily in *Ty3/Gypsy*, showed higher signal density than that of *PgTork*, the major family in *Ty1/Copia* (Fig. 2-4b and c). Dense FISH signals of the *PgDel1* probe were observed along whole pachytene chromosomes (Fig. 2-4b). The FISH signals of *PgTork* were less dense than that of *PgDel1*, but the signals were distributed in heterochromatic regions as well as euchromatic regions (Fig. 2-4c).

FISH analyses on somatic metaphase chromosomes showed differential distribution patterns for each family and subfamily (Fig. 2-5 and 6). In *Ty3/Gypsy*, the most abundant subfamily *PgDel1* exhibited dense FISH signals throughout the whole chromosomes, except 45s rDNA satellite loci (Fig. 2-5a). *PgTat1* showed dispersed FISH signals in whole chromosomes, except subtelomeric and telomeric areas (Fig. 2-5b). The FISH signals of *PgTork* belonging to *Ty1/Copia* were observed on interstitial regions (Fig. 2-5c).

Peculiar signal patterns were detected in the FISH analysis for *PgDel2*. The FISH signals of *PgDel2* showed strong intensities in pericentromeric regions of 24 chromosomes which represent a half of the chromosome complement (Fig. 2-6). This phenomenon was more distinct when the high concentration of antibody for the detection of labeled probes was applied (Fig. 2-6b and d).

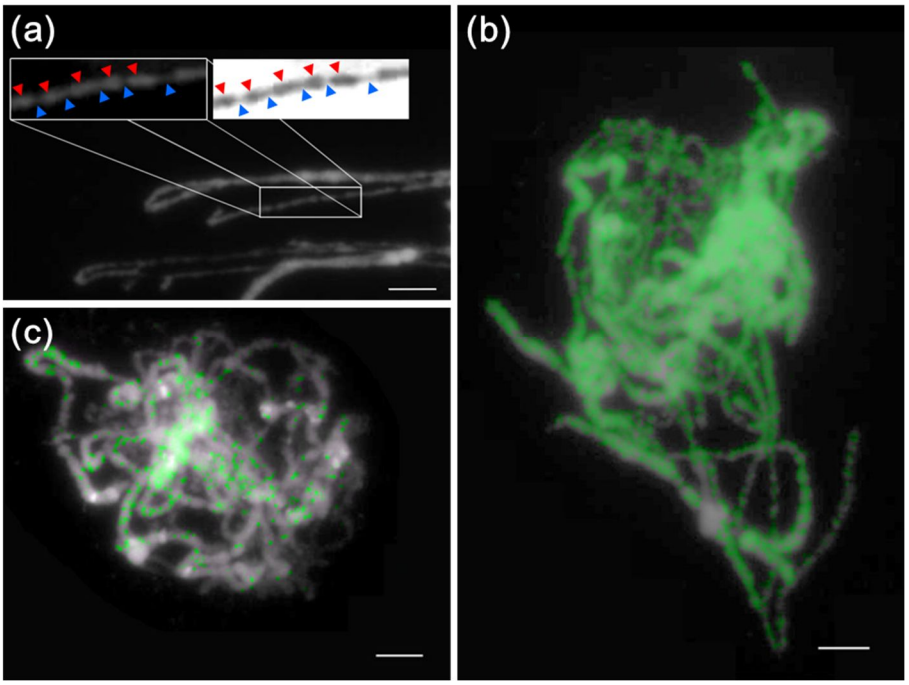

Figure 2-4. FISH analysis using ginseng pachytene chromosomes. (a) Partial section of the ginseng pachytene chromosome. Heterochromatic and euchromatic regions are indicated in red and blue arrowheads, respectively, in the enlarged portion. (b) FISH result of *PgDel1*. (c) FISH result of *PgTork*. Bar, 5μm.

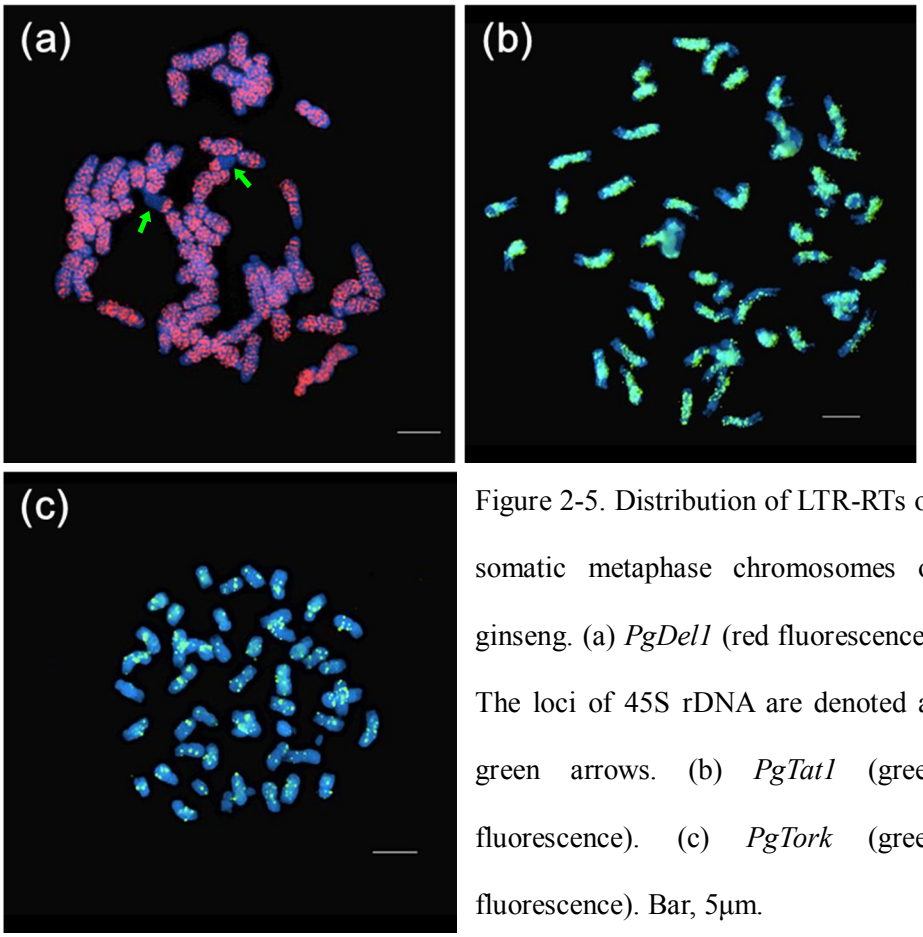

Figure 2-5. Distribution of LTR-RTs on somatic metaphase chromosomes of ginseng. (a) *PgDel1* (red fluorescence). The loci of 45S rDNA are denoted as green arrows. (b) *PgTat1* (green fluorescence). (c) *PgTork* (green fluorescence). Bar, 5μm.

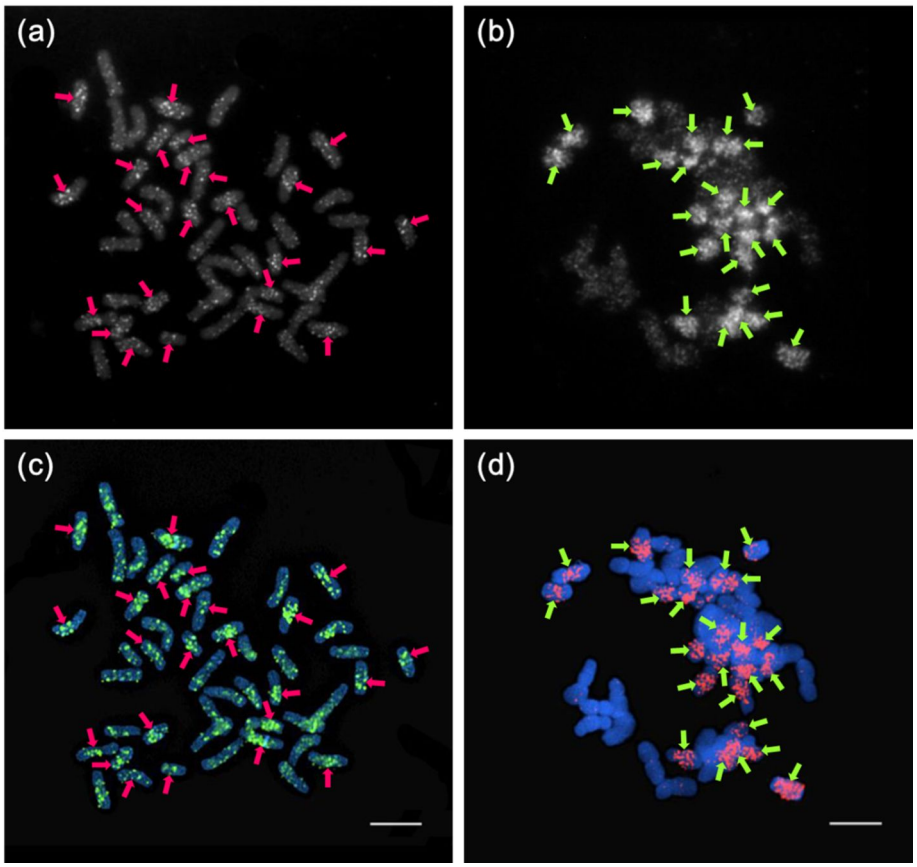

Figure 2-6. Distribution of *PgDel2* on somatic metaphase chromosomes of ginseng. Chromosomes showing intense signals are denoted as arrows. (a) and (b) are raw images. (c) and (d) are pseudo-color images. The FISH signals were detected using 1:500 AD-FITC (a, c) and 1:100 concentration of AD-rhodamine (b, d). Bar, 5μm.

The unique gene structure in 5J07

Using the coding region of the gene structure characterized in 5J07 (Fig. 2-1c), four scaffold sequences containing paralogous genes were identified by BLASTN search against whole genome draft sequences, whereas the counterpart gene was found only one copy in the Arabidopsis genome (Lamesch et al. 2012). Excluding one sequence corresponding to the BAC sequence, the other three scaffold sequences were aligned with the BAC sequence (Fig. 2-7). Although Scaffold1 had short length of 12,807 bp and a 1,098-bp gap, the entire sequence was aligned to the BAC sequence (Fig. 2-7). The exons were conserved in length and showed 99% similarity and 0.008 of the Ks value between the two sequences. Both Scaffold2 and Scaffold3 showed significant sequence matches only to the gene region of the BAC sequence (Fig. 2-7). The first introns of the two scaffolds were obviously shorter than that of the BAC sequences, which was approximately 1 kb. The exons of Scaffold2 and Scaffold3 were also conserved in length and showed 97% similarity and 0.068 of the Ks value between them and average 90% similarity and 0.298 of the Ks value against the BAC sequence and Scaffold1. On the other hand, Scaffold2 and Scaffold3 were aligned throughout the whole sequences (Fig. 2-7).

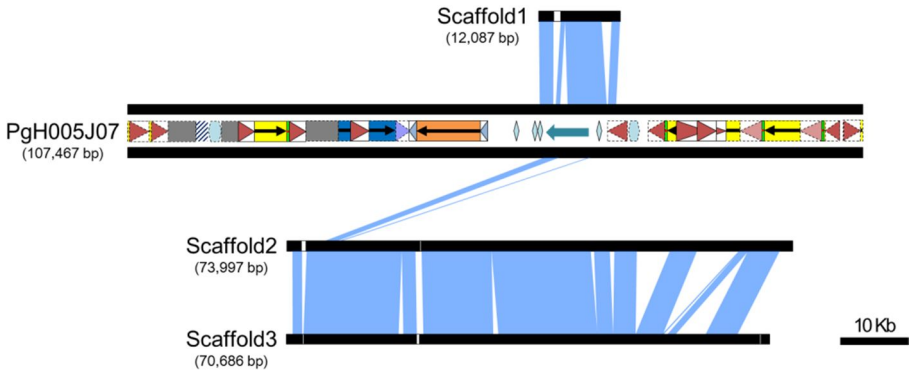

Figure 2-7. Alignment of four sequences containing paralogous gene copies in the ginseng genome using BLASTZ algorithm. Homologous regions are connected with blue boxes. Gaps in the whole genome draft sequences are depicted with white boxes.

Duplicated copies of non-repetitive regions in ginseng genome

Probes were designed from the non-repetitive regions which showed low mapping frequency to investigate the number of loci in the ginseng genome. Two adjacent regions covering the gene region were selected from 5J07, and one region was selected from 8D23 (Fig. 2-1a, c, and Table 2-1). Pseudo-colored composite image merged from the signals of the two probes from 5J07 showed four clear signals (Fig. 2-8). The FISH result with the probes from 8D23 was also showed four signals on interphase nucleus (Fig. 2-9).

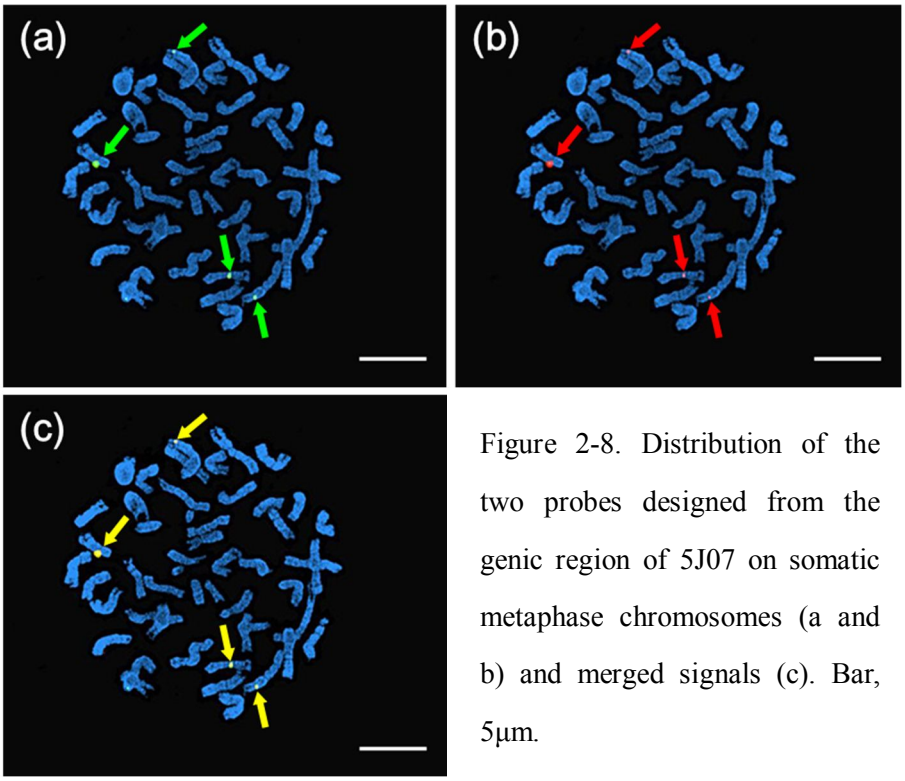

Figure 2-8. Distribution of the two probes designed from the genic region of 5J07 on somatic metaphase chromosomes (a and b) and merged signals (c). Bar, 5μm.

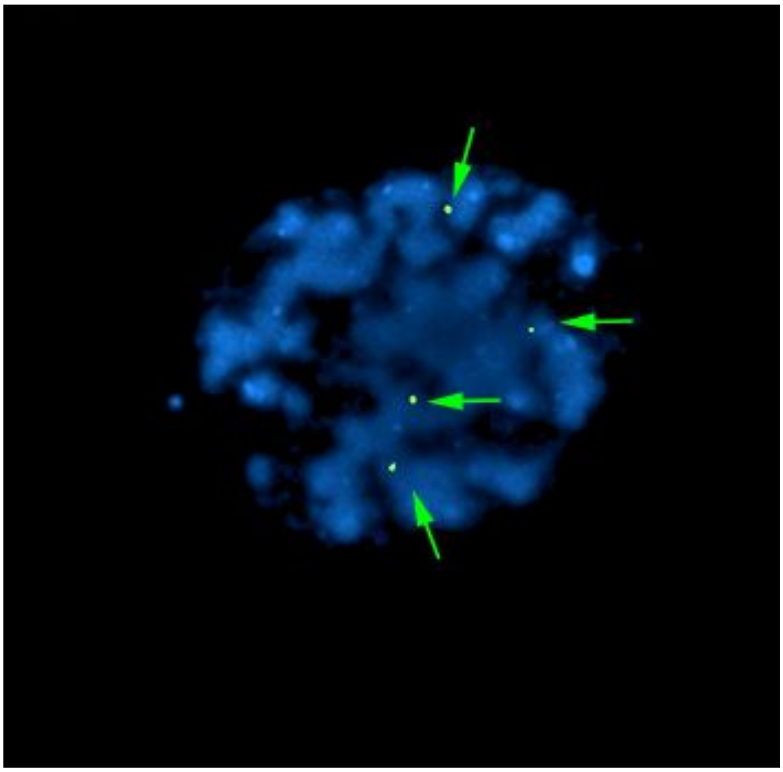

Figure 2-9. Distribution of the probe designed from the non-repetitive region of 8D23 on interphase nucleus.

DISCUSSION

Here, I first characterized the major components of the ginseng genome, LTR-RTs, by sequence analysis of the repeat-rich BAC clones. High abundance of the LTR-RTs was revealed by WGS read mapping and FISH. Among transposable elements, LTR-RTs have been known as the main factor in size variation of plant genomes (Bennetzen et al. 2005; Kumar and Bennetzen 1999; Vitte and Panaud 2005). Many studies in the plant species which have large genomes, such as maize (SanMiguel et al. 1996; Schnable et al. 2009), barley (Suoniemi et al. 1996; Vicient et al. 1999), wheat (Charles et al. 2008; Salina et al. 2011), and pepper (Park et al. 2011; Park et al. 2012b), have been reported that their genome sizes were expanded by accumulation of LTR-RTs. This results suggest that LTR-RTs have played an important role in the formation and evolution of the large ginseng genome.

LTR-RTs and genome size expansion in ginseng

The sequence analysis revealed complex insertion patterns of repetitive elements, mainly caused by LTR-RTs and their derivatives (Fig. 2-1). Nested insertion of LTR-RTs has been discussed in intergenic regions of the plant species with large genomes (Ma and Bennetzen 2004; Park et al. 2012b; SanMiguel et al. 1996; Wicker et al. 2001; Wicker et al. 2005). In this study, the two BACs containing no genic region, 8D23 and 6J17, could represent the context of intergenic or heterochromatic regions in the ginseng genome. Complex and consecutive nested insertion patterns, estimated high copy numbers and dense FISH signals of the LTR-RTs indicate that they are the main components of heterochromatin regions in the ginseng genome. Above all, the *PgDel1* elements are considered the most major components, as it is indicated from the estimated genome proportion of 25% and intensely

dispersed FISH signals along the entire chromosomes.

Although the genome proportion of the LTR-RT families based on the WGS mapping frequency was calculated more than 34% (Table 2-6), this could be underestimated due to not only their sequence divergence, but also disruption of elements by nested insertion or recombination in the ginseng genome. Besides, it is plausible that other major repeat families should be present in the ginseng genome because of its large size.

PgDel elements and remodeling of euchromatic regions in ginseng genome

The 5J07 sequence containing the only gene structure also showed nested insertion patterns but less complex structure than the two BACs (Fig. 2-1c). This may reflect a case of nested insertion in adjacent genic regions of the ginseng genome. Accumulation of LTR-RTs around genic regions has been continuously reported in maize (Kronmiller and Wise 2009; SanMiguel et al. 1998), wheat (Salina et al. 2011; Wicker et al. 2001), barley (Wicker et al. 2005) and pepper (Park et al. 2011; Park et al. 2012b). Among them, the studies in pepper proposed that accumulation of the *Ty3/Gypsy* elements in *Tat* and *Athila*-like subgroups has resulted in expansion of euchromatic regions by random insertion (Park et al. 2011), while accumulation of the elements in a *Del*-like subgroup has resulted in expansion of constitutive heterochromatic regions such as pericentromeric regions (Park et al. 2012b). In contrast, the sequence analysis revealed the existence of the *Del*-like elements and their solo-LTRs in the areas adjacent the genic region (Fig. 2-1c). Solo-LTRs have been postulated to be byproducts of LTR-RTs resulting from the internal deletion caused by unequal homologous recombination that they are vestiges of LTR-RT insertion (Shirasu et al. 2000; Vitte and Panaud 2003). Chromodomain motifs were found in most of the *PgDel* elements (Fig. 2-1)

that they are known to have a function of targeting integration of retrotransposons in heterochromatic regions by recognizing histone H3 K9 methylation (Gao et al. 2008). Extensive deoxycytidine-methylation, an epigenetic DNA modification, elevates C to T transition rate that leads to higher Ts/Tv between two LTRs of an LTR-RT (SanMiguel et al. 1998; Vitte and Bennetzen 2006). In this study, the average Ts/Tv of 2.2 was calculated from the LTR pairs of the 11 elements, and the ratio of the two elements in 5J07 was respectively 2.3 and 4.3 (Table 2-4). This high ratio could indicate active methylation of the repetitive elements in the ginseng genome. In addition, the pachytene FISH analysis showed intermingling distribution of heterochromatic and euchromatic regions and non-preferential distribution of *PgDel1* (Fig. 2-3). Taken together, accumulation of *Del*-like elements in the ginseng genome may have contributed to not only expansion of constitutive heterochromatic regions, but also conversion of euchromatic regions to heterochromatic regions and formation of genic islands by insertion into the methylated repeat regions in the vicinity of genic regions.

PgDel2 and alloploidy-like nature of ginseng genome

Previous studies have inferred that the *Panax* species with $2n = 48$ are likely to be allopolyploids, because analyses from chloroplast intergenic regions versus internal transcribed spacer (ITS) datasets exhibited incongruent phylogenies (Yi et al. 2004; Lee and Wen 2004). The FISH result with *PgDel2* represented the denser signals on a half of the chromosome complement (Fig. 2-5), which is a first piece of evidence for alloploidy nature of the ginseng genome that *PgDel2* may have been activated in one of the ginseng ancestors in old times. Subgenome-specific or -differential proliferation of transposable elements have been reported in *Brassica* (Yang et al. 2007) cotton (Hanson et al. 2000; Zhao et al. 1998), *Oryza* (Ammiraju et al. 2008) and wheat (Sabot et

al. 2006; Salina et al. 2011). Further inspection of ginseng and related species with *PgDel2* and/or other chromosome-specific elements will clarify the origin and authentication of subgenomes, and evolutionary relationship of *Panax* species.

Evolution of ginseng genome

The ginseng whole genome draft sequences contained four paralogous copies of the gene structure observed in 5J07 that can be divided into two pairs on the basis of the sequence alignment, exon sequence similarity, and Ks values (Fig. 2-7). The former study reported two rounds of genome duplication events coincidently observed between ginseng and American ginseng from the distribution of Ks values among paralogous genes (Choi et al. 2012). The Ks value of the coding regions from each pair is close to the recent Ks peak and the Ks values between the two pairs are close to the ancient Ks peak. These results indicate a recent paralogous relationship between each pair and an ancient paralogous relationship between the two pairs, produced by the two rounds of genome duplication events. The four FISH signals from the two probes designed in the genic region (Fig. 2-8) are regarded as the two loci of recent paralogous pair, because the sequence homology between the two pairs was limited only in the exons and very adjacent regions (Fig. 2-8).

However, associating the allopolyploidy nature and the two Ks peaks is needed to address two caveats. First, various processes are possible for allopolyploidy formation (Hegarty and Hiscock 2008). Second, while Ks peaks allow us to get evidence for the existence of genome duplication events, they do not inform about the point of the events because Ks peaks derived from homoeologous pairs indicates the divergence time of the two progenitors (Doyle and Egan 2010).

The degree of sequence divergence between an LTR pair is proportional

to the insertion time of the LTR-RT (Vitte and Bennetzen 2006). In the former study (Choi et al. 2012), the time of the recent Ks peak was estimated at ca. 1.6 to 3.3 MYA, which slightly precedes and overlaps the time of LTR-RT insertion, 1.1 to 2.3 MYA (Table 2-4). Activation of retrotransposons in all eukaryote genomes including plants are well-known to be induced by certain stresses (Grandbastien 1998; Kumar and Bennetzen 1999). The so called "genome shock" is a kind of responses to stress that hybridization and/or chromosome doubling triggers activation of repetitive elements (Jones and Pasakinskiene 2005; McClintock 1984). Although the number of the elements is not enough to assure, the insertion time of the LTR-RTs could be related to the genome shock that the ginseng genome experienced recently. However, the times converted from the Ks and K values should be accepted as relative scales because ginseng is perennial that its generation time is speculated to be more than eight to ten years (Choi et al. 2012). It should be correct that the evolutionary events revealed in the ginseng genome actually took placed much earlier than the estimated times.

The evolutionary events which have been unveiled so far in the ginseng genome are summarized in Fig. 2-10. As either of the Ks peaks cannot be presumed to be an index related to the allopolyploidy, the time calculated from the Ks peak may indicate when the two progenitors of ginseng diverged. In this case, the time of the Ks peak is prior to that of the genome duplication. Meanwhile, if the Ks peak is derived from autopolyploidy, the time of it indicates the point of genome duplication.

In this study, I first characterized the LTR-RTs which are major components in the ginseng genome. Of them, *PgDel1* have played an important role in the formation of the large ginseng genome. The differential intensity of *PgDel2* FISH signals on half the chromosome complement suggests alloploidy-like genome structure of ginseng. The insertion time represents proliferation of LTR-RTs after the recent polyploidy in ginseng.

Upcoming large-scale genome information will provide an insight into the ginseng genome which is large and complex.

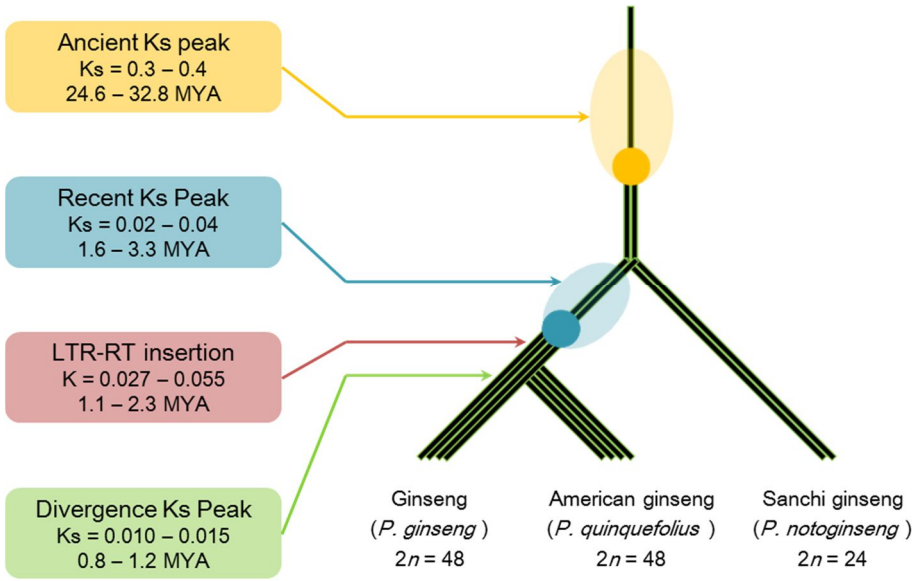

Figure 2-10. Evolutionary scenario of ginseng genome. The evolutionary events which have been revealed by Choi et al. (2012) and this study are listed in the left boxes. Deep circles represent the times of genome duplication events. Semi-translucent ovals behind the circles represent the time ranges which Ks peaks indicate.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang JH, 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

Ammiraju JSS, Lu F, Sanyal A, Yu Y, Song X, Jiang N, Pontaroli AC, Rambo T, Currie J, Collura K, Talag J, Fan CZ, Goicoechea JL, Zuccolo A, Chen J, Bennetzen JL, Chen MS, Jackson S, Wing RA (2008) Dynamic evolution of oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. Plant Cell 20:3191-3209

Bang KH, Lee JW, Kim YC, Kim DH, Lee EH, Jeung JU (2010) Construction of Genomic DNA library of Korean ginseng (*Panax ginseng* CA MEYER) and development of sequence-tagged sites. Biol Pharm Bull 33:1579-1588

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621-627

Bennetzen JL, Ma JX, Devos K (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95:127-132

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573-580

Charles M, Belcram H, Just J, Huneau C, Viollet A, Couloux A, Segurens B, Carter M, Huteau V, Coriton O, Appels R, Samain S, Chalhoub B (2008) Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. Genetics 180:1071-1086.

Chen S, Luo H, Li Y, Sun Y, Wu Q, Niu Y, Song J, Lv A, Zhu Y, Sun C, Steinmetz A, Qian Z (2011) 454 EST analysis detects genes putatively involved in ginsenoside biosynthesis in *Panax ginseng*. Plant Cell Rep 30:1593-1601

Choi HI, Kim NH, Lee J, Choi B, Kim K, Park J, Lee SC, Yang TJ (2012) Evolutionary relationship of *Panax ginseng* and *P. quinquefolius* inferred from sequencing and comparative analysis of expressed sequence tags. Genet Res Crop Evol doi:10.1007/s10722-012-9926-3

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

Doyle JJ, Egan AN (2010) Dating the origins of polyploidy events. New Phytol

186:73-85

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186-194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175-185

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat rev Genet 3:329-341

Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. Trends Genet 5:103-107

Gao X, Hou Y, Ebina H, Levin HL, Voytas DF (2008) Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res 18:359-369

Gaut BS, d'Ennequin ML, Peek AS, Sawkins MC (2000) Maize as a model for the evolution of plant nuclear genomes. Proc Natl Acad Sci USA 97:7008-7015.

Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence finishing. Genome Res 8:195-202

Grandbastien MA (1998) Activation of plant retrotransposons under stress conditions. Trends Plant Sci 3:181-187

Han YJ, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38:e199

Hanson RE, Islam-Faridi MN, Crane CF, Zwick MS, Czeschin DG, Wendel JF, McKnight TD, Price HJ, Stelly DM (2000) *Ty1-copia*-retrotransposon behavior in a polyploid cotton. Chromosome Res 8:73-76

Hegarty MJ, Hiscock SJ (2008) Genomic clues to the evolutionary success of review polyploid plants. Curr Biol 18:R435-R444

Heslop-Harrison JS (2000) Comparative genome organization in plants: From sequence and markers to chromatin and chromosomes. Plant Cell 12:617-635

Ho ISH, 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

Jo IH, Bang KH, Kim YC, Lee JW, Seo AY, Seong BJ, Kim HH, Kim DH, Cha SW,

Cho YG, Kim HS (2011) Rapid identification of ginseng cultivars (*Panax ginseng* Meyer) using novel SNP-based probes. J Ginseng Res 35:504-513

Jones N, Pasakinskiene I (2005) Genome conflict in the gramineae. New Phytol 165:391-409

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH (2004) Large retrotransposon derivatives: Abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166:1437-1450

Kidwell MG, Lisch D (1997) Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci USA 94 (15):7704-7711

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

Kim SK, Park JH (2011) Trends in ginseng research in 2010. J Ginseng Res 35:389-398

Kim TD, Han JY, Huh GH, Choi YE (2011) Expression and functional characterization of three squalene synthase genes associated with saponin biosynthesis in *Panax ginseng*. Plant Cell Physiol 52:125-137

Kim YJ, Shim JS, Krishna PR, Kim SY, In JG, Kim MK, Yang DC (2008) Isolation and characterization of a glutaredoxin gene from *Panax ginseng* C. A. Meyer. Plant Mol Biol Rep 26:335-349

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. J Mol Evol 16:111-120

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7: 474

Kronmiller BA, Wise RP (2009) Computational finishing of large sequence contigs reveals interspersed nested repeats and gene islands in the *rf1*-associated region of maize. Plant Physiol 151:483-495

Kubis S, Schmidt T, Heslop-Harrison JS (1998) Repetitive DNA elements as a major component of plant genomes. Ann Bot 82:45-55

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Ann Rev Genet 33:479-532

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH,

Nelson WD, Ploetz L, Singh S, Wensel A, Huala E (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40:D1202-1210

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

Lim KB, Yang TJ, Hwang YJ, Kim JS, Park JY, Kwon SJ, Kim JA, Choi BS, Lim MH, Jin M, Kim HI, de Jong H, Bancroft I, Lim Y, Park BS (2007) Characterization of the centromere and peri-centromere retrotransposons in *Brassica rapa* and their distribution in related *Brassica* species. Plant J 49:765-765

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155

Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ (2009) The dynamic ups and downs of genome size evolution in Brassicaceae. Mol Biol Evol 26:85-98

Ma JX, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101:12404-12410

Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 37:D205-D210

McClintock B (1984) The significance of responses of the genome to challenge. Science 226:792-801

Obae SG, West TP (2012) Nuclear DNA content and genome size of American ginseng. J Med Plant Res 6:4719-4723

Park HJ, Kim DH, Park SJ, Kim JM, Ryu JH (2012a) Ginseng in traditional herbal prescriptions. J Ginseng Res 36:225-241

Park M, Jo S, Kwon JK, Park J, Ahn JH, Kim S, Lee YH, Yang TJ, Hur CG, Kang BC, Kim BD, Choi D (2011) Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. BMC Genomics 12:85

Park M, Park J, Kim S, Kwon JK, Park HM, Bae IH, Yang TJ, Lee YH, Kang BC,

Choi D (2012b) Evolution of the large genome in Capsicum annuum occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. Plant J 69:1018-1029

Plunkett GM, Wen J, Lowry PP (2004) Infrafamilial classifications and characters in Araliaceae: Insights from the phylogenetic analysis of nuclear (ITS) and plastid (*trn*L-*trn*F) sequence data. Plant Syst Evol 245:1-39

Qi LW, Wang CZ, Yuan CS (2011) Isolation and analysis of ginseng: advances and challenges. Nat Prod Rep 28:467-495

Sabot F, Sourdille P, Chantret N, Bernard M (2006) Morgane, a new LTR retrotransposon group, and its subfamilies in wheats. Genetica 128:439-447

Saitou N, Nei M (1987) The Neighbor-Joining method - a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425

Salamov AA, Solovyev VV (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Res 10:516-522

Salina EA, Sergeeva EM, Adonina IG, Shcherban AB, Belcram H, Huneau C, Chalhoub B (2011) The impact of *Ty3-gypsy* group LTR retrotransposons Fatima on B-genome specificity of polyploid wheats. BMC Plant Biol 11:99

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, MelakeBerhan 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

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du FY, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B et al. (2009) The B73 maize genome: Complexity, diversity, and dynamics. Science 326:1112-1115

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003) Human-mouse alignments with BLASTZ. Genome Res 13:103-107

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000) PipMaker - A web server for aligning two genomic DNA sequences. Genome Res 10:577-586

Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res 10:908-915

Smyth DR, Kalitsis P, Joseph JL, Sentry JW (1989) Plant retrotransposon from *Lilium Henryi* is related to *Ty3* of yeast and the *Gypsy* group of *Drosophila*. Proc Natl Acad Sci USA 86:5015-5019

Soltis DE, Soltis PS, Bennett MD, Leitch IJ (2003) Evolution of genome size in the angiosperms. Am J Bot 90:1596-1603

Stack SM (1984) Heterochromatin, the synaptonemal complex and crossing over. J Cell Sci 71:159-176

Sun H, Wang HT, Kwon WS, Kim YJ, In JG, Yang DC (2011) A simple and rapid technique for the authentication of the ginseng cultivar, Yunpoong, using an SNP marker in a large sample of ginseng leaves. Gene 487:75-79

Suoniemi A, Narvanto A, Schulman AH (1996) The *BARE-1* retrotransposon is transcribed in barley from an LTR promoter active in transient assays. Plant Mol Biol 31:295-306

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739

Tansakul P, Shibuya M, Kushiro T, Ebizuka Y (2006) Dammarenediol-II synthase, the first dedicated enzyme for ginsenoside biosynthesis, in *Panax ginseng*. FEBS Lett 580:5143-5149

Thompson JD, Higgins DG, Gibson TJ (1994) ClustalW - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680

Vicient CM, Suoniemi A, Anamthawat-Jonsson K, Tanskanen J, Beharav A, Nevo E, Schulman AH (1999) Retrotransposon *BARE-1* and its role in genome evolution in the genus *Hordeum*. Plant Cell 11:1769-1784

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638-17643

Vitte C, Panaud O (2003) Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in

rice Oryza sativa L. Mol Biol Evol 20:528-540

Vitte C, Panaud O (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res 110:91-107

Waminal N, Park HM, Ryu KB, Kim JH, Yang TJ, Kim HH (2012) Karyotype analysis of *Panax ginseng* C.A. Meyer, 1843 (Araliaceae) based on rDNA loci and DAPI band distribution. Comp Cytogenet 6:425-441.

Wen J (1999) Evolution of eastern Asian and eastern North American disjunct distributions in flowering plants. Annu Rev Ecol Syst 30:421-455

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 Bot 26:144-167

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973-982

Wicker T, Stein N, Albar L, Feuillet C, Schlagenhauf E, Keller B (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. Plant J 26:307-316

Wicker T, Zimmermann W, Perovic D, Paterson AH, Ganal M, Graner A, Stein N (2005) A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley *Hv-eIF4E* locus: recombination, rearrangements and repeats. Plant J 41:184-194

Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 98:13778-13783

Wu B, Wang MZ, Ma YM, Yuan LC, Lu SF (2012) High-throughput sequencing and characterization of the small RNA transcriptome reveal features of novel and conserved microRNAs in *Panax ginseng*. PLOS One 7.

Yang TJ, Kwon SJ, Choi BS, Kim JS, Jin M, Lim KB, Park JY, Kim JA, Lim MH, Kim HI, Lee HJ, Lim YP, Paterson AH, Park BS (2007) Characterization of terminal-repeat retrotransposon in miniature (TRIM) in Brassica relatives. Theor Appl Genet 114:627-636

Yang TJ, Lee S, Chang SB, Yu Y, Jong H, Wing RA (2005) In-depth sequence analysis of the tomato chromosome 12 centromeric region: identification of a

large CAA block and characterization of pericentromere retrotranposons. Chromosoma 114:103-117

Yi TS, Lowry PP, Plunkett GM, Wen J (2004) Chromosomal evolution in Araliaceae and close relatives. Taxon 53:987-1005

Yun TK (2001) Brief introduction of *Panax ginseng* C.A. Meyer. J Korean med sci 16:S3-5

Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, Stelly DM, Wendel JF, Paterson NH (1998) Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res 8:682-682

초록 (ABSTRACT IN KOREAN)

최 홍 일

인삼(*Panax ginseng* C. A. Meyer)은 동아시아 지역에서 수천 년 간 약용으로 사용되어 온 우리나라의 귀중한 자원식물이다. 인삼의 약리적 성분과 그 효능에 대해서는 연구가 다양하게 진행되어 d 왔지만, 유전학이나 유전체학 분야에서는 다른 식물 종에 비해 연구가 미진한 상황이었다. 본 연구는 발현유전자단편(EST, expressed sequence tag)과 세균인공염색체(BAC, bacterial artificial chromosome) 분석을 통하여 인삼의 유전체 구조와 진화과정을 밝히는 것을 목적으로 수행되었다. EST 데이터로부터 paralog와 ortholog 쌍을 동정하고, synonymous substitution rate(Ks)값을 계산하여 분포한 결과, 인삼과 미국삼(*P. quinquefolius* L.)의 공통선조에서 두 차례의 유전체 배가 사건이 일어났고, 그 이후 두 종이 분화하였음을 확인하였다. 반복인자를 다량 함유한 BAC 서열의 분석을 통하여 인삼의 유전체를 구성하는 주요 요소인 long terminal repeat retrotransposon(LTR-RT)들을 동정하였으며, 그 중 *Ty3/Gypsy* 유사인자들이 *Ty1/copia* 유사인자들에 비해 훨씬 많이 존재하는 것을 확인하였다. Whole genome shotgun (WGS) read mapping을 통해 동정

된 LTR-RT들이 인삼 유전체의 많은 부분을 차지함을 밝혀내었다. 특히 *PgDel1* 인자들은 인삼 유전체 내에서 이질염색질 지역의 확장 뿐만 아니라 진정염색질 지역의 개조에도 중요한 역할을 한 것으로 보인다. *PgDel2* 인자들의 경우 전체 중 절반의 염색체에 형광동소보합법(FISH, fluorescence in situ hybridization) 신호가 분포하였는데, 이는 인삼 유전체가 이질배수성의 특성을 가지고 있음을 시사하였다. LTR-RT들의 삽입시기 계산 결과, LTR-RT들이 인삼 유전체의 최근 배가 사건 이후에 증식하였을 가능성을 확인하였다. 이러한 결과들은 오늘날의 거대한 인삼 유전체가 유전체 배가 및 LTR-RT들의 누적에 의해 확장되고 진화하여 왔음을 보여준다.

주요어: 발현유전자단편, 배수성, 세균인공염색체, 인삼 유전체, 형광동소보합법, long terminal repeat retrotransposon (LTR-RT)

학번: 2007-21304