A DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Genome structure and evolution of *Panax ginseng* C. A. Meyer revealed by comprehensive cytogenomic studies

By

NOMAR ESPINOSA WAMINAL

FEBRUARY, 2017

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

COLLEGE OF AGRICULTURE AND LIFE SCIENCES

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Genome structure and evolution of *Panax ginseng* C. A. Meyer revealed by comprehensive cytogenomic studies

UNDER THE DIRECTION OF DR. TAE-JIN YANG

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

BY

NOMAR ESPINOSA WAMINAL

MAJOR IN CROP SCINECE AND BIOTECHNOLOGY
DEPARTMENT OF PLANT SCIENCE

APPROVED AS QUALIFIED DISSERTATION OF NOMAR ESPINOSA WAMINAL FOR THE DEGREE OF DOCTOR OF PHILOSOPHY BY THE COMMITTEE MEMBERS

DECEMBER, 2017

CHAIRMAN	Suk-Ha Lee, Ph.D.
VICE-CHAIRMAN	Tae-Jin Yang, Ph.D.
MEMBER	Hee-Jong Koh, Ph.D.
MEMBER	Hyun Hee Kim, Ph.D.
MEMBER	Nam-Soo Kim, Ph.D.

Genome structure and evolution of *Panax ginseng* C. A. Meyer revealed by comprehensive cytogenomic studies

NOMAR ESPINOSA WAMINAL

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

GENERAL ABSTRACT

Panax ginseng Meyer (Asian ginseng) is highly valued for its diverse pharmacological properties, and it has been regarded as the "king of herbs" in East Asia for centuries. Vigorous researches on its medicinal properties have contributed ample pharmacological data in the literature. Although genetic and genomic researches are recently gaining pace by providing unprecedented insights on the genome composition and corresponding historical events that could have shaped the extant ginseng genome, very limited molecular cytogenetic data are available to support these claims. This study was conducted to provide a cytogenetic foundation in understanding the genome structure and evolution of ginseng by conducting a comprehensive cytogenomic study that involved molecular cytogenetic techniques in conjunction with genome-wide analysis of repetitive elements (REs) and genic blocks. These analyses included characterization of the ginseng genome at the chromosomal level, identification of a high-copy TE, *PgCACTA1*, and a high-copy

tandem repeat, Pg167TR, whose sequence variants were unevenly amplified at the 3' distal region of *PgCACTA1*, cytogenetic mapping of major ginseng REs, and validation of scaffold assembly and recent genome duplication.

In the first chapter, fluorescence *in situ* hybridization (FISH) was utilized to analyze the chromosome composition and karyotype of ginseng using inherent DAPI bands observed in ginseng chromosomes, and three molecular probes namely, 5S and 45S ribosomal RNA genes (rDNA) and Pg167TR. The exact chromosome number of ginseng was determined to be $2n = 48$, a tetraploid considering a basic chromosome number of $x = 12$ in Araliaceae. The combination of these four cytogenetic markers, along with chromosome size and arm ratio, was efficient in characterizing individual ginseng chromosomes. In particular, Pg167TR provided considerable cytogenetic marks that enabled identification of homologous pairs and the establishment of the first FISH-based ginseng karyotype.

In the second chapter, *in silico* analysis revealed several ginseng *CACTA* transposons (*PgCACTA1*) bearing variable Pg167TR unit copy numbers at their 3' distal region, with some carrying over 1,000 copies. Further genome-wide characterization of Pg167TR was carried out, and a putative amplification pathway that gave rise to several long tandem array loci distributed in different chromosomal regions was described. The Pg167TR was highly amplified in the ginseng genome compared with other species in the genus *Panax* and related genera. Two sequence variants were identified, namely Pg167TRa and Pg167TRb. Pg167TRa was more abundant, diverse and associated with amplified Pg167TR arrays than Pg167TRb. While there was a higher ratio of Pg167TRa to Pg167TRb among species in the genus *Panax*, an opposite pattern was observed in species from related genera. Putative *cis*-regulatory elements related to biotic and abiotic stress responses were identified in Pg167TR, implying a functional role of Pg167TR in ginseng physiology. FISH analysis using a transposase domain and Pg167TR regions as separate probes supported an amplification of Pg167TR array from *CACTA* elements. This presents

an alternative pathway of satDNA evolution from Class II TE, particularly *CACTA* DNA transposons.

In the third and final chapter, cytogenetic mapping and analysis of the distribution pattern of the major ginseng REs that were characterized previously were carried out. In addition, cytogenetic techniques were used to investigate the recent whole-genome duplication (WGD) event involved in shaping the ginseng genome by pooling PCR-amplified assembly scaffold-linked genic blocks as FISH probes. Cytogenetic mapping of major ginseng REs showed a more comprehensive genomic distribution of different repeat families, which revealed their distribution in different chromosomal niches. Some preferentially localized in pericentromeric area, and some predominantly in subtelomeric regions. FISH analysis with paralogous gene blocks supports a recent WGD while simultaneously validating the assembly of these two paralogous blocks.

These analyses established the first FISH-based ginseng karyotype, presented an alternative pathway for satDNA evolution from *CACTA* transposons, and revealed the preferential chromosomal localization of different TE families. Altogether, this enabled further understanding of the structure and evolutionary history of the ginseng genome. This information will further provide a framework for future cytogenetic analyses of ginseng and its related species, allow validation of the assembly of the ginseng genome, and facilitate integration of genetic and cytogenetic maps for better ginseng crop improvement programs.

Keywords: *Panax ginseng* karyotype, fluorescence *in situ* hybridization, Pg167TR, *PgCACTA1*, allopolyploidization, single-copy FISH

Student number: 2012-30756

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BAC	Bacterial artificial chromosomes
BLAST	Basic local alignment search tool
CDS	Coding sequences
dnaLCW	*de novo assembly* of low-coverage whole genome sequences
FISH	Fluorescence *in situ* hybridization
GR	Genome representation
LC	Length of contig
LTR	Long terminal repeat
OP	oligoprobe
PCR	Polymerase chain reactions
RD	Read depth
rDNA	ribosomal DNA
REs	Repetitive elements
satDNA	satellite DNA
TEs	Transposable elements
TIRs	Terminal inverted repeats
TRs	Tandem repeats
TSDs	Target site duplications
WGD	Whole genome duplication
WGS	Whole genome sequences

GENERAL INTRODUCTION

Panax ginseng (Asian ginseng or ginseng) is highly valued for its medicinal properties owing to ginsenosides found in the plant (Court 2000; Leung and Wong 2010; Zhang *et al.* 2011; Park *et al.* 2012a). While numerous studies have been conducted on the pharmacological aspects of ginseng for several decades (Hu 1976; Wu and Zhong 1999; Court 2000; Yun 2001a; Ki *et al.* 2013; Liu *et al.* 2013; Kim *et al.* 2014a), dedicated research aimed at understanding its genome structure, composition, and history have just gained pace in recent years (Choi *et al.* 2014; Kim *et al.* 2014b; Kim *et al.* 2014c; Kim *et al.* 2014d; Li *et al.* 2015; Shi *et al.* 2015b). We now understand, that, like many other angiosperm species, the ginseng genome is also replete with transposable elements (TEs) (Choi *et al.* 2014). Moreover, what and how transposable elements localize in their respective chromosomal regions often provide clues about their functions in genome maintenance and evolution (Fedoroff 2013; Fedoroff and Bennetzen 2013). Comprehensive analysis of TE distribution in a genome, therefore, furnishes important information not only on how a genome is organized but also on how an extant genome came to be.

Cytogenomics approaches involve the use of molecular cytogenetics, most commonly fluorescence *in situ* hybridization (FISH), and genomics techniques in studying genome-wide abundance or distribution of, but not limited to, repetitive elements (Macas *et al.* 2007; Macas *et al.* 2009; Lou *et al.* 2014; Waminal *et al.* 2015; Waminal *et al.* 2016b). In fact, this approach has elucidated the genome composition and organization of several plant genomes including those of *Pisum sativum* (Macas *et al.* 2007) and some *Brassica* spp. (Waminal *et al.* 2015). While there is very limited chromosomal, more so, cytogenomic, data for ginseng, the on-going ginseng genome analysis for crop improvement provides opportunities to carry out a comprehensive cytogenomic study that aims to understand the genome organization and evolution of ginseng, and by so doing, identify elements that play important roles in ginseng genome structure and function.

The extant genomes of most species have undergone numerous rounds of rearrangements through expansion and subsequent contraction which are often facilitated by TE interplays (Soltis *et al.* 2009; Tank *et al.* 2015). Similarly, a recent study has shown some molecular evidence suggesting that the ginseng genome has undergone two rounds of whole genome duplications (WGD), with the more recent duplication being an allotetraploidization event responsible for the doubling of ginseng chromosome number to $2n = 48$ compared with other diploid species (Choi *et al.* 2014; Kim *et al.* 2014c). A previous BAC sequencing analysis identified LTR retrotransposons of the *PgDel* family as major players in ginseng genome expansion and evolution (Choi *et al.* 2014). In addition, a 167-bp tandem repeat (TR) DNA in few copies was also identified, but a comprehensive cytogenomic distribution has not been carried out.

In this study, a cytogenomic mapping of this TR, which was named Pg167TR, was carried out. A description on how its distinct genomic distribution in different chromosomes facilitated an efficient discrimination of the 24 ginseng chromosome, thus, establishing the first refined ginseng karyotype was presented. Considering the several studies that have shown the evolutionary pathways of TRs from TEs (Sharma *et al.* 2013; Mehrotra and Goyal 2014), The sequence features of Pg167TR were analyzed, and its associated TE family, *PgCACTA1*, was identified. Data which linked *PgCACTA1*-bound Pg167TR units as 'seed' repeat for its expansion to long satellite arrays were presented.

Cytogenomic study of Pg167TR and other major REs in the ginseng genome gave a clear picture about the structure and evolution of the ginseng genome. Moreover, an attempt to validate the contiguity of two assembly scaffolds did not only confirm the quality of the current ginseng genome assembly but also showed the distribution of paralogous genic blocks from the recent WGD. Understanding the genomic structure and evolution of ginseng and the establishment of its karyotype will be necessary in further cytogenomic, genetic, and TE functional studies.

LITERATURE REVIEW

Panax ginseng

Panax ginseng C. A. Meyer belongs to the family Araliaceae (Fig. 1) (Yi *et al.* 2004). It is a very slow-growing, light- and temperature-sensitive perennial plant that is propagated from seeds obtained from healthy 5-year old plants (Court 2000). *P. ginseng* has been used for millennia in eastern Asia to treat various ailments and is considered as the king of eastern medicinal plants (Yun 2001b; Xie *et al.* 2005; Shi *et al.* 2015a). Saponins called ginsenosides are the main biochemical compound that give *P. ginseng* its medicinal value (Leung and Wong 2010).

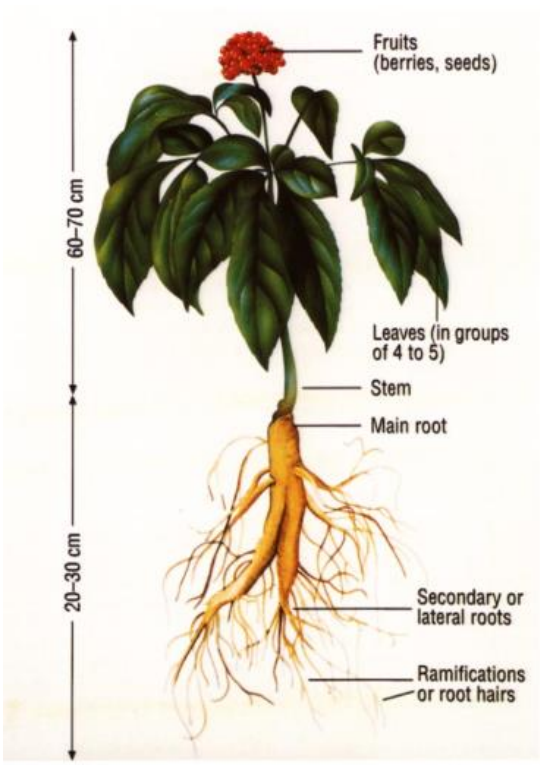


Fig. 1-1. *Panax ginseng* morphology. Adapted from Court *et al.* 2000.

The genus *Panax*

There are currently 12 species in the genus *Panax* (Wen and Zimmer 1996), and three of these (*P. ginseng*, *P. quinquefolius*, and *P. notoginseng*) are commercially cultivated (Zuo *et al.* 2011). Eastern Asia is considered as the center

of diversity while two species, *P. quinquefolius* and *P. trifolius,* are geographically isolated in eastern North America (Wen and Zimmer 1996). The name "Panax" is derived from the Greek word "pan" and "axos" which means "all-heal" or "all-cure" (Court 2000; Yun 2001b).

The genus *Panax* have both diploid and tetraploid species, and those with 2*n* = 24 chromosomes are considered diploid while those with 2*n* = 48 are considered tetraploids considering a basic chromosome number of *x* = 12 in the family Araliaceae (Yi *et al.* 2004). *P. ginseng* and *P. quinquefolius* have 2*n* = 48 chromosomes and are considered allotetraploid species based on cytogenomic studies (Yi *et al.* 2004; Waminal *et al.* 2012; Choi *et al.* 2014).

Two known major rounds of whole-genome duplications (WGD) have shaped the extant *P. ginseng* genome (Fig. 2). An ancient WGD was estimated to have occurred around 24.6 ~ 32.8 million years ago (MYA) and a recent one occurred around 1.6 ~ 3.3 MYA (Choi *et al.* 2013). The divergence of tetraploid species, *P. ginseng* and *P. quinquefolius*, from diploid *P. notoginseng* and that between the two tetraploids were estimated to have occurred around 3.5 ~ 5.2 MYA and 0.8 ~ 1.2 MYA, respectively (Choi *et al.* 2013; Kim *et al.* 2013). *In silico* analyses have revealed the association of the recent WGD with rapid amplification of transposable elements (TEs), mostly of the LTR type (Choi *et al.* 2014; Kim *et al.* 2014b). Currently, the assembly of the *P. ginseng* genome is ongoing and more relevant information should be available soon.

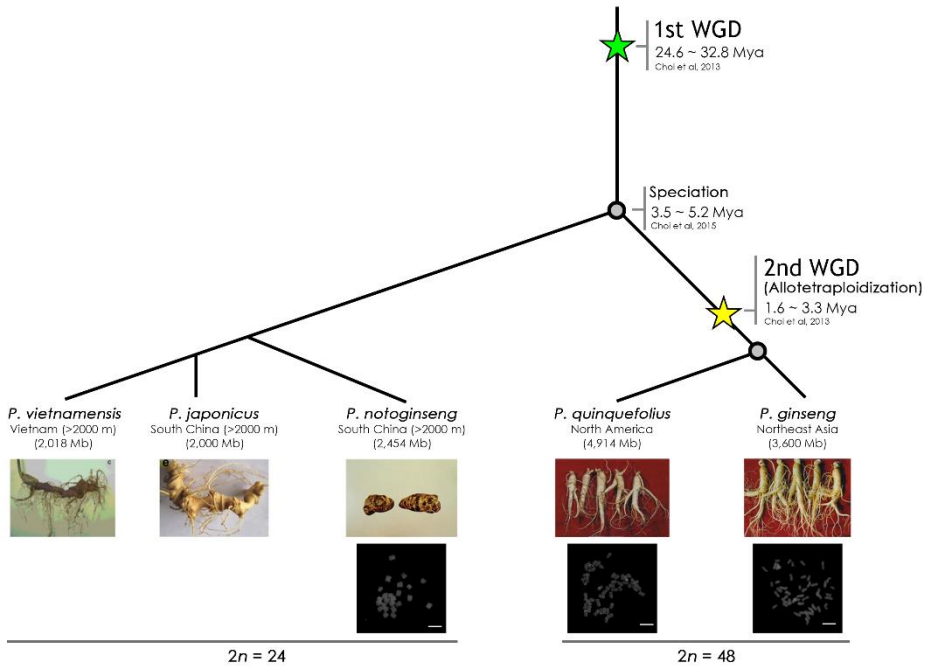


Fig. 1-2. Whole genome duplications in *Panax*. Diploid and tetraploid species with corresponding geographical distribution, genome sizes and chromosome numbers are presented. Bars in inset = 10 μm.

Cytogenomics and FISH

An integral part of genomic research is to understand the structure of a genome. In fact, having sufficient information about a genome's structure would greatly facilitate in elucidating the function of its components as well as in exploiting these various data to achieve the desired product. With the abundance of whole-genome sequence (WGS) information and the advancement of bioinformatics pipelines to analyze these data, more and more genome scale repetitive and unique gene sequences are assembled (Novák *et al.* 2013; Waminal *et al.* 2016b). Coupled with molecular cytogenetic techniques, mainly fluorescence *in situ* hybridization (FISH), a quicker and more informative approach called 'cytogenomics' has been carried out successfully in plants and animals (Macas *et al.* 2007; Dias *et al.* 2014; McPherson *et al.* 2014).

Fluorescence *in situ* hybridization (FISH) is a powerful technique in visualizing the actual chromosomal location of a target DNA sequence in its native cellular environment (Fig. 3) (Levsky and Singer 2003). It exploits the nature of DNA to hybridize with complementary sequences, and by incorporating reporter molecules to the DNA probe, the location of the target DNA in a cell, more often at metaphase stage chromosomes, can be identified (Levsky and Singer 2003; Bishop 2010). It has a wide array of applications that include chromosomal detection of repetitive elements and single-genes (Khrustaleva and Kik 2001; Choi *et al.* 2014), integration of physical and genetic maps (Capdeville *et al.* 2008; Szinay *et al.* 2008; Tang *et al.* 2009; Chamala *et al.* 2013), detection of chromosomal translocations (Huang *et al.* 2009), understanding phylogenetic relationships (Siljak-Yakovlev *et al.* 2014; Sousa *et al.* 2014; Fajkus *et al.* 2016), quantification of mRNA transcripts (Trcek *et al.* 2012), and diagnosis of some hematologic cancer (Bishop 2010; Hu *et al.* 2014).

In the context of genomics, FISH has been very useful in mapping repetitive elements, physical mapping of BAC clones, and validating genome assemblies (Chester *et al.* 2010; Chamala *et al.* 2013; Waminal *et al.* 2016a; Waminal *et al.* 2016b). The use of FISH especially in chromosomal mapping of repetitive elements is of particular importance in structural and evolutionary genomics because it allows understanding of the dynamics of repetitive elements; hence, their phylogenetic impact (Choi *et al.* 2014; Waminal *et al.* 2016a; Waminal *et al.* 2016b).

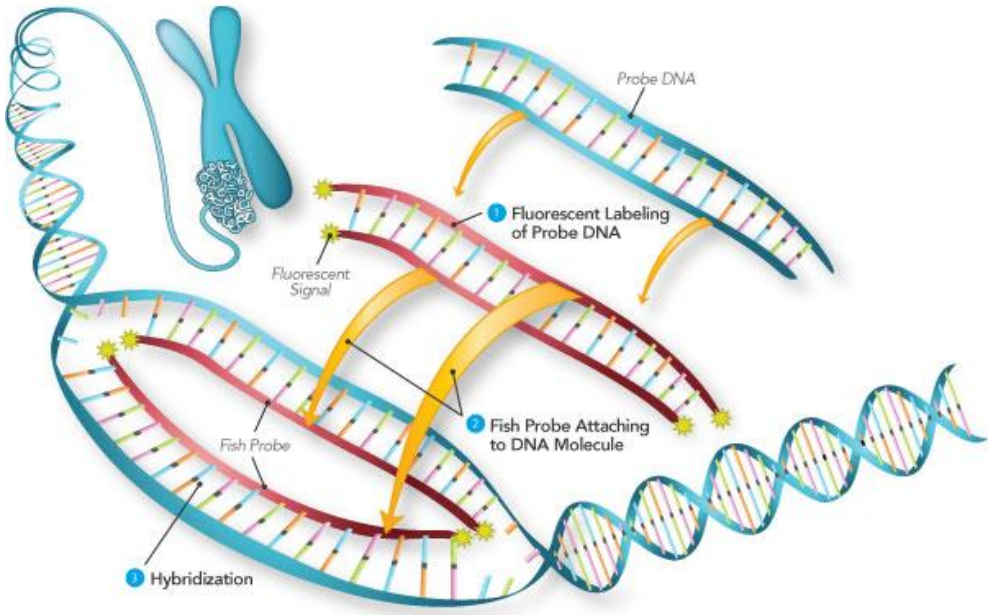


Fig. 1-3. Simplified concept of FISH. Adapted from Semrock Inc. (https://www.semrock.com/fish.aspx)

Repetitive DNA elements in plant genomes

Chromosomal DNA can be classified as either euchromatin or heterochromatin (Allis and Jenuwein 2016). The euchromatic regions contain most of the protein-coding genes while heterochromatic regions are packed with repetitive elements (REs) (Strålfors and Ekwall 2006). In addition, heterochromatin can be subdivided into two groups, constitutive and facultative, according to their distribution and condensation characteristics. Constitutive heterochromatins are mainly composed of REs, and can often be observed in centromeric, pericentromeric, telomeric, and subtelomeric regions (Park *et al.* 2012b). In addition, they can also be seen as condensed chromatins throughout the cell cycle in cytological observations. In contrast, facultative heterochromatins are present in gene-rich regions and can change from heterochromatin to euchromatin depending on cellular responses to particular signals (Park *et al.* 2012b; Allis and Jenuwein 2016).

Heterochromatins make up a considerable proportion of plant genomes, even more than 80% of some plant genomes (Michael and Jackson 2013). REs in heterochromatins can be categorized into two major types based on their distribution in a genome – transposable elements (TEs) and tandem repeats (TRs) (Fig. 4) (Waminal *et al.* 2016a). The former are distributed as dispersed elements either in constitutive heterochromatic regions or in facultative heterochromatin regions as gene regulators (Allis and Jenuwein 2016). The latter have more distinct distribution in chromosomal regions and are arranged in a head-to-tail fashion.

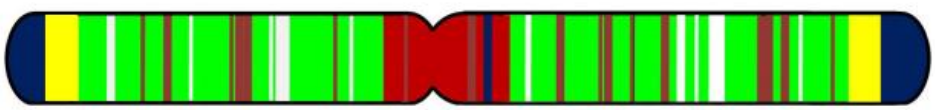


Fig. 1-4. Typical distribution of REs and genes in plants. Red, centromeric tandem repeats; blue, telomeric repeats; yellow, subtelomeric tandem repeats; green, intercalary tandem repeats; brown, dispersed repeats; white, genes and low-copy sequences. Adapted from Mehrotra *et al.* 2014.

TEs classification is still progressing owing to more and more elements that are being discovered as more genomes are assembled (Piegu *et al.* 2015). However, according to a widely accepted classification, TEs can be grouped as either Class I or Class II based on their transposition intermediate (Wicker *et al.* 2007)]. Class I or retrotransposons use RNA intermediate for transposition and are not excised from their original donor site; thus, they are also referred to as the "copy-and-paste" TEs and may explain their predominance in most plant genomes. On the other hand, Class II often transpose via "cut-and-paste" mechanism and are often less abundant than Class I. In most plant genomes, the LTR order of retrotransposons are often the most abundant among all TEs (Tenaillon *et al.* 2010; Estep *et al.* 2013; Choi *et al.* 2014; Macas *et al.* 2015). In Class II, the order TIR is characterized by an inverted repeats at both ends; hence, TIR for Terminal Inverted Repeat. Within this order, the *CACTA* superfamily has a characteristic 5-bp start in its TIR. In more recent classification,

this superfamily is categorized in the *CMC* superfamily owing to high similarity with *Mirage* and *Chapaev* groups; hence, *CMC* for *CACTA*, *Mirage*, and *Chapaev*. Although this superfamily is often not as abundant as *Gypsy* or *Copia* superfamilies in the order LTR, members are often known to have important roles as gene regulators (Yuan and Wessler 2011).

TRs are classified as either microsatellites, minisatellites, or satellites (satDNA) depending on the length of the repeat unit and the size of repeat array in each locus (Mehrotra and Goyal 2014). Microsatellite repeat units range from 2 ~ 5 bp and have 10 ~ 100 copies per locus. On the other hand, minisatellites have 6 ~ 100 bp units and 0.5 ~ 30 kb copies per locus, while satellites have 101 ~ 500 bp units and up to 5 Mb array size. Most satDNA evolution is often associated with LTR retrotransposons, perhaps due to their more abundance in most plant genomes, which could increase their likelihood of initiating homology-based satDNA expansion. On the other hand, very few reports have associated Class II elements with satDNA amplification, although some wheat satDNA have shown to have high homology with repeats in *CACTA* elements (Nagaki *et al.* 1998; Wicker *et al.* 2003).

REFERENCES

Allis CD, and Jenuwein T. 2016. The molecular hallmarks of epigenetic control. Nat Rev Genet 17: 487-500.

Bishop R. 2010. Applications of fluorescence in situ hybridization (fish) in detecting genetic aberrations of medical significance. Bioscience Horizons: hzq009.

Capdeville G, Souza Júnior MT, Szinay D, Diniz LEC, Wijnker E, Swennen R, *et al.* 2008. The potential of high-resolution bac-fish in banana breeding. Euphytica 166: 431-443.

Chamala S, Chanderbali AS, Der JP, Lan T, Walts B, Albert VA, *et al.* 2013. Assembly and validation of the genome of the nonmodel basal angiosperm amborella. Science 342: 1516-1517.

Chester M, Leitch AR, Soltis PS, and Soltis DE. 2010. Review of the application of modern cytogenetic methods (fish/gish) to the study of reticulation (polyploidy/hybridisation). Genes 1: 166-192.

Choi HI, Kim NH, Lee J, Choi BS, Do Kim K, and Park JY. 2013. Evolutionary relationship of panax ginseng and p. Quinquefolius inferred from sequencing and comparative analysis of expressed sequence tags. Genet Resour Crop Evol 60.

Choi HI, Waminal NE, Park HM, Kim NH, Choi BS, Park M, *et al.* 2014. Major repeat components covering one-third of the ginseng (panax ginseng c.A. Meyer) genome and evidence for allotetraploidy. Plant J 77: 906-916.

Court WE. 2000. Ginseng: The genus *panax*, Amsterdam, Netherlands, Hardwood Academic Publishers.

Dias GB, Svartman M, Delprat A, Ruiz A, and Kuhn GC. 2014. Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of drosophila virilis. Genome Biol Evol 6: 1302-13.

Estep MC, Debarry JD, and Bennetzen JL. 2013. The dynamics of ltr retrotransposon accumulation across 25 million years of panicoid grass evolution. Heredity 110: 194-204.

Fajkus P, Peska V, Sitova Z, Fulneckova J, Dvorackova M, Gogela R, *et al.* 2016. Allium telomeres unmasked: The unusual telomeric sequence (ctcggttatggg)n is synthesized by telomerase. Plant J 85: 337-47.

Fedoroff NV 2013. Mcclintock and epigenetics. *Plant transposons and genome dynamics in evolution.* Wiley-Blackwell.

Fedoroff NV, and Bennetzen JL 2013. Transposons, genomic shock, and genome evolution. *Plant transposons and genome dynamics in evolution.* Wiley-Blackwell.

Hu L, Ru K, Zhang L, Huang Y, Zhu X, Liu H, *et al.* 2014. Fluorescence in situ hybridization (fish): An increasingly demanded tool for biomarker research and personalized medicine. Biomarker Research 2: 3-3.

Hu SY. 1976. The genus panax (ginseng) in chinese medicine. Economic botany 30: 11-28.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, *et al.* 2009. The genome of the cucumber, *cucumis sativus* l. Nature Genetics 41: 1275-81.

Khrustaleva LI, and Kik C. 2001. Localization of single-copy t-DNA insertion in transgenic shallots (allium cepa) by using ultra-sensitive fish with tyramide signal amplification. Plant J 25: 699-707.

Ki SH, Yang JH, Ku SK, Kim SC, Kim YW, and Cho IJ. 2013. Red ginseng extract protects against carbon tetrachloride-induced liver fibrosis. Journal of Ginseng Research 37: 45-53.

Kim JH, Jung J-Y, Choi H-I, Kim N-H, Park JY, Lee Y, *et al.* 2013. Diversity and evolution of major panax species revealed by scanning the entire chloroplast intergenic spacer sequences. Genetic resources and crop evolution 60: 413-425.

Kim JR, Choi J, Kim J, Kim H, Kim EH, Chang JH, *et al.* 2014a. 20-o-beta-d-glucopyranosyl-20(s)-protopanaxadiol-fortified ginseng extract attenuates the development of atopic dermatitis-like symptoms in nc/nga mice. Journal of Ethnopharmacology 151: 365-371.

Kim N-H, Lee S-C, Choi H-I, Kim K, Choi B-S, Jang W, *et al.* Genome sequences and evolution of *panax ginseng*. *In:* KIM, S.-K. & RHEE, D.-K., eds. Evidence-based Efficacy of Ginseng 2014, Proceedings of the 11th International symposium on Ginseng, 2014b Seoul, Korea. The Korean Society of Ginseng, 125-146.

Kim NH, Choi HI, Kim KH, Jang W, and Yang TJ. 2014c. Evidence of genome duplication revealed by sequence analysis of multi-loci expressed sequence

tag-simple sequence repeat bands in panax ginseng meyer. J Ginseng Res 38: 130-5.

Kim NH, Choi HI, Kim KH, Jang W, and Yang TJ. 2014d. Evidence of genome duplication revealed by sequence analysis of multi-loci expressed sequence tag-simple sequence repeat bands in panax ginseng meyer. J Ginseng Res 38.

Leung KW, and Wong AS. 2010. Pharmacology of ginsenosides: A literature review. Chinese Medicine 5: 20.

Levsky JM, and Singer RH. 2003. Fluorescence in situ hybridization: Past, present and future. J Cell Sci 116: 2833-2838.

Li MR, Shi FX, Zhou YX, Li YL, Wang XF, and Zhang C. 2015. Genetic and epigenetic diversities shed light on the domestication of cultivated ginseng (panax ginseng). Mol Plant 8.

Liu QM, Jung HM, Cui CH, Sung BH, Kim JK, Kim SG, *et al.* 2013. Bioconversion of ginsenoside rc into rd by a novel alpha-l-arabinofuranosidase, abf22-3 from leuconostoc sp 22-3: Cloning, expression, and enzyme characterization. Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology 103: 747-754.

Lou Q, Zhang Y, He Y, Li J, Jia L, Cheng C, *et al.* 2014. Single-copy gene-based chromosome painting in cucumber and its application for chromosome rearrangement analysis in cucumis. Plant J 78: 169-79.

Macas J, Koblížková A, Navrátilová A, and Neumann P. 2009. Hypervariable 3′ utr region of plant ltr-retrotransposons as a source of novel satellite repeats. Gene.

Macas J, Neumann P, and Navratilova A. 2007. Repetitive DNA in the pea (pisum sativum l.) genome: Comprehensive characterization using 454 sequencing and comparison to soybean and medicago truncatula. BMC Genomics 8: 427.

Macas J, Novak P, Pellicer J, Cizkova J, Koblizkova A, Neumann P, *et al.* 2015. In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe fabeae. Plos One 10.

Mcpherson MC, Robinson CM, Gehlen LP, and Delany ME. 2014. Comparative cytogenomics of poultry: Mapping of single gene and repeat loci in the japanese quail (coturnix japonica). Chromosome Res 22: 71-83.

Mehrotra S, and Goyal V. 2014. Repetitive sequences in plant nuclear DNA: Types, distribution, evolution and function. Genomics, Proteomics & Bioinformatics 12: 164-171.

Michael TP, and Jackson S. 2013. The first 50 plant genomes. Plant Genome 6.

Nagaki K, Tsujimoto H, and Sasakuma T. 1998. Dynamics of tandem repetitive afa-family sequences in triticeae, wheat-related species. J Mol Evol 47: 183-9.

Novák P, Neumann P, Pech J, Steinhaisl J, and Macas J. 2013. Repeatexplorer: A galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29: 792-793.

Park HJ, Kim DH, Park SJ, Kim JM, and Ryu JH. 2012a. Ginseng in traditional herbal prescriptions. J Ginseng Res 36: 225-41.

Park M, Park J, Kim S, Kwon JK, Park HM, Bae IH, *et al.* 2012b. Evolution of the large genome in *capsicum annuum* occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. The Plant Journal 69: 1018-29.

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

Sharma A, Wolfgruber TK, and Presting GG. 2013. Tandem repeats derived from centromeric retrotransposons. BMC Genomics 14: 142.

Shi F-X, Li M-R, Li Y-L, Jiang P, Zhang C, Pan Y-Z, *et al.* 2015a. The impacts of polyploidy, geographic and ecological isolations on the diversification of panax (araliaceae). BMC Plant Biology 15: 297.

Shi F-X, Li M-R, Li Y-L, Jiang P, Zhang C, Pan Y-Z, *et al.* 2015b. The impacts of polyploidy, geographic and ecological isolations on the diversification of panax (araliaceae). BMC Plant Biology 15: 297.

Siljak-Yakovlev S, Pustahija F, Vicic V, and Robin O. 2014. Molecular cytogenetics (fish and fluorochrome banding): Resolving species relationships and genome organization. Methods Mol Biol 1115: 309-23.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, *et al.* 2009. Polyploidy and angiosperm diversification. Am J Bot 96: 336-48.

Sousa A, Cusimano N, and Renner SS. 2014. Combining fish and model-based predictions to understand chromosome evolution in typhonium (araceae). Ann Bot 113: 669-80.

Strålfors A, and Ekwall K 2006. Heterochromatin and euchromatin—organization, boundaries, and gene regulation. *Reviews in cell biology and molecular medicine.* Wiley-VCH Verlag GmbH & Co. KGaA.

Szinay D, Chang SB, Khrustaleva L, Peters S, Schijlen E, Bai Y, *et al.* 2008. High-resolution chromosome mapping of bacs using multi-colour fish and pooled-bac fish as a backbone for sequencing tomato chromosome 6. Plant J 56: 627-37.

Tang X, De Boer JM, Van Eck HJ, Bachem C, Visser RG, and De Jong H. 2009. Assignment of genetic linkage maps to diploid solanum tuberosum pachytene chromosomes by bac-fish technology. Chromosome Res 17: 899-915.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, *et al.* 2015. Nested radiations and the pulse of angiosperm diversification: Increased diversification rates often follow whole genome duplications. New Phytol 207: 454-67.

Tenaillon MI, Hollister JD, and Gaut BS. 2010. A triptych of the evolution of plant transposable elements. Trends in Plant Science 15: 471-478.

Trcek T, Chao JA, Larson DR, Park HY, Zenklusen D, Shenoy SM, *et al.* 2012. Single-mrna counting using fluorescent in situ hybridization in budding yeast. Nat Protoc 7: 408-19.

Waminal N, Park HM, Ryu KB, Kim JH, Yang TJ, and Kim HH. 2012. Karyotype analysis of panax ginseng c.A.Meyer, 1843 (araliaceae) based on rdna loci and dapi band distribution. Comparative Cytogenetics 6: 425-441.

Waminal NE, Choi H-I, Kim N-H, Jang W, Lee J, Park JY, *et al.* 2016a. A refined panax ginseng karyotype based on an ultra-high copy 167-bp tandem repeat and ribosomal dnas. Journal of Ginseng Research.

Waminal NE, Perumal S, Lee J, Kim HH, and Yang T-J. 2016b. Repeat evolution in brassica rapa (aa), b. Oleracea (cc), and b. Napus (aacc) genomes. Plant Breeding and Biotechnology 4: 107-122.

Waminal NE, Perumal S, Lim K-B, Park B-S, Kim HH, and Yang T-J 2015. Genomic survey of the hidden components of the b. Rapa genome. *The brassica rapa genome.* Springer.

Wen J, and Zimmer EA. 1996. Phylogeny and biogeography of *panax* l. (the ginseng genus, araliaceae): Inference from its sequences of nuclear ribosomal DNA. Molecular Phylogenetics and Evolution 6: 167-177.

Wicker T, Guyot R, Yahiaoui N, and Keller B. 2003. Cacta transposons in triticeae. A diverse family of high-copy repetitive elements. Plant Physiol 132: 52-63.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, *et al.* 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973-982.

Wu J, and Zhong JJ. 1999. Production of ginseng and its bioactive components in plant cell culture: Current technological and applied aspects. Journal of Biotechnology 68: 89-99.

Xie J-T, Mehendale S, and Yuan C-S. 2005. Ginseng and diabetes. The American journal of Chinese medicine 33: 397-404.

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

Yuan Y-W, and Wessler SR. 2011. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proceedings of the National Academy of Sciences 108: 7884-7889.

Yun T-K. 2001a. Brief introduction of *panax ginseng* c.A. Meyer. Journal of Korean Medical Science 16(suppl): S3-5.

Yun TK. 2001b. Brief introduction of panax ginseng ca meyer. Journal of Korean medical science 16: S3.

Zhang JY, Bae TW, Boo KH, Sun HJ, Song IJ, Pham CH, *et al.* 2011. Ginsenoside production and morphological characterization of wild ginseng (panax ginseng meyer) mutant lines induced by gamma-irradiation ((60)co) of adventitious roots. J Ginseng Res 35: 283-93.

Zuo Y, Chen Z, Kondo K, Funamoto T, Wen J, and Zhou S. 2011. DNA barcoding of panax species. Planta Med 77.

CHAPTER I

Characterization of *Panax ginseng* chromosomes by FISH using a novel satellite DNA, Pg167TR

ABSTRACT

Panax ginseng Meyer (Asian ginseng) has a large nuclear genome of > 3.5 Gbp in a haploid genome equivalent of 24 chromosomes. Tandem repeats (TRs), particularly satellite DNAs (satDNA) occupy significant portions of the genome in many plants and are often found in specific chromosomal loci, making them a valuable molecular cytogenetic tool in discriminating chromosomes. In an effort to understand the *P. ginseng* genome structure, we characterized an ultrahigh copy 167-bp satDNA (Pg167TR) and explored its chromosomal distribution as well as its utility for chromosome identification. Polymerase chain reaction amplicons of Pg167TR were labeled, along with 5S and 45S rDNA amplicons, using a direct nick-translation method. Direct fluorescence *in situ* hybridization (FISH) was carried out to analyze the chromosomal distribution of Pg167TR and both 5S and 45S rDNAs. Here, a unique distribution of Pg167TR in all 24 *P. ginseng* chromosomes was observed, allowing easy identification of individual homologous chromosomes. Additionally, three 5S and one 45S rDNA loci were observed. Identification of individual *P. ginseng* chromosomes was achieved using cytogenetic markers such as chromosome arm ratio, 5S and 45S rDNA, DAPI bands, and mostly from Pg167TR signals. Chromosome identification is important in understanding the *P. ginseng* genome structure, and our method will be useful for future integration of genetic linkage maps and genome scaffold anchoring. Additionally, it is a good tool for comparative studies with related species in efforts to understand the evolution of *P. ginseng.*

INTRODUCTION

The presence of repetitive elements (REs) is mainly responsible for the huge variations in nuclear genome size among angiosperms (Macas *et al.* 2009; Michael and Jackson 2013). REs are categorized as either dispersed repeats, i.e. transposable elements (TEs), or tandem repeats (TRs). As their names denote, dispersed REs are often loosely distributed throughout the genome (Choi *et al.* 2014), while TRs are organized in a head-to-tail fashion in distinct chromosomal regions (Coluccia *et al.* 2011; Sharma *et al.* 2013).

Satellite DNAs (satDNAs) are a special class of TRs that consist of monomers, often of 150~400 bp but occasionally reaching a few thousand bp, whose repeat array size can extend to several hundred Mb (Macas *et al.* 2009; Mehrotra and Goyal 2014). Numerous satDNA families have been discovered in plants, and most of these include the ribosomal RNA gene families (5S and 45S rDNAs), centromeric and subtelomeric TRs, and heterochromatin 'knobs' (Ananiev *et al.* 1998; Palomeque and Lorite 2008; Plohl *et al.* 2008; Koo *et al.* 2011; Melters *et al.* 2013; Sharma *et al.* 2013; Mehrotra and Goyal 2014; Mondin *et al.* 2014; Waminal *et al.* 2016). The resulting distinct chromosomal distribution of satDNAs allows their exploitation as efficient cytogenetic markers in identifying homologous and homeologous chromosomes, thus facilitating karyotyping and comparative cytogenetics among closely related taxa (Matyasek *et al.* 1997; Albert *et al.* 2010; Mendes *et al.* 2011b; Xiong and Pires 2011).

Fluorescence *in situ* hybridization (FISH) is a powerful tool to physically localize genes and repetitive elements in chromosomes (Pita *et al.* 2014). Probes can be prepared through random priming (Smith 1993), PCR (Lion and Haas 1990), or nick translation (Rigby *et al.* 1977) labeling methods, which incorporate haptens or fluorochromes indirectly or directly, respectively (Pita *et al.* 2014).

Panax ginseng Meyer (Asian ginseng) is a perennial herb highly valued for its ginsenosides, which are reputed to have a wide range of medicinal effects (Court 2000; Leung and Wong 2010; Park *et al.* 2012). *P. ginseng* is the most widely studied species in the genus *Panax*, but most studies have focused on the plant's pharmacological effects (Leung and Wong 2010; Park *et al.* 2012). Nevertheless, there has been a growing interest in *P. ginseng* molecular cytogenetic and genomic studies in recent years (Waminal *et al.* 2012; Choi *et al.* 2014).

P. ginseng has a large haploid nuclear genome of over 3.5 Gbp in 24 chromosomes (Yi *et al.* 2004; Waminal *et al.* 2012; Choi *et al.* 2014; Kim *et al.* 2014). The large genome size compared with other species in the genus, is attributed to two rounds of whole genome duplications (WGD) that occurred 24.6∼32.8 and 1.6∼3.3 million years ago prior to its divergence from *P. quinquefolius* (Choi *et al.* 2013). Consequently, both *P. ginseng* and *P. quinquefolius* are known tetraploids having $2n = 4x = 48$ chromosomes. Despite the shared WGD, *P. quinquefolius* has approximately 4.9 Gbp haploid genome equivalent, which is about 1.4 Gbp larger than that of *P. ginseng* although both have the same number of chromosomes (24 chromosome pairs). The big variation in genome size between these two closely related *Panax* species, *P. ginseng* and *P. quinquefolius*, is an interesting feature to be elucidated further.

A large proportion of the *P. ginseng* genome is covered by REs (Choi *et al.* 2014). In our previous analysis of three repeat-rich *P. ginseng* BAC clones, we identified a 167-bp TR (Choi *et al.* 2014). Here, we present the sequence characteristics, chromosomal distribution and cytogenetic marker potential of this *P. ginseng* TR (Pg167TR). Additionally, we demonstrated the efficiency of direct nick-translation labeling of FISH probes to detect smaller 5S rDNA loci. Altogether, these analyses enabled the establishment of the first comprehensive karyotype of *P. ginseng*.

MATERIALS AND METHODS

Root sample preparation

Stratified seeds of three ginseng cultivars 'Sunun' were provided by the Korea Ginseng Corporation (KGC) Natural Resources Research Institute (Daejeon, Korea). Stratified seeds were allowed to germinate in petri dishes with wet filter papers at 10-15ºC. The root meristems were then excised (about 2 cm from the root tips), pretreated with 0.002M 8-hydroxyquinoline for 5 hours at 18ºC, fixed in 90% acetic acid for 15 min at room temperature (RT, ~24ºC), and then stored in 70% ethanol until use.

Fluorescence *in situ* hybridization (FISH) analysis

Chromosome spread preparation

Somatic chromosome spreads were obtained using a modified version of the technique described by Kato *et al.* (2004). After thorough washing with distilled water, the meristematic regions of the fixed root tips (~2 mm) were excised and digested in a pectolytic enzyme mix [2% cellulase (MB Cell, Korea), 1.5% macerozyme (Maxim Bio, USA) and 1% pectolyase (Sigma, Japan) in 150 mM Citrate Buffer, pH 4.5] for 75 min at 37ºC. The digested meristems were then pipetted into a petri dish with chilled distilled water and incubated on ice for 15 min to wash out the enzymes. Using a stereomicroscope, the root epidermis was removed, and the protoplasts were gently pipetted into a 1.5 ml tube containing 40 µl chilled Carnoy's fixative. The protoplasts were then suspended by gently vortexing the tube for about 30 sec at room temperature, after which the sample was centrifuged at 4,000 ×*g* for ~3 min and the pellet was resuspended in acetic acid-ethanol (9:1) solution. Finally, the protoplast suspension was pipette-mounted onto ethanol-

cleaned glass slides, which were placed in a humid chamber to facilitate spreading of the chromosomes, and then allowed to dry.

Probe preparation

PCR amplicons from 18S rDNA and 5S rDNA were obtained using primers previously designed (Koo *et al.* 2002; Matoba *et al.* 2007). Primers flanking the entire Pg167TR locus in BAC PgH005J07 (KF357942) were designed using the online tool Primer3 (Rozen and Skaletsky 1998). Another set of primers was designed from the internal region of the Pg167TR repeat unit to amplify Pg167TR from genomic DNA (gDNA) template. Fluorochromes were directly conjugated to the DNA amplicons (Table 1-1) through direct nick translation labeling. Pg167TR and *B. oleracea* 5S rDNA amplicons were labeled with Texas Red-5-dUTP (Perkin Elmer, NEL417001EA), and 18S rDNA with Alexa Fluor® 488-5-dUTP (Invitrogen, C11397).

Table 1-1. List of primers used to amplify *P. ginseng* satDNAs.

Target	Forward	Reverse	Modified nucleotide	Ref.
18S	AACCTGGTTGATCCTGCCAGT	TGATCCTTCTGCAGGTTCACCTAC	Alexa Fluor 488-5-dUTP	Matoba et al., 2007
5S_F	GATCCCATCAGAACTCC	GGTGCTTTAGTGCTGGTAT	Texas Red-5-dUTP	Koo et al., 2002
Pg167TR_bac	ATTTGAGTTTGTATTCTTCAAGTTAGGTG	AACTGGACACAAAGATCCATGTTATTC	n/a	This study
Pg167TR_genome	GAGGCGGGTTTTGACCTATT	CCACGCAAACACACACGTA	n/a	This study

Hybridization and image analysis

FISH procedures were as described previously (Waminal *et al.* 2012). Briefly, slides were immediately used for FISH after fixation with 4% paraformaldehyde, without pepsin and RNase pretreatment. The hybridization mixture contained 50% formamide, 10% dextran sulfate, 2× SSC, 5 ng μl^{-1} salmon sperm DNA and 20 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. The mixture was denatured at 90ºC for 10

min and immediately kept on ice for at least 5 min prior to mounting on slides. After covering with a glass coverslip, the chromosomes were denatured at 80ºC for 3-5 min on a hot plate. The slides were then immediately transferred into a humid chamber preset at 37ºC and incubated overnight (~16 hr). The following day, the slides were washed in 2× SSC (15 min at RT), 0.1× SSC (35 min at 42ºC), and finally 2× SSC (30 min at RT).

Images were captured with an Olympus BX53 fluorescence microscope equipped with a Leica DFC365 FS CCD camera, and processed using Cytovision ver. 7.2 (Leica Microsystems, Germany). Further image enhancements and creation of the idiogram were performed in Adobe Photoshop CC.

Chromosome numbering and pairing

Chromosome number assignment was based on the decreasing order of chromosome lengths, while homologous chromosome pairing was achieved according to the centromeric position (Levan *et al.* 1964), DAPI band and rDNA loci distribution. Chromosomes were grouped according to the number of DAPI bands in each arm. As demonstrated by Costa Silva *et al.* (2011), the estimated DNA content in each chromosome was calculated by distributing the 1C DNA content of *P. ginseng* (3.12×10^3 Mb, (Hong *et al.* 2004)) relative to the length of each chromosome.

Restriction digest

The Pg167TR array from BAC PgH005J07 was searched *in silico* using CLC Main Workbench ver. 7.6.4 (Qiagen, Denmark) for restriction sites appearing at regular intervals among repeat units. PCR products were digested with *Taq*I (Thermo Fisher Scientific Inc., ER0671, USA) following the manufacturer's instructions. High-resolution fragment analysis was done using a Fragment Analyzer™ (Advanced Analytical Technologies, Inc., USA).

RESULTS

Chromosome complement composition and rDNA localization

P. ginseng chromosome complement was confirmed to be 2*n* = 48 (Fig. 2-1). With reference to the centromere position (i.e. arm ratio), the complement comprised 12 metacentric (1-7, 11-13, 15, and 18), 9 submetacentric (8-10, 16-17, 19, and 22-24), and 3 subtelocentric (14 and 20-21) homologous chromosome pairs with a karyotype formula of 12m+9sm+3st. The chromosome lengths ranged from 3.27 to 6.30 μm (Table 1-2).

Only one pair of satellite chromosomes (pair 14) was observed, and the only locus of 45S rDNA in the genome was localized at the secondary constriction of this subtelocentric chromosome (Fig. 2-1). Moreover, one major locus of 5S rDNA signal at the intercalary region of the short arm of chromosome 11 and two minor loci at the short arms of chromosome 2 and 22 were detected. The major 5S rDNA locus was flanked by two DAPI bands.

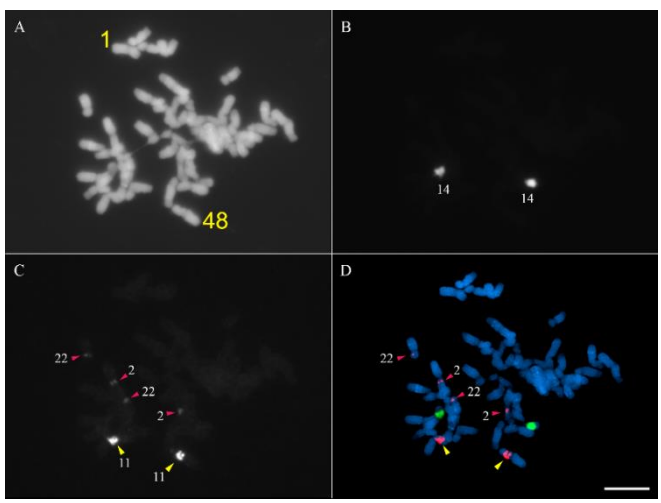


Fig. 2-1. Minor 5S rDNA loci. Two previously unidentified (Waminal et al. 2012) 5S rDNA loci were observed using direct-labeled probes. These minor loci were localized in the proximal region of 2S and 22S. Panels A–D: raw DAPI, raw 45S rDNA, raw 5S rDNA, and merged signals, respectively. Bar = 10 μm.

Table 1-2. Summary of *P. ginseng* chromosome features.

Chr. no.	Short arm (p)	Long arm (q)	Total	Arm ratio (q/p)	Type	RDC (Mb)[a]
1	3.16 ± 0.12	3.17 ± 0.11	6.3 ± 0.22	1.002	m	196.77
2[b]	2.64 ± 0.08	3.27 ± 0.12	6.05 ± 0.06	1.237	m	188.98
3	2.52 ± 0.23	3.61 ± 0.11	5.88 ± 0.18	1.434	m	183.72
4	2.54 ± 0.2	3.27 ± 0.22	5.64 ± 0.07	1.289	m	176.29
5	2.23 ± 0.27	3.35 ± 0.12	5.41 ± 0.17	1.506	m	168.89
6	2.05 ± 0.11	3.3 ± 0.06	5.31 ± 0.23	1.609	m	165.71
7	2.09 ± 0.07	3.35 ± 0.22	5.3 ± 0.14	1.605	m	165.45
8	1.54 ± 0.32	3.66 ± 0.13	5.23 ± 0.4	2.378	sm	163.34
9	1.52 ± 0.19	3.82 ± 0.13	5.08 ± 0.21	2.515	sm	158.69
10	1.77 ± 0.04	3.49 ± 0.06	5.04 ± 0.28	1.965	sm	157.45
11[b]	2.13 ± 0.12	2.91 ± 0.13	4.94 ± 0.12	1.363	m	154.25
12	1.96 ± 0.07	3.03 ± 0.12	4.83 ± 0.28	1.547	m	150.88
13	2.04 ± 0.05	3.05 ± 0.04	4.82 ± 0.07	1.492	m	150.41
14[c]	1.99c ± 0.21	3.21 ± 0.14	4.8 ± 0.31	1.612[d]	st	149.95
15	2.26 ± 0.17	2.58 ± 0.28	4.73 ± 0.49	1.143	m	147.64
16	1.55 ± 0.09	3.33 ± 0.1	4.72 ± 0.08	2.157	sm	147.33
17	1.59 ± 0.15	3.05 ± 0.07	4.5 ± 0.11	1.919	sm	140.50
18	2.09 ± 0.25	2.54 ± 0.19	4.5 ± 0.06	1.214	m	140.46
19	1.39 ± 0.12	2.78 ± 0.17	4.11 ± 0.21	1.998	sm	128.28
20	1.05 ± 0.04	3.24 ± 0.07	4.09 ± 0.06	3.067	st	127.67
21	0.9 ± 0.05	3.02 ± 0.21	3.8 ± 0.13	3.355	st	118.70
22[b]	1.32 ± 0.06	2.32 ± 0.1	3.56 ± 0.09	1.761	sm	111.13
23	1.25 ± 0.11	2.3 ± 0.22	3.38 ± 0.09	1.836	sm	105.52
24	1.13 ± 0.25	2.08 ± 0.24	3.27 ± 0.1	1.84	sm	101.98

[a] Relative DNA content based on 3.6 Gb genome size. [b] 5S rDNA, [c] 45S rDNA, [d] value obtained using satellite instead of short arm, m: metacentric, sm: submetacentric, st: subtelocentric (Levan et al. 1964).

PCR and sequence analysis of Pg167TR satDNA

Repeat unit analysis of the Pg167TR locus in BAC PgH005J07 (nt 10,048~11,650) with Tandem Repeat Finder (Benson 1999) revealed 9.6 repeat units in the array (Fig. 2-2A) (Choi *et al.* 2014). Sequence alignment of the nine repeat units showed high GC content at the nucleotide positions from 80 to 100 (Fig. 2-2B), and *in silico* restriction digestion revealed *Taq*I restriction sites at regular intervals, which would result in fragments ranging from 149 to 169 bp (Fig. 2-3A, C and D). PCR with primers flanking Pg167TR in BAC PgH005J07 amplified the main 1,682-bp target with additional ladder-pattern amplicons, and a smear pattern of high molecular weight amplicons. The same primer set with gDNA template amplified a similar pattern as found with the BAC template (data not shown), indicating that such amplification of this Pg167TR locus is independent of PCR template. In addition, PCR with primers designed from the internal region of the Pg167TR unit and using gDNA template to amplify other units in the genome also amplified a smear pattern from about 150 bp to much longer fragments. This smear pattern in both BAC and gDNA indicates that there are multiple annealing sites, as is generally the case for repetitive elements (Fig. 2-3B). Further analysis may help us understand the nature, origin, and impact of Pg167TR in the *P. ginseng* genome, i.e. whether it is associated with other TEs as in the case of other satDNAs (Sharma *et al.* 2013; Dias *et al.* 2014).

Restriction digestion of both BAC and gDNA amplicons with *Taq*I enzyme supported the *in silico* prediction of restriction size fragments, indicating that the amplicons are Pg167TR-associated (Fig. 2-3A-D).

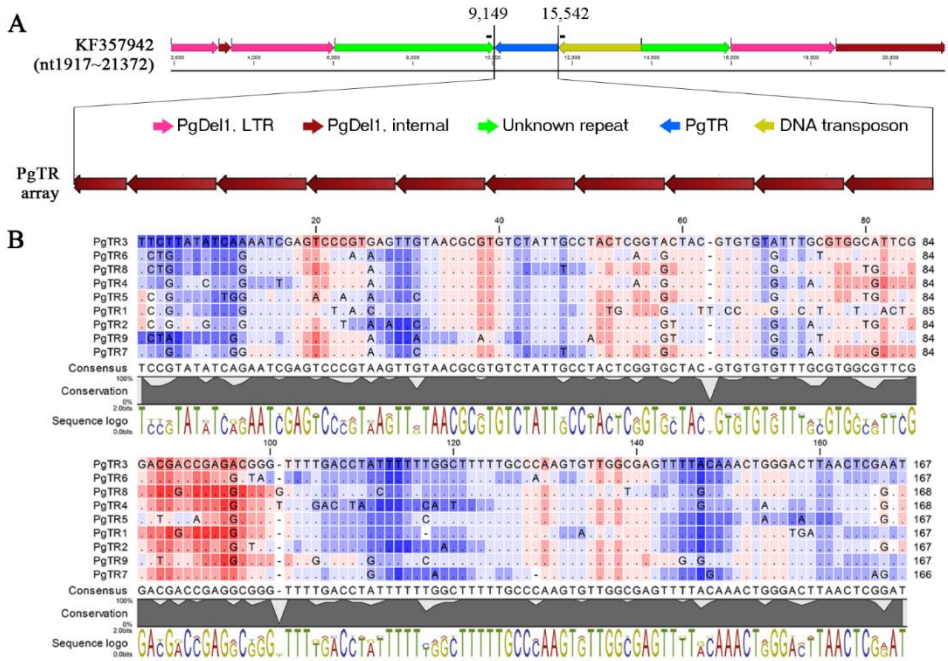


Fig. 2-2. Sequence characterization of Pg167TR repeats identified in BAC H005J07 (Pg167TR_KF357942) shows heterochromatin-related features. A) Portion of BAC H005J07 showing the location of 9.6 Pg167TR units flanked by DNA transposon and unknown sequences reported previously (Choi et al. 2014). B) Multiple sequence alignment of the nine complete Pg167TR units in BAC H005J07 showing regions of high GC content in red shade.

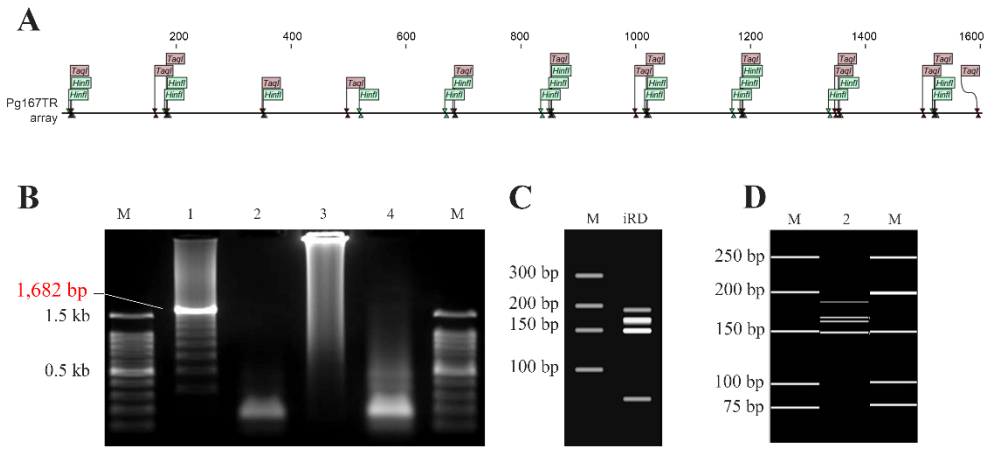


Fig. 2-3. Pg167TR sequence characterization. A) *In silico* mapping of *Taq*I (blush red) and *Hinf*I (green) restriction sites in the Pg167TR array in BAC PgH005J07 showing regular intervals of both restriction sites. B) PCR amplification of Pg167TR. Lane 1: Amplification of the 1,682-bp Pg167TR array from BAC PgH005J07 using BAC-derived primer sequences. Additional bands showing a ladder-like pattern were obtained from partial annealing primer, 3: Another set of Pg167TR primers were designed from internal regions and, with gDNA as template, revealed an expected smear pattern, 2 and 4: Restriction enzyme digestion with *Taq*I revealed several <200-bp fragments corresponding to Pg167TR unit lengths. C) *In silico* restriction enzyme digestion (iRD) of Pg167TR array revealed a similar pattern with the gel analysis in panel B. D) Partially *Taq*I-digested amplicons from panel B were analyzed with higher-resolution Fragment Analyzer and expected results from iRD were obtained.

Identification of homologous chromosomes

DAPI band distribution

Numerous DAPI-binding heterochromatic regions were dispersed along all chromosomes and were visible as DAPI dots. These dots, similar to those in chromosomes 5 and 8, did not form distinct DAPI bands. Both the DAPI dots and bands were made more easily visible by inverting the images (Fig. 2-4). In addition to the rDNA loci, the presence of several observable DAPI bands along the chromosome complement made identification of homologous pairs possible. A total of 38 bands were identified (Fig. 2-4).

Pg167TR probe as an efficient cytogenetic marker

With Pg167TR as a cytogenetic marker, all 24 pairs of homologous chromosomes were identified successfully (Fig. 2-5A and B, Table 1-3). While certain chromosomes required more detailed analysis, some signals were readily distinguishable. For instance, chromosome 5 was easily distinguished by the presence of three very intense signals, two of which were localized respectively in the pericentromeric and intercalary regions of the long arm (5L), while the third was located in the subtelomeric region of the short arm (5S). Another example was chromosome 7, identifiable by a large block of Pg167TR loci at the intercalary region of 7L. Although some chromosomes (e.g., 22–24) bore similar Pg167TR signal patterns, they could be distinguished via other characteristics, such as chromosome length, DAPI bands, and centromeric indices. This Pg167TR-based method thus enabled identification of individual chromosomes from different chromosome spreads (Fig. 2-6A-C). A karyotypic idiogram of the *P. ginseng* chromosomes showing the DAPI bands, rDNA and Pg167TR signals is shown in Fig. 2-7.

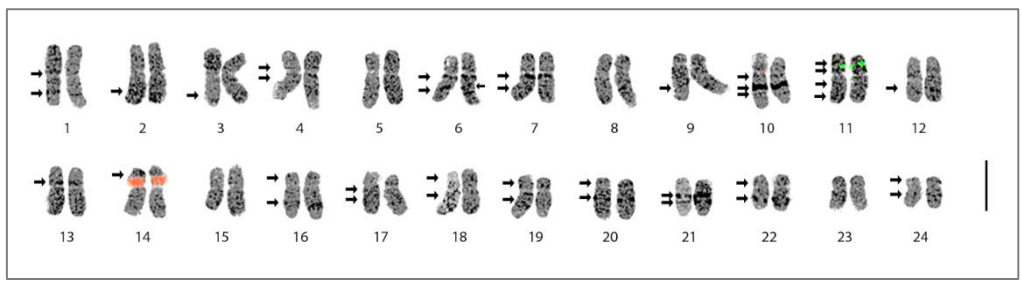


Fig. 2-4. DAPI band distribution. Arrows depict the DAPI bands along ginseng chromosomes.

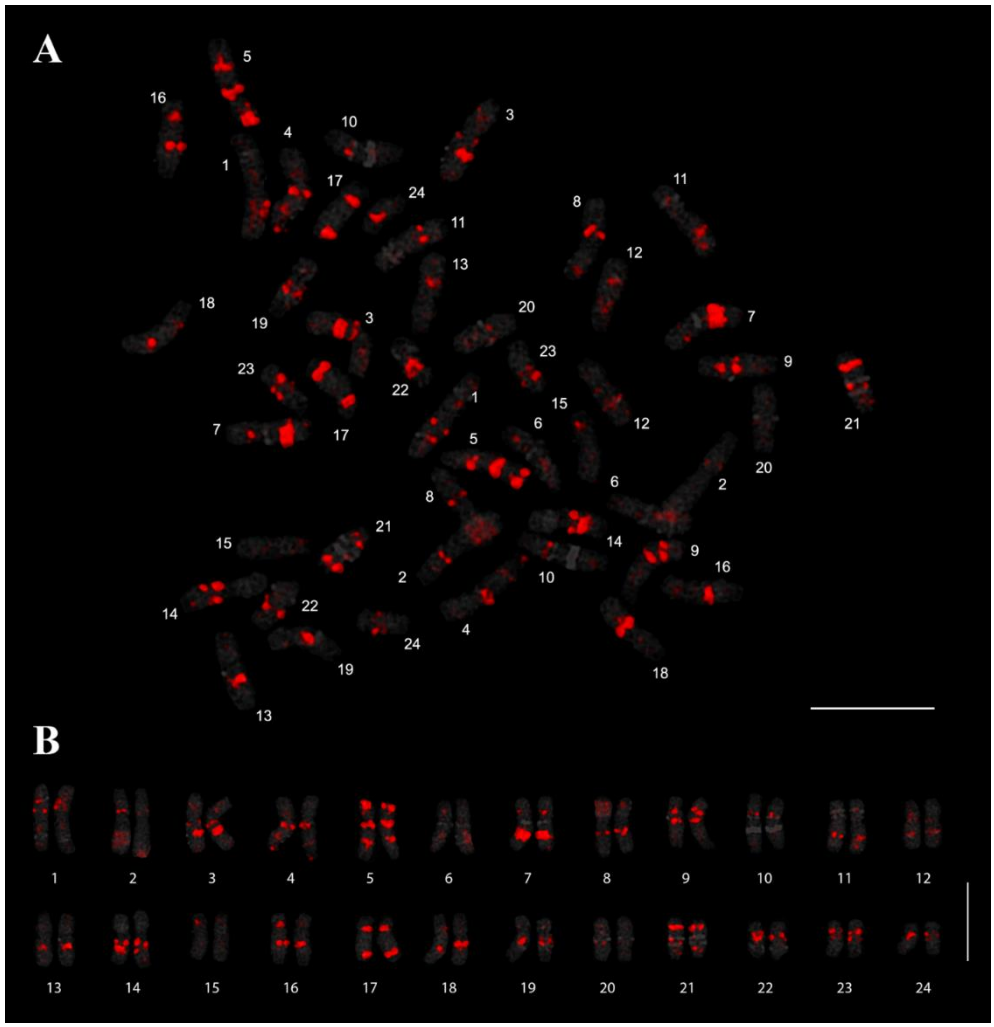


Fig. 2-5. Genomic distribution of Pg167TR and homologous chromosome identification. A) FISH analysis—using the BAC-amplified 1,862-bp Pg167TR PCR product as a probe—shows distinct chromosome distribution and abundance, allowing for easy identification. B) Karyogram of *P. ginseng* using only Pg167TR signals and the recognizable DAPI bands. This allows refinement of the previously reported *P. ginseng* karyotype. Bars = 10 μm.

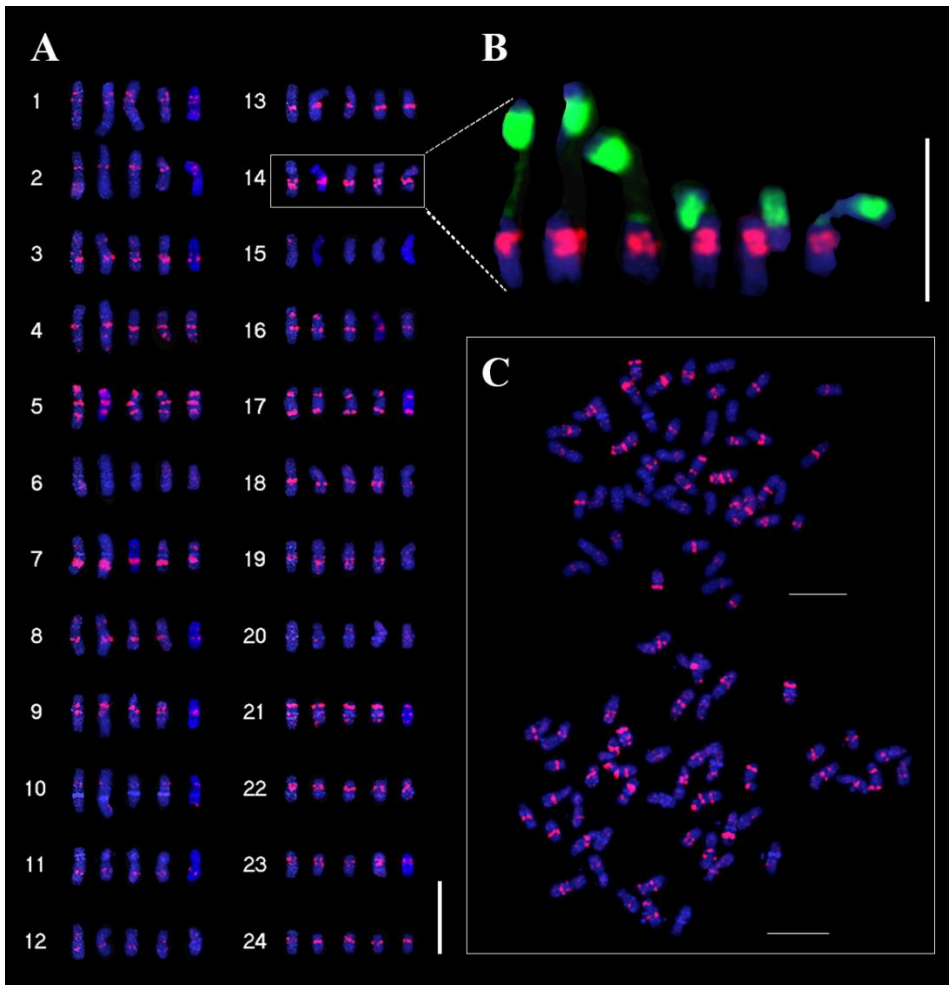


Fig. 2-6. Cytogenetic mapping of Pg167TRs. A) Identification of ginseng chromosomes from several root mitotic chromosome spreads using Pg167TR, B) Consistent rDNA and Pg167TR patterns for chromosome 14. C) Two mitotic chromosome spreads with Pg167TR signals. Bars = 10 μm.

Table 1-3. Summary of DAPI, rDNA, and Pg167TR distribution in 24 *Panax ginseng* chromosomes.

Chr. no.	Chromosome features
1	Paracentromeric DAPI band on short arm (S), more intense intercalary on long arm (L). Two medium-intense Pg167TR loci on 1S, one being paracentromeric and the other intercalary. Another two weak Pg167TR loci on 1L flanking the 1L DAPI band.
2	Weak subtelomeric DAPI band on 2L. One paracentromeric Pg167TR locus on 2S. Minor 5S rDNA locus on paracentromeric region of 2S, proximal to the Pg167TR locus.
3	Subtelomeric and average intensity DAPI band on 3L. Large and intense Pg167TR on intercalary 3L, with small and weak loci on paracentromeric 3L and intercalary 3S.
4	Three Pg167TR loci on 4L, pericentromeric, intercalary (about 50% from centromere and telomere), and telomeric, with the pericentromeric signal being the most intense.
5	This chromosome is easily distinguishable owing to its intense Pg167TR signals localized in the pericentromeric and intercalary regions of 5L and subtelomeric region of 5S. Another weak Pg167TR signal can be seen just proximal to the 5S intense signal.

6	One intense paracentromeric DAPI band, two intercalary with weaker proximal. Weak Pg167TR signal proximal to the intercalary DAPI band on 6L. Another two weak Pg167TR signals on 6S, one at the intercalary region, and one at the paracentromeric area.
7	Average intensity DAPI bands on 7L. This is easily distinguishable for its large and intense Pg167TR signal at the intercalary region of 7L, one of the most intense signals in the genome. Additional Pg167TR signal is localized at the intercalary region of 7S.
8	Intercalary 8L Pg167TR signal.
9	Weak intercalary DAPI band in 9L. Pg167TR signals localized at the paracentromeric regions of 9S and 9L.
10	Weak paracentromeric DAPI band, two intercalary with very intense middle and weak distal on 10L and weak at intercalary on 10S. Only one Pg167TR signal at the centromeric region.
11	Two moderate intensity flanking 5S rDNA on 11S, one weak subtelomeric on 11L. 5S rDNA moderate intensity One intercalary Pg167TR signal proximal to the subtelomeric DAPI band on 11L.
12	Intercalary moderate intensity DAPI band on 12L. Intercalary 12L DAPI band. PgDel2-rich, concentrated at the centromeric up to the intercalary regions of 12L. Weak paracentromeric 12S and intercalary 12L Pg167TR signals.

13 Weak paracentromeric DAPI band on 13S. One intercalary Pg167TR signal on 13L, proximal to the DAPI band.

14 Weak DAPI band on the satellite chromosome. Two intense Pg167TR loci can be found in the intercalary region of 14L, which sometimes overlap and can be seen as one large signal in some spreads.

15 Weak intercalary Pg167TR signal on 15S.

16 Weak subtelomeric DAPI band on 16S, more intense intercalary on 16L. One intense intercalary Pg167TR signal proximal to the DAPI band on 16L, another weak signal on 16L.

17 Weak paracentromeric and weak intercalary DAPI bands. Intense Pg167TR signals that correspond to the DAPI bands on both arms. This chromosome is easily distinguishable owing to the intense Pg167TR signals.

18 Weak intercalary on 18S, weak paracentromeric on 18L. Intercalary 18L and weak paracentromeric 18S Pg167TR signals.

19 Intercalary DAPI bands on both arms, more intense on 19L. Two closely localized intercalary 19L Pg167TR signals that sometimes overlap to be seen as one signal. Another weak intercalary Pg167TR signal seems to colocalize with the 19S DAPI band.

20 Intercalary DAPI bands on both arms, more intense on 20L. One weak 20L Pg167TR signal just distal to the 20L DAPI band.

21	Two intercalary DAPI bands on 21L, proximal more intense than distal. Intense centromeric Pg167TR locus is observable, plus two other weaker loci on 21L, one in between the DAPI bands, and one at the subtelomeric area.
22	Intercalary DAPI bands on both arms, more intense on 22L, 22L signal more intense than that on chromosome 20L. Two Pg167TR loci, one at the pericentromeric region of each arm, that can overlap to be seen as one signal. Minor 5S rDNA locus at the paracentromeric region of 22S.
23	Two Pg167TR loci, one at the pericentromeric region of each arm, that can overlap to be seen as one signal. This looks similar to chromosome 22 except for the DAPI band that is absent here.
24	Weak subtelomeric DAPI band on 24S, more intense intercalary on 24L. Two closely localized Pg167TR at the paracentromeric area of 24L.

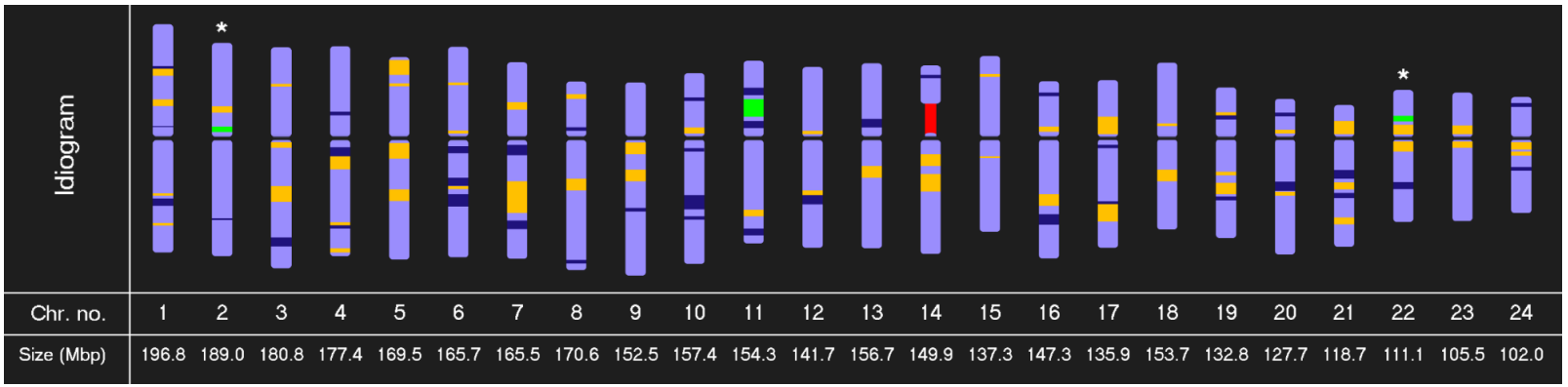


Fig. 1-7. FISH idiogram of *P. ginseng*. 5S and 45S rDNA are indicated by green and red bars, respectively. DAPI bands are indicated by dark blue bars. PgTR signals are indicated by orange bars. Asterisks indicate chromosomes with minor 5S rDNA signals detected. Chromosome sizes are based on the 3.6 Gb genome size of ginseng.

DISCUSSION

Available cytogenetic information for ginseng is still currently limited despite the increasing interest in genomic research. Consequently, there are no established cytogenetic markers for the identification of homologous chromosomes. This lack of data has limited our understanding of the karyotype of ginseng and its concomitant phylogenetic relationship with other related species in and out of the genus *Panax*. In this study, I exploited the usefulness of major satDNA that are highly abundant in ginseng as well as the inherent DAPI bands in ginseng chromosomes as molecular cytogenetic markers in pairing homologous chromosomes and establishing the first molecular cytogenetic-based karyotype of *P. ginseng.*

Pg167TR, rDNAs and DAPI bands are reliable cytogenetic markers

The 5S and 45S rDNA, which are highly conserved among eukaryotes, were only localized in four chromosomes. Only one locus of 45S rDNA was detected, which is in agreement with the results reported by Choi *et al.* (2009). However, although these authors also detected one locus for 5S rDNA, three loci were detected in this study. The disparity may be attributed to the sensitivity and efficiency of the current FISH protocol which allowed detection of two additional minor 5S signals in chromosomes 2 and 22 that might not have been detected in the previous report.

Intense and reliable DAPI bands enabled furhter identification of a few more chromosomes, such as chromosomes 6, 10, and 20. With only the rDNA and DAPI bands used, many other chromosomes remained difficult to categorize due to lack of cytogenetic markers. However, cytogenetic mapping of Pg167TR amplicon from BACH05J07 produced unique signal patterns in each chromosome, thus providing an efficient cytogenetic marker, that, together with rDNA, DAPI, and chromosome arm ratio, enabled the establishment of the first ginseng molecular cytogenetic based karyotype. This is of particular importance considering the fact that not only has the ginseng genome been organized into high number of chromosomes but also most

chromosomes are challenging to identify owing to a fairly small and uniform size of *P. ginseng* chromosomes. A similar success in identifying short and homogenous chromosome size was achieved in soybean using highly abundant repetitive elements (Findley *et al.* 2010).

Efficient identification of homologous chromosomes provides a more robust cytogenetic map of *P. ginseng* for future integration with genetic linkage maps and pseudo-chromosomes from genome assembly scaffolds. It also provides a platform for comparative cytogenetics among species within and without the genus *Panax*. This comparative study will allow a cytogenetic view of the history of *P. ginseng* and related genomes.

A major satDNA in *P. ginseng* with features that resemble those of Pg167TR was previously reported (Ho and Leung 2002). Although these authors observed a 170-bp repeat unit (compared with up to 168 bp for Pg167TR), validation could not be carried out because precise sequence information was not available (personal communication). Nevertheless, they reported that this repeat is one of the most abundant satDNA repeats in *P. ginseng* and showed a gel ladder pattern after digestion with *Hinf*I that was similar to our *Taq*I digestion results. To explore this similarity further, we carried out *in silico* restriction analysis of Pg167TR, which revealed *Hinf*I restriction sites near the *Taq*I sites (Fig 1-3A), suggesting that the satDNA identified in the previous report is most likely Pg167TR.

Ginseng karyotype and tetraploidy

Karyotype data are essential in understanding the phylogenetic relationship among species belonging to the same family (Heslop-Harrison and Schwarzacher 2011; Mendes *et al.* 2011a), making them useful to cyto-taxonomic studies (Pinto *et al.* 2012). Additionally, comparative cytogenetics provide knowledge the relationship between related diploid and polyploid species (Kovarik et al. 2005, Leflon et al. 2006, Snowdon 2007, Wang et al. 2007, Kolano et al. 2008, Xiong and Pires 2011).

Most species belonging to the family Araliaceae have chromosomes of $2n$ = 24 or $2n$ = 48, except for a few genera that have little chromosomal number variation (Yi et al. 2004). In a review of the chromosomal evolution of the family Araliaceae, Yi et al. (2004) discussed that, although the actual basic chromosome number of the family was thought to be x = 12, some species were observed to be $2n$ = 36. Considering the family's basic chromosome number, these species would be regarded triploids, but triploids are often genetically unstable. This challenged the establishment of the basic chromosome number of the family. The x = 12 hypothesis was further challenged after the genus *Hydrocotyle* which has several taxa with $2n$ = 18, 36, and 60 were moved into Araliaceae from Apiaceae, giving an alternative basic chromosome number x = 9 and x = 6. Nevertheless, x = 12 is generally accepted as the basic chromosome number in the family, but this does not eliminate the possible ancestral x = 6 (Yi et al. 2004), especially when considering the recent discovery that all angiosperms underwent rounds of genome duplications and diploidization (Tank *et al.* 2015). Recently, Choi *et al.* (2011) showed the high replication of homologous genes in ginseng using SSR markers and suggested that the polyploidy could range from tetra- to octoploidy. Nevertheless, in practice, *Panax ginseng* is regarded as a tetraploid species with a basic chromosome number of 12 (Wen and Zimmer 1996, Court 2000, Yi et al. 2004, Choi et al. 2009).

Our data showed a somatic cell chromosome complement of $2n$ = 48, supporting previously reported chromosome numbers (Ko et al. 1993, Choi et al. 2009) and polyploidy (Wen and Zimmer 1996, Court 2000, Yi et al. 2004, Choi et al. 2009). Although homologous chromosome pairs were easily identified, homeologous chromosome pairs were not, perhaps owing to the extensive genome reshuffling in ginseng and its ancient allopolyploidization event (Choi *et al.* 2014). On the contrary, homeologues of recent allopolyploids like *Tragopogon mirus* could easily be distinguished by simple molecular cytogenetic analysis.

The reduction of rDNA loci may be explained by the non-additive nature of rDNA loci and other genomic DNA segments after polyploidization (Snowdon et al.

1997, Ozkan et al. 2003, Yoshikazu et al. 2006). Many polyploid species do not reflect a correlation between the rDNA loci number and the level of ploidy; in fact, polyploids can even have half the number of rDNA signals than their diploid counterparts (Yoshikazu et al. 2006). This rDNA reduction phenomenon has been well-documented in the *Artemisia* species (Pellicer et al. 2010).

Summary

The chromosomal distribution patterns of DAPI bands, rDNA, and Pg167TR, along with the chromosome arm ratio, were efficient in establishing the first FISH-based karyotype of ginseng. The discovery of Pg167TR reveals its considerable abundance in the *P. ginseng* genome. Importantly, Pg167TR distribution enabled identification of each *P. ginseng* chromosome despite their relatively uniform lengths, demonstrating its utility in *P. ginseng* karyotyping, which will promote future integration of *P. ginseng* genetic maps, the validation of pseudo-chromosomes from the *P. ginseng* genome assembly, and comparative analysis with related species to elucidate the evolutionary history of the *P. ginseng* genome.

REFERENCES

Albert PS, Gao Z, Danilova TV, and Birchler JA. 2010. Diversity of chromosomal karyotypes in maize and its relatives. Cytogenet Genome Res 129: 6-16.

Ananiev EV, Phillips RL, and Rines HW. 1998. Complex structure of knob DNA on maize chromosome 9: Retrotransposon invasion into heterochromatin. Genetics 149: 2025-2037.

Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res 27: 573-80.

Choi H-I, Kim N-H, Lee J, Choi B, Kim K, Park J, *et al.* 2013. Evolutionary relationship of panax ginseng and p. Quinquefolius inferred from sequencing and comparative analysis of expressed sequence tags. Genetic Resources and Crop Evolution: 1-11.

Choi HI, Kim NH, Kim JH, Choi BS, Ahn IO, Lee JS, *et al.* 2011. Development of reproducible est-derived ssr markers and assessment of genetic diversity in panax ginseng cultivars and related species. Journal of Ginseng Research 35: 399-412.

Choi HI, Waminal NE, Park HM, Kim NH, Choi BS, Park M, *et al.* 2014. Major repeat components covering one-third of the ginseng (panax ginseng c.A. Meyer) genome and evidence for allotetraploidy. Plant J 77: 906-916.

Choi HW, Choi DH, Bang KH, Paek KY, Seong NS, and Bak JW. 2009. Fish and gish analysis of the genomic relationships among *panax* species. Genes & Genomics 31: 99-105.

Coluccia E, Pichiri G, Nieddu M, Coni P, Manconi S, Deiana AM, *et al.* 2011. Identification of two new repetitive elements and chromosomal mapping of repetitive DNA sequences in the fish gymnothorax unicolor (anguilliformes: Muraenidae). Eur J Histochem 55: e12.

Costa Silva S, Marques A, Santos Soares Filho W, Mirkov TE, Pedrosa-Harand A, and Guerra M. 2011. The cytogenetic map of the *poncirus trifoliata* (l.) raf.—a nomenclature system for chromosomes of all citric species. Tropical Plant Biology 4: 99-105.

Court WE. 2000. Ginseng: The genus *panax*, Amsterdam, Netherlands, Hardwood Academic Publishers.

Dias GB, Svartman M, Delprat A, Ruiz A, and Kuhn GC. 2014. Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of drosophila virilis. Genome Biol Evol 6: 1302-13.

Findley SD, Cannon S, Varala K, Du J, Ma J, Hudson ME, *et al.* 2010. A fluorescence in situ hybridization system for karyotyping soybean. Genetics 185: 727-44.

Heslop-Harrison JS, and Schwarzacher T. 2011. Organisation of the plant genome in chromosomes. The Plant Journal 66: 18-33.

Ho IS, and Leung FC. 2002. Isolation and characterization of repetitive DNA sequences from *panax ginseng*. Moleclar Genetics and Genomics 266: 951-61.

Hong C, Lee S, Park J, Plaha P, Park Y, Lee Y, *et al.* 2004. Construction of a bac library of korean ginseng and initial analysis of bac-end sequences. Molecular Genetics and Genomics 271: 709-716.

Kato A, Lamb JC, and Birchler JA. 2004. Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. Proceedings of the National Academy of Sciences 101: 13554-9.

Kim N-H, Lee S-C, Choi H-I, Kim K, Choi B-S, Jang W, *et al.* Genome sequences and evolution of *panax ginseng*. *In:* KIM, S.-K. & RHEE, D.-K., eds. Evidence-based Efficacy of Ginseng 2014, Proceedings of the 11th International symposium on Ginseng, 2014 Seoul, Korea. The Korean Society of Ginseng, 125-146.

Koo DH, Hong CP, Batley J, Chung YS, Edwards D, Bang JW, *et al.* 2011. Rapid divergence of repetitive dnas in brassica relatives. Genomics 97: 173-85.

Koo DH, Hur Y, Jin DC, and Bang JW. 2002. Karyotype analysis of a korean cucumber cultivar (cucumis sativus l. Cv. Winter long) using c-banding and bicolor fluorescence in situ hybridization. Mol Cells 13: 413-8.

Leung KW, and Wong AS. 2010. Pharmacology of ginsenosides: A literature review. Chinese Medicine 5: 20.

Levan A, Fredga K, and Sandberg AA. 1964. Nomenclature for centromeric position on chromosomes. Hereditas 52: 201-220.

Lion T, and Haas OA. 1990. Nonradioactive labeling of probe with digoxigenin by polymerase chain reaction. Analytical biochemistry 188: 335-337.

Macas J, Koblížková A, Navrátilová A, and Neumann P. 2009. Hypervariable 3′ utr region of plant ltr-retrotransposons as a source of novel satellite repeats. Gene.

Matoba H, Mizutani T, Nagano K, Hoshi Y, and Uchiyama H. 2007. Chromosomal study of lettuce and its allied species (lactuca spp., asteraceae) by means of karyotype analysis and fluorescence in situ hybridization. Hereditas 144: 235-243.

Matyasek R, Gazdova B, Fajkus J, and Bezdek M. 1997. Ntrs, a new family of highly repetitive dnas specific for the t1 chromosome of tobacco. Chromosoma 106: 369-79.

Mehrotra S, and Goyal V. 2014. Repetitive sequences in plant nuclear DNA: Types, distribution, evolution and function. Genomics, Proteomics & Bioinformatics 12: 164-171.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, *et al.* 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14: R10.

Mendes MM, Da Rosa R, Giuliano-Caetano L, and Dias AL. 2011a. Karyotype diversity of four species of the *incertae sedis* group (characidae) from different hydrographic basins: Analysis of agnors, cma3 and 18s rdna. Genetics and Molecular Research 10: 3596-3608.

Mendes S, Moraes AP, Mirkov TE, and Pedrosa-Harand A. 2011b. Chromosome homeologies and high variation in heterochromatin distribution between citrus l. And poncirus raf. As evidenced by comparative cytogenetic mapping. Chromosome Research 19: 521-530.

Michael TP, and Jackson S. 2013. The first 50 plant genomes. Plant Genome 6.

Mondin M, Santos-Serejo JA, Bertao MR, Laborda P, Pizzaia D, and Aguiar-Perecin ML. 2014. Karyotype variability in tropical maize sister inbred lines and hybrids compared with kys standard line. Front Plant Sci 5: 544.

Palomeque T, and Lorite P. 2008. Satellite DNA in insects: A review. Heredity 100: 564-573.

Park HJ, Kim DH, Park SJ, Kim JM, and Ryu JH. 2012. Ginseng in traditional herbal prescriptions. J Ginseng Res 36: 225-41.

Pinto MMPD, Calixto MD, De Souza MJ, De Araujo APT, Langguth A, and Santos N. 2012. Cytotaxonomy of the subgenus *artibeus* (phyllostomidae, chiroptera) by characterization of species-specific markers. Comparative Cytogenetics 6: 17-28.

Pita M, Orellana J, Martinez-Rodriguez P, Martinez-Ramirez A, Fernandez-Calvin B, and Bella JL. 2014. Fish methods in cytogenetic studies. Methods Mol Biol 1094: 109-35.

Plohl M, Luchetti A, Mestrovic N, and Mantovani B. 2008. Satellite dnas between selfishness and functionality: Structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409: 72-82.

Rigby PW, Dieckmann M, Rhodes C, and Berg P. 1977. Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase i. Journal of molecular biology 113: 237-251.

Rozen S, and Skaletsky HJ. 1998. Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Sharma A, Wolfgruber TK, and Presting GG. 2013. Tandem repeats derived from centromeric retrotransposons. BMC Genomics 14: 142.

Smith DR. 1993. Random primed labeling of DNA. Methods Mol Biol 18: 445-7.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, *et al.* 2015. Nested radiations and the pulse of angiosperm diversification: Increased diversification rates often follow whole genome duplications. New Phytol 207: 454-67.

Waminal N, Park HM, Ryu KB, Kim JH, Yang TJ, and Kim HH. 2012. Karyotype analysis of panax ginseng c.A.Meyer, 1843 (araliaceae) based on rdna loci and dapi band distribution. Comparative Cytogenetics 6: 425-441.

Waminal NE, Perumal S, Lee J, Kim HH, and Yang T-J. 2016. Repeat evolution in brassica rapa (aa), b. Oleracea (cc), and b. Napus (aacc) genomes. Plant Breeding and Biotechnology 4: 107-122.

Xiong Z, and Pires JC. 2011. Karyotype and identification of all homoeologous chromosomes of allopolyploid brassica napus and its diploid progenitors. Genetics 187: 37-49.

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

CHAPTER II

Characterization of a high-copy *CACTA* transposon with unevenly amplified Pg167TR tandem repeats in *Panax ginseng*

ABSTRACT

Genome size variations among angiosperms have been attributed to differential abundance of transposable elements (TEs) and satellite DNAs (satDNAs). TEs have been associated with satDNA formation and amplification. The differential amplification of a few TE or satDNA families in genomes of related species provide information not only about the structure of a genome, but also its history and a potential function of particular TE or satDNA families in a particular genome. In a previous report, we have shown the high abundance of Pg167TR, a 167-bp high-copy satDNA from *Panax ginseng*, through cytogenetic mapping. Here, I identified a ginseng *CACTA* transposon family (*PgCACTA1*) harboring Pg167TR sequences. Genome wide analysis revealed that the *PgCACTA1* is high copy with large size variation derived from amplification of Pg167TR. Further analysis was conducted to understand the evolution of *PgCACTA1* and Pg167TR sequences and its impact for ginseng genome via *in silico* and FISH analyses. Variable Pg167TR copy number was localized at the last intron of a transposase-related gene at the 3' distal region of *PgCACTA1* ranging from about three to more than 1,000 causing expansion of the Pg167TR locus in some elements. Two Pg167TR sequence variants, Pg167TRa and Pg167TRb with >90% and ~86% similarity within and between groups were identified. A polymorphic region at nt 100~112 highlights the difference between the two variants. *PgCACTA1* elements carrying either only one Pg167TR variant or both were identified. Pg167TRa was associated with highly expanded Pg167TR loci, while Pg167TRb was more limited to loci with fewer copies, but a biased homogenizing force seemed to favor amplification of Pg167TRa. Fluorescence *in situ* hybridization data supported the amplification of Pg167TR into long satellite arrays as well as the differential abundance between Pg167TRa and Pg167TRb. Pg167TRa was three times more abundant than Pg167TRb in species within the genus *Panax* but showed an inverse but highly reduced total abundance in related species outside the genus. Pg167TR showed a relatively balanced nucleotide content

but with an uneven AT distribution along the sequence increasing sequence curvature propensity associated with heterochromatin coiling. The presence of *cis*-regulatory element motifs in Pg167TR sequences suggests underlying regulatory functions. This study showed an association of Pg167TR with *PgCACTA1* elements, and demonstrated how *PgCACTA1* spurred amplification of Pg167TR to chromosomally distinct long satDNA arrays. This also provides a platform for further functional studies on *PgCACTA1* to understand its role in the *P. ginseng* genome.

Keywords: *Panax ginseng*, *CACTA*, Pg167TR, *PgCACTA1*, FISH, tandem repeat, karyotype

INTRODUCTION

Genome size variation among angiosperms are mainly attributed to the differential amplification or contraction of different repetitive element (REs) families (Macas *et al.* 2009; Michael and Jackson 2013). REs cover different types of repeats based on their repeat unit organization as either dispersed or tandem repeats (TRs) (Kubis *et al.* 1998; Choi *et al.* 2014). Dispersed repeats include transposable elements (TEs), which are distributed throughout a genome or subgenome or specific chromosomal regions (Lim *et al.* 2007; Choi *et al.* 2014). TRs, on the other hand, are organized in a head-to-tail fashion in distinct chromosomal regions (Coluccia *et al.* 2011; Sharma *et al.* 2013). They include microsatellites (1-5 bp monomers), minisatellites (6-100 bp monomers), and satellites (mostly 150~400 bp, but can reach up to thousands of bp monomers) (Macas *et al.* 2009; Mehrotra and Goyal 2014).

Satellite DNAs (satDNAs) can extend up to 100-Mb array sizes and are the most abundant class among these three. They include ribosomal DNA (rDNA) families, centromeric and subtelomeric TRs, and heterochromatin 'knobs;" (Ananiev *et al.* 1998; Palomeque and Lorite 2008; Plohl *et al.* 2008; Koo *et al.* 2011; Melters *et al.* 2013; Sharma *et al.* 2013; Mehrotra and Goyal 2014; Mondin *et al.* 2014). Consequently, they occupy a substantial proportion in large plant genomes (Kelly *et al.* 2015), but also in plants with small genomes. For instance, about 24% and 8% of the fat duckweed (*Lemna gibba*) and *Arabadopsis thaliana* genomes, respectively, consist of centromeric satellite DNA alone (Melters *et al.* 2013).

TEs and satDNAs are evolving rapidly but follow different modes of evolution (Bennetzen and Wang 2014; Kelly *et al.* 2015). TEs code their own protein for transposition and amplification, i.e. reverse transcriptase with RNaseH and integrase for retrotransposons, and transposase for DNA transposons (Bennetzen and Wang 2014), while satDNAs do not. However, various mechanisms, such as unequal cross over, rolling circle replication, and gene conversion are supposed to have spurred evolution of satDNAs (Dover 1982; Walsh 1987; Charlesworth *et al.* 1994; Cohen

et al. 2003; Jo *et al.* 2009). Because of the different mechanisms involved in TEs and satDNAs amplification, they have mostly been analyzed independently (Macas *et al.* 2009). However, sequence similarity of some satDNAs to TE segments (Cheng and Murata 2003; Macas *et al.* 2009; Sharma *et al.* 2013; Dias *et al.* 2014), suggest an evolutionary association of satDNAs to TEs, often derived from Class I TEs, mostly of the long terminal repeat (LTR) order (Macas *et al.* 2009; Sharma *et al.* 2013), only few have been described from Class II or DNA transposons (Nagaki *et al.* 1998; Wicker *et al.* 2003; Dias *et al.* 2014).

Distinct chromosomal distribution is an important characteristic of satDNAs, allowing them to be utilized as efficient cytogenetic markers in identifying homologous and homeologous chromosomes; thus, facilitating karyotyping and comparative cytogenetics among closely related taxa (Matyasek *et al.* 1997; Albert *et al.* 2010; Mendes *et al.* 2011; Xiong and Pires 2011).

Ginseng (*Panax ginseng* C. A. Meyer) is a perennial herb highly valued for its ginsenosides, which are reputed to have a wide range of medicinal effects (Court 2000; Leung and Wong 2010; Park *et al.* 2012). It is the most widely studied species in the genus *Panax*, but most studies have largely focused on the plant's pharmacological effects (Leung and Wong 2010; Park *et al.* 2012). Nevertheless, we have seen a growing interest in ginseng molecular cytogenetic and genomic studies in recent years (Waminal *et al.* 2012; Choi *et al.* 2014; Jayakodi *et al.* 2014; Kim *et al.* 2014; Kim *et al.* 2016; Waminal *et al.* 2016a).

In a previous report, we have shown the high abundance and cytogenetic marker efficiency of Pg167TR, a 167-bp high-copy satDNA from *Panax ginseng* (Waminal *et al.* 2016a). In this work, we conducted *in silico* and FISH analyses to characterize the sequence and evolution of Pg167TR. Our results suggest that long Pg167TR arrays that are chromosomally distinct evolved from a resident Pg167TR of a ginseng *CACTA* element, *PgCACTA1*, through expansion of this locus most likely driven by homology-dependent unequal crossovers. This information broadens our understanding of the Pg167TR evolution and provide an evolutionary

pathway of plant satDNA expansion from a Class II DNA transposon, particularly from a *CACTA* element. This furthers our understanding of the ginseng genome structure and should be a platform for further functional studies of *PgCACTA* and Pg167TR in understanding their roles in gene regulation, taking into consideration reports showing the involvement of *CACTA* transposons and satDNAs in gene regulation (He *et al*. 2000; Ugarkovic 2005; Zabala and Vodkin 2007; Alix *et al.* 2008; Nosaka *et al*. 2013; Fambrini *et al*. 2014; Zabala and Vodkin 2014).

MATERIALS AND METHODS

Pg167TR classification and whole-genome sequence (WGS) read mapping

Repeat units from Pg167TR array, a 1,603-bp region in BAC PgH005J07 (KF357942, Fig. 1A) (Choi *et al.* 2014) were identified using Tandem Repeats Finder (Benson 1999) and used as a query for genomic blast against *P. ginseng* assembly scaffolds to extract representative sequences for genomic Pg167TR classification. Top 200 elements with 95% sequence length coverage and 80% identity along with Pg167TR units from identified *PgCACTA1* elements (a total of 344 sequences) were grouped for classification. Multiple sequence alignments, sequence comparisons, and k-mer phylogenetic tree construction were carried out using CLC Main Workbench ver. 7.6.4 (Qiagen, Denmark). Whole genome sequence (WGS) reads for *P. ginseng* cv. Chunpoong and other related species were generated using the Illumina HiSeq 2000 system in the National Instrumentation Center for Environmental Management at Seoul National University (Table 2-1). Randomly extracted WGS reads from each species were mapped on ginseng Pg167TRa and Pg167TRb consensus sequences with CLC Assembly Cell ver. 4.21 (http://www.clcbio.com/products/clc-assembly-cell/), using default parameters.

Identification of highly abundant contigs

The top 30 highly abundant WGS assembly contigs were identified using a CLC Assembly Cell ver. 4.21 following the genome-skimming approach called *de novo* assembly of low-coverage WGS (dnaLCW) method (Kim *et al.* 2015). *De novo* assembly of 0.74x haploid genome-equivalent WGS data of *P. ginseng* (Chunpoong) from the quality filtered reads by the CLC-quality trim tool were then assembled by a CLC genome assembler (ver. 4.06, CLC Inc, Rarhus, Denmark) with parameters of 200 to 600 bp autonomously controlled overlap size. Genomic abundance in terms

of average read depth (RD) along with the length of the contig (LC) were done using clc-reference assembly approach. Top 30 high-depth contigs were retrieved based on the high genome representation (RD x LC). The top 30 high-depth contigs were then annotated by BLASTn (best hit) against the Plant Repeat Database (Ouyang and Buell 2004) and a custom database with previously reported REs of *P. ginseng* (Choi *et al.* 2014) and classified as known repeats if contigs shares 80% homology and 80% sequence alignment (Wicker *et al.* 2007).

Table 2-1. Comparative WGS mapping of Pg167TR variants among *Panax ginseng* samples.

Symbol	Sample	WGS (Mb)	x[a]
cp	Chunpoong	2662.14	0.74
cr	Cheonryang	2489.72	0.69
cs	Cheongsun	1950.77	0.54
go	Gopoong	1557.42	0.43
gu	Gumpoong	1234.14	0.34
sh	Sunhyang	1623.51	0.45
so	Sunwon	1526.30	0.42
sp	Sunpoong	1584.15	0.44
su	Sunun	1734.36	0.48
yp	Yunpoong	1739.50	0.48
hs	Hwangsook	1821.13	0.51
jk	Jakyung	1361.62	0.38
hamyang	wild ginseng-hamyang	2749.61	0.76
kw2	wild ginseng-kangwon-2	5746.83	1.60
kw8	wild ginseng-kangwon-8	5761.35	1.60
kw13	wild ginseng-kangwon-13	5871.86	1.63
kw16	wild ginseng-kangwon-16	5837.51	1.62
kw17	wild ginseng-kangwon-17	5791.93	1.61
kw18	wild ginseng-kangwon-18	5532.79	1.54

[a] Equivalent genome coverage of extracted WGS reads used for analysis

Table 2-2. Comparative WGS mapping of Pg167TR variants among *Panax ginseng* related species.

Species	**2*n***	**Ploidy**	**Genome size (Mb)**	**wgs size (Mbp)**	**x[a]**
P. ginseng (cp)	48	4x	3,600	2,662	0.74
P. quinquefolius	48	4x	4,914	1,183	0.24
P. notoginseng	24	2x	2,454	1,989	0.81
P. vietnamensis	24	2x	2,018	3,182	1.58
P. japonicus	24	2x	2,000	1,998	1.00
Acanthopanax divaricatus var. *albeofructus*	48	4x	2,200	1,718	0.78
A. senticosus for. *inermis*	48	4x	2,200	1,698	0.77
A. koreanum	48	4x	2,200	1,211	0.55
A. sessiliflorus	48	4x	2,200	1,544	0.70
Aralia elata	24	2x	1,500	3,936	2.62
Dendropanax morbifera	48	4x	--	2,144	--
Kalopanax septemlobus	48	4x	1,700	1,904	1.12

[a] Equivalent genome coverage of extracted WGS reads used for analysis

PCR amplification of Pg167TR sequence variants and oligoprobe design

Primers were designed from consensus sequences using the online tool Primer3 (Rozen and Skaletsky 1998) to amplify Pg167TRa and Pg167TRb sequence variants (Table 2-2). Additionally, fluorochrome-conjugated Pg167TR sequence variants were designed and synthesized by Bioneer (Seoul, Korea).

Table 2-3. List of primers and oligoprobes (OP) used in this study.

Name	Sequence	Mer	Modification
Pg167TRa_unit_f	GATATACAGAATCCGAGTTATCG	23	-
Pg167TRa_unit_r	ATCAAGTCCCGTATGTTCTAA	21	-
Pg167TRb_unit_f	GATATATGGTATCCAAGTTAAGT	23	-
Pg167TRb_unit_r	ATCAAGTCCCGTATGTTGTAA	21	-
Pg167TRa_OP	ATCGCCCAGTTTGCAAAACTCGC CAACAC	29	5'-FAM
Pg167TRb_OP	AAGTCCCAGTTTGTAAAACTCGC CAACAC	29	5'-Cy3

Homology search for *cis*-regulatory elements and curvature propensity analysis

Analysis for putative c*is*-acting regulatory elements in Pg167TR elements were carried out by homology search with c*is*-acting regulatory elements identified in other plant satellite DNAs using Pg167TR units from BACH05J07 representing both Pg167TRa and Pg167TRb (Ugarkovic 2005; Pezer *et al.* 2011; Brajkovic *et al.* 2012; Sharma *et al.* 2013) via the PlantCARE (Lescot *et al.* 2002) and PLACE databases (Higo *et al.* 1999). A curvature propensity analysis was carried out using the bend.it and model.it tools (Vlahovicek *et al.* 2003).

Sequence characterization and genome quantification of *PgCACTA1*

Unknown repeat regions flanking the Pg167TR array in BAC PgH005J07 were analyzed for TE domains through Conserved Domain Database (www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). A putative transposase domain

was identified and manual curation of regions flanking the transposase domain and Pg167TR array revealed terminal inverted repeats (TIRs), target site duplications (TSDs), and abundant subterminal repeat at both ends, leading to the identification of putative *CACTA* elements. The GC/AT content of the *PgCACTA1* flanking regions designating insertion sites were analyzed using the Artemis Comparison Tool (Carver *et al.* 2005). Coding sequences (CDS) were predicted using Fgenesh gene finder set to dicot *Zea mays* parameters after initial analysis revealed a closer relationship of *PgCACTA1* to *Z. mays En/Spm* than other related dicot species (data not shown) (Solovyev *et al.* 2006), and transposase domains were identified through the NCBI Conserved Domain Database (Marchler-Bauer *et al.* 2015). Phylogenetic analysis of TnpD- and TnpA-like transposases were done using predicted protein sequences from Fgenesh using ClustalW alignment and tree builder in CLC Main Workbench ver. 7.7.2 (Qiagen, Denmark). Reference mapping of WGS reads and estimation of *PgCACTA1* copy number were carried out with CLC Assembly Cell ver. 4.21 and Microsoft Excel 2016. Mean WGS read mapping coverage of conserved domains were normalized to account the genome size, and values of TnpA domain were used to estimate active *PgCACTA1* transposons in the ginseng genome.

Transcriptome read mapping

Raw reads from transcriptome sequencing of leaves from a one-year old ginseng plant were generated using Illumina HiSeq 2000 system in the National Instrumentation Center for Environmental Management at Seoul National University. Quality trimming and read pre-processing was carried out using CLC Assembly Cell ver. 4.21. Transcriptome raw reads were mapped with default parameters.

Fluorescence *in situ* hybridization (FISH) analysis

Root mitotic chromosome spreads were prepared from stratified seeds provided by the Korea Ginseng Corporation Natural Resources Research Institute (Daejeon, Korea), according to the methods of Waminal *et al.* (2012). BAC PgH005J07 single-

locus Pg167TR amplicons were labeled with Texas Red-5-dUTP (Perkin Elmer, NEL417001EA) through direct-labeling nick translation. Oligoprobes for each Pg167TR variant were synthesized by Macrogen (South Korea, http://dna.macrogen.com) with fluorochrome modification. To check co-hybridization of a *PgCACTA1* transposase region with BAC PgH005J07 single-locus Pg167TR, a 2,975-bp transposase domain was PCR amplified and the product was labeled with Alexa Fluor488-5-dUTP (Invitrogen, C11397). FISH procedures were done according to Waminal *et al.* (2012). The Pg167TR was labeled with Texas Red-5-dUTP (Perkin Elmer, NEL417001EA). For directly labeled probes and oligoprobes, slides were immediately used for FISH after fixation with 4% paraformaldehyde, without subsequent pepsin and RNase pretreatment. Images were captured with an Olympus BX53 fluorescence microscope equipped with a Leica DFC365 FS CCD camera, and processed using Cytovision ver. 7.2 (Leica Microsystems, Germany). We performed further image enhancements and FISH-based estimation of Pg167TR genomic proportion in Adobe Photoshop CC.

RESULTS

Identification of variable *CACTA* transposons harboring variable copy number of Pg167TR

Sequence characterization of regions flanking the Pg167TR locus in BACH05J07 was carried out to characterize the heterochromatin features of surrounding DNA. A putative transposase domain upstream the Pg167TR locus was identified. Further manual sequence characterization allowed the identification of a 6,394-bp *En/Spm*-like (*CACTA*) element with a 29-bp TIR that starts with a 'CACTA' motif, a 'TAA' 3-bp TSD flanking the element and subterminal repeats (STR) in both ends (Fig. 3-1). The 3' STR region was longer (~400 bp) than that of the 5' (~300 bp).

A whole-genome survey was carried out to identify other members of the *PgCACTA1* family. Several elements were identified with sizes ranging from about 6 kb to >280 kb with various TSD sequences. The 'TAA' sequence was most frequently observed while some elements did not have conserved and identifiable TSD (Table 2-4). The TIR sequences ranged from 7 to 99 bp with a mode of 31 bp, and sequence homology between the left and right TIR ranging from 64 to 100% and a mean homology of 80%. One common feature of these elements was the presence of Pg167TR repeat units at the 3' region of all these elements but in variable copy number ranging from three to about 1,500 copies (Table2-6). Some elements do have some N-gaps at the 5' regions, but mostly at the Pg167TR region.

A putative autonomous element, *PgCACTA1_1058*, which has a 31-bp and 84% homologous flanking TIR, was shown to have two adjacent genes with domains matching transposase_21 and transposase_24 (Fig. 3-2). Additionally, its 5' and 3' flanking regions showed considerably higher AT content (~90%) than the 40% AT content of its internal region (Fig. 3-2) and 65% of the whole ginseng genome (unpublished) , a common insertion preference of *CACTA* elements (Alix *et al.* 2008;

Gao *et al.* 2016). Blast with *PgCACTA1_1058* against the ginseng genome assembly also revealed a high sequence conservation of the transposase_21-like gene, which is about 95% homologous with other *PgCACTA1* elements, compared with the transposase_24-like gene (Fig. 3-2). Likewise, WGS read mapping revealed a more abundant copy of the transposase_21-like gene (~750 copies) compared with the transposase_24-like gene (~460 copies). However, transcriptome read mapping showed a more abundant number of transcripts mapped on the transposase_24-like gene (~20x) compared to those on transposase_21-like gene (~5x), suggesting a generally higher expression of transposase_24-like gene (Fig. 3-2).

The Pg167TR satDNA was observed to be localized at the last intron of the transposase_24-like gene which was localized downstream of the transposase_21-like gene towards the 3' end of the *PgCACTA1* element. The variability of Pg167TR copy number at the 3' end of the *PgCACTA1* elements caused extensive sequence expansion in some elements (Fig. 3-3). Comparative FISH data between diploid (*P. notoginseng*) and tetraploid (*P. ginseng*) *Panax* species supports the *in silico* prediction of Pg167TR amplification from *PgCACTA1* elements (Fig.2-5, 2-8).

Table 2-4. List of CACTA elements characterized in this analysis.

No.	Source	S. start	S. end	Strand	SIZE	Flank	TSD	Left TIR sequence	Length (bp)	TIR homology	Pg167TR units	Poly-N
1	BAC_H005J07	15542	9149	-	6394	Left	TAA	CACTACAAAGAAATGGGTCATTGGCGACG	29	0.83	9.7	-
						Right	TAA	CACTACAAGAAAAAGTGTCATTGCCGACG	29			
2	Pg_scaffold1058	689160	700123	+	10964	Left	TTA	CACTACAAGGAAATTGGTCATTGGCGACGGG	31	0.84	11.3	-
						Right	TTA	CACTACAAGAAAAAGTGTCATTGCCGACGGG	31			
3	Pg_scaffold1156	653694	639247	-	14448	Left	AAA	CACTACAAGGAAATTGGTCATTGGCGACGGG	31	0.84	9.3	-
						Right	AGA	CACTACAAGAAAAAGTGTCATTGCCGACGGG	31			
4	Pg_scaffold3595	167850	161464	-	6387	Left	TAA	CACTACAAAGAAATGGGTCATTGGCGACG	29	0.83	9.7	-
						Right	TAA	CACTACAAGAAAAAGTGTCATTGCCGACG	29			
5	Pg_scaffold5380	13373	19790	+	6418	Left	AAG	CACTACAAGGAAATTGGTCATTGGCGACGGG	31	0.81	5.3	yes
						Right	AAG	CACTACAAGAAAAAGTGTCATTGCCAACGGG	31			
6	Pg_scaffold1695	206966	222748	+	15783	Left	TAC	GTACTACAAAAAATAGGTCATTGGCGACGGC	31	0.87	43.4	-
						Right	TAC	CACTACAAAAAAACAGCTCATTGTCGACGGC	31			
7	Pg_scaffold0720	617354	630314	+	12961	Left	ATT	CACTACAAAAAATTAGGTCATT	22	0.77	14.5	yes
						Right	ATT	CACTACAAGAAAATCAGTCATT	22			
8	Pg_scaffold1087	148817	162129	+	13313	Left	CCC	CACTACAAAAAAAAAGGGTCATTGCTGACGG	31	0.69	5.2	yes
						Right	CCC	CATTAAAAGAGAACCAGTCATTGCCGATGAG	31			
9	Pg_scaffold0568	807447	820826	+	13380	Left	---	---	---	---	7.6	yes
						Right	---	---	---			
10	Pg_scaffold0628	852265	865939	+	13675	Left	AAA	CACTACAAAAAAGGGTCATTCACGACGGGGTCATTAGCGTCGGCGAA	47	0.64	36.5	-
						Right	---	AAGAGTGTCATTACCGAAGTGGTCATTGGCGACGGCTAA	39			
11	Pg_scaffold0502	625943	636400	+	10458	Left	AAA	CACTACA	7	1	14.6	-
						Right	AAA	CACTACA	7			
12	Pg_scaffold0145	1454956	1464481	+	9526	Left	AAT	CACTACAAAAAAGAGGGTCATTGCCGA	27	0.82	3.4	-
						Right	AAT	CACTACAAGAAAACCAGTCATTGCCGA	27			
13	Pg_scaffold3809	59119	230084	+	170966	Left	AGA	CACTACAAGGAAATTGGTCATTGGCGACGG	30	0.73	894.6	yes
						Right	AGA	CACTACAAAAAAAACCTATCATTACCGACGG	30			
14	Pg_scaffold1025	110161	120310	+	10150	Left	ATA	CACTACAAGGAAAAGGGTCATTGACGACGGGC	32	0.81	16.3	-
						Right	ATA	CACTACAAGAAAACCAGTCATTGCCGACAGGC	32			
15	Pg_scaffold1261	234767	247623	+	12857	Left	ATC	CACTACAAAAAAACAGGTCATTGGCGACGG	30	0.83	23.9	-
						Right	ATC	CACTACAAGAAAATCAGTCATTGCCGACGG	30			

16	Pg_scaffold0638	403656	513190	+	109535	Left	ATT	CACTACAAAAAAATAGGTCATTGGCGACAGC	31	INC	629.1	yes
						Right	---	---	---			
17	Pg_scaffold1200	30681	40000	+	9320	Left	CTT	CACTACAAGAAAAAGGGTCATTGGTGACGGGC	32	0.81	7.7	-
						Right	CTT	CACTACAAGAAAATCAGTCATTGCCAACGGGC	32			
18	Pg_scaffold1807	24772	73873	+	49102	Left	GAA	CACTACAAAAATAAGGGTCAATGGCGACGGGCACTCTGCCG	41	0.85	121.1	yes
						Right	GAA	CACTACAAAAAAAGGGTCATTGCCGACGGCCTCTGCCG	38			
19	Pg_scaffold0233	36230	47448	+	11219	Left	GTT	CACTACAAGGAAATTGGTCATTGGAGACAGG	31	0.77	12.7	-
						Right	GTT	CACTACAAAAAAAGGTCATTGCCGACGGG	29			
20	Pg_scaffold0018	925063	1206529	+	281467	Left	TAA	CACTACAAAAAAATTGGTCATTGCCGACGG	30	0.77	1491.7	yes
						Right	TAA	CACTACAAGAAAACCAATCATTGGCGATGGG	31			
21	Pg_scaffold1760	235926	246804	+	10879	Left	TAA	CACTACAAGGAAATTGGTCATTGGCGACGG	30	0.81	19.7	yes
						Right	TAA	CACTACAAAAAAATAAGGTCATTGCCGACGG	30			
22	Pg_scaffold0991	640752	653140	+	12389	Left	TAT	CACTACAAGGAAATTGGTCATTGGCGACG	29	0.83	13.3	yes
						Right	TAT	CACTACAAGAAAAAGTGTCATTGCCGACG	29			
23	Pg_scaffold0820	383122	390270	+	7149	Left	TCA	CACTACAAAAAAACATGTCATAGGCGACGGG	31	0.81	7.6	-
						Right	TCA	CACTACAAGAAAACCAGTCATTACCGACGGG	31			
24	Pg_scaffold0796	7467	17201	+	9735	Left	TTA	CACTACAAGAAAAAGGTTCATTGGCGACGGGAGTATTGGCGACGACCACTTTACCGTCGATAATAACGTCACTGTCGACGGGGATGCCGTCGCC	94	0.66	9.4	-
						Right	TTA	CACTACAAGAAAACCAGCCATTGCCAACAGGCACTGCCGTCGCCAATGACCATAGTGCCGTCGACAGTGAGGTCATTGGCGATGGGCCCTGCCGTCGCC	99			
25	Pg_scaffold1130	595774	601676	+	5903	Left	TTC	CACTACAAAAAAGTAGGTCATTGGCGACGGCA	32	0.88	4.3	yes
						Right	TTC	CACTACCAAAAAATAAGGTCATTGCCGACGGCA	33			
26	Pg_scaffold1471	13668	71905	+	58238	Left	TTC	CACTACAAGAAAATAGGTCATTGGCGACGGCGTCATTAGCGACGGTAGTATCCCCGTCGCCAATAATATCATTCTCGACGGCATAGCCGTCGCC	94	0.7	295.6	-
						Right	TTC	CACTACAAAAAAATAGCTCATTGCGGACGACCAATGCCGTCGCCAATGACCCTCTTGCCGTCGCCAATGATGTCATTGGCGACGGCACTGTCCGTCGCC	99			
27	Pg_scaffold1178	144917	156883	+	11967	Left	TTT	CACTATAAGGAAACTGGTCATTGGCAACGGGC	32	0.81	17	-
						Right	TTT	CACTACAAGAAAACCAGTCATTGCCGACGGGC	32			
28	Pg_scaffold1087	547303	532845	-	14457	Left	TAA	CACTACAAAAAAAACAGTTCATTGGCGACGGCATCATTAGCGACGGTAG	49	0.74	37.7	yes
						Right	TAA	ACTACAAAAAAACAGGTCATTGTCGTCGGCAATGAGGGCATTGGCGACGGCAG	53			

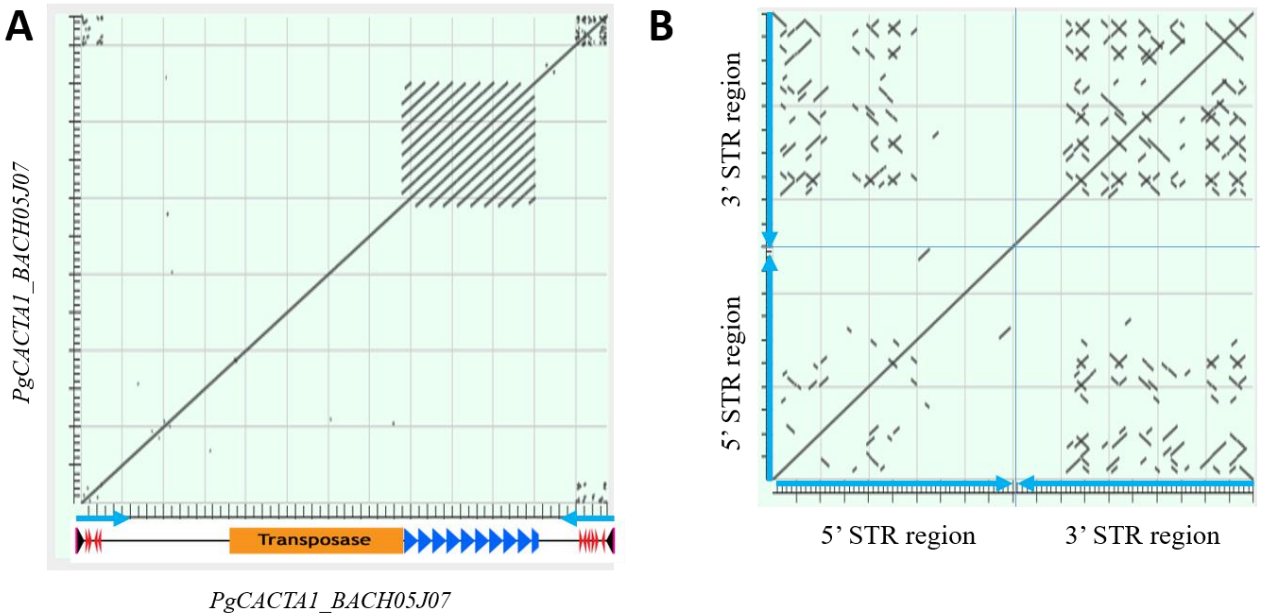


Fig. 3-1. Sequence characterization of *PgCACTA1* element from BACH05J07. A) Dot plot of *PgCACTA1_BACH0J07* bound by conserved TIRs and TSDs showing Pg167TR (blue arrowheads) region at the 3' region, subterminal repeats (pink arrowheads) at both ends, and a transposase domain upstream of the Pg167TR locus B) Dot plot of a 500-bp region from both ends including the TIR regions show the more abundance of STRs at the 3' than the 5'.

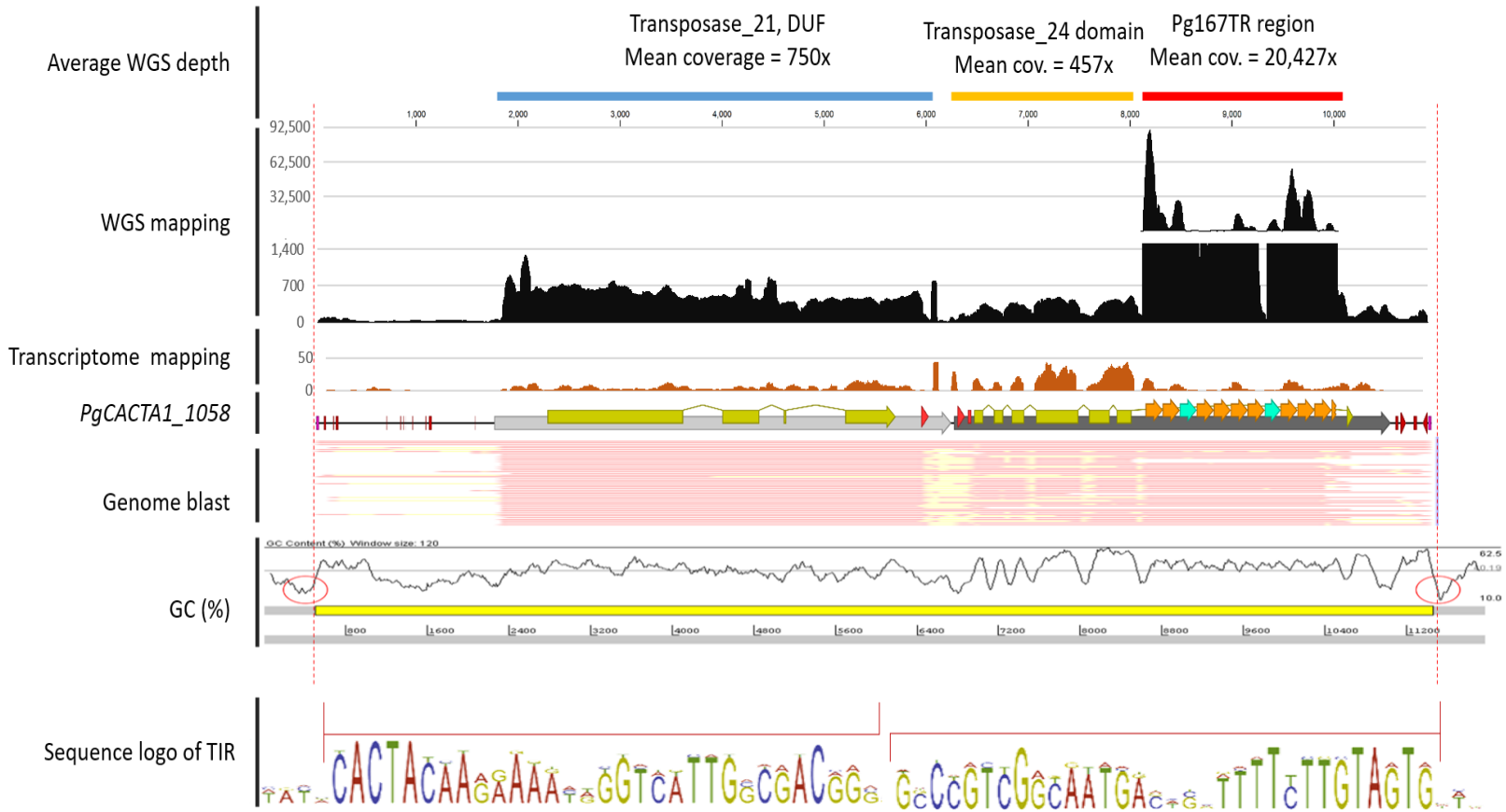

Average WGS depth
Transposase_21, DUF
Mean coverage = 750x
Transposase_24 domain
Mean cov. = 457x
Pg167TR region
Mean cov. = 20,427x
WGS mapping
92,500
62,500
32,500
1,400
700
0
Transcriptome mapping
50
0
PgCACTA1_1058
Genome blast
GC (%)
Sequence logo of TIR

Fig. 3-2. Characterization of a full length autonomous *PgCACTA1_1058* element showing the Pg167TR locus at the 3’ end.

WGS read mapping, CDS and repeat annotation, diagram of genomic blast hits, and GC composition and insertion site of *PgCACTA1_1058* identified from flanking regions of the Pg167TR in BACH05J07. Two genes (light and dark grey arrows on *PgCACTA_1058* sequence) were predicted using *Z. mays* model in Fgenesh gene prediction. The upstream gene matched with transposase_21 superfamily and domain unknown function (DUF), while the downstream gene matched with transposase_24 superfamily. The upstream gene showed a relatively higher genome coverage, but the CDS of the downstream gene showed higher mapping of transcriptome reads. Pg167TR sequences (Blue and orange arrows, Pg167TRa and Pg167TRb, respectively) were inserted at the last intron of the downstream gene; hence at the 3’ region of the element. Insertion sites of *PgCACTA1_1058* showed high AT composition relative to its internal regions. The 31-bp TIR region of several *PgCACTA1* elements are shown.

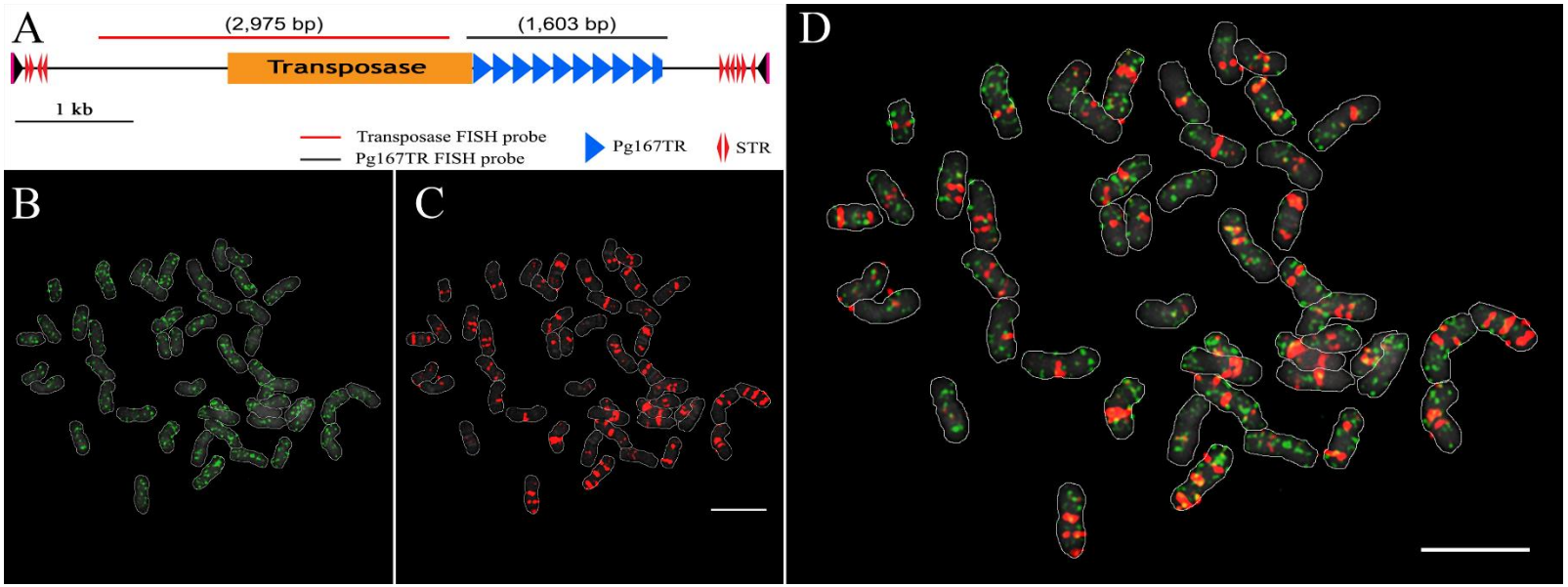

A
(2,975 bp)
(1,603 bp)
Transposase
1 kb
Transposase FISH probe
Pg167TR FISH probe
Pg167TR
STR
B
C
D

Fig. 3-3. Dual-color FISH with transposase and Pg167TR array shows highly amplified Pg167TR loci. A) Diagram of *PgCACTA1_BACH05J07* element showing the transposase, Pg167TR, and STR regions. Bars above the sequence indicate PCR products and corresponding sizes used as FISH probes, B) Transposase domain FISH signals, C) PgTR array signals., and D) merged signals from B and C, showing chromosomal regions with highly amplified PgTR loci. The highly amplified Pg167TR chromosomal loci from the *PgCACTA1* element are observable. Bar = 10 μm.

Autonomous *PgCACTA1* codes for two putative transposases

In silico prediction of transposase coding genes in *PgCACTA1* elements was carried out to predict their transposition autonomy. A common pattern of two adjacent genes were observed in several elements, (Fig. 3-10), like those identified in *PgCACTA_1058* (Fig. 3-2). However, elements with a single gene but still carry both predicted gene products like those in *PgCACTA1_1058* were observed (*PgCACTA1_0502* and *PgCACTA1_0502*). Both gene products (TnpD and TnpA in tobacco) are necessary for autonomous transposition (Masson *et al.* 1991). Domain search with the NCBI CDD showed different transposase family in either gene. Although the exon number of TnpD-like gene varied from four to nine with seven being more common, all genes showed specific hits to transposon associated domain (pfam13963), and superfamily hits to Transposase_21 (pfam02992) and to domains of unknown function, DUF416 and DUF418 superfamilies (pfam13952 and pfam13960, respectively). On the other hand, there were six to eight exons for the TnpA-like gene which showed high homology to a Transposase_24 (pfam03004).

However, some elements, like *PgCACTA1_0991*, *PgCACTA1_0502*, and *PgCACTA1_1471*, have their TnpD-like and TnpA-like CDS fused into one single gene similar to the bicistronic gene observed in tobacco (Masson *et al.* 1991). In addition, some elements carry internal deletions, such as *PgCACTA1_3595*, disrupting proper transposase gene sequence. We predict these truncated sequences to be non-autonomous elements.

A pattern of 75-bp direct repeat (TCCTTTTAGTTATATTCAAATGTAA TAGTTGTCTCCTTTTAGTTATATCTGCATGTTGTATCCTTTTAGTTGTAT) in between the two genes was often observed, and even though some elements have fused genes, this pattern often indicated the boundary between the genes (Fig. 3-10).

Phylogenetic analysis of TnpD- and TnpA-like transposase protein sequences from different *PgCACTA1* elements obtained from *in silico* prediction, classified the two transposase into two distinct groups. TnpD-like transposases in ginseng were more related to TnpD-like transposases from other species despite low sequence

homology (41, 37, and 36% with TDC1, TGM5, and TnpD, respectively) than they were to predicted TnpA-like proteins in ginseng (Fig. 3-4). Like the low conservation of TnpA-like protein sequence in ginseng, very low sequence homology was observed between ginseng TnpA-like transposase with those from the outgroup species (22, 22, and 20% DOPA, TDC1 and TGM5, respectively), further demonstrating the higher sequence variation of TnpA- than the TnpD-like region.

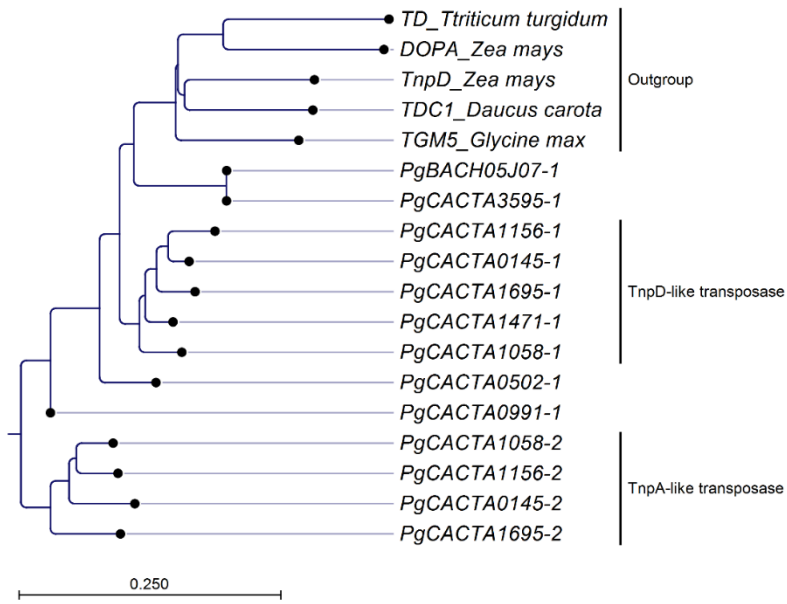


Fig. 3-4. Phylogenetic analysis of TnpD- and TnpA-like protein sequences. Proteins derived from the first gene of each element are denoted '1' while those from the second are denoted '2'. Elements like *PgCACTA1_1471*, *PgCACTA1_0502*, and *PgCACTA1_0991* have only one gene with fused TnpD- and TnpA-like domains. Truncation of the TnpA-like CDS region in *PgCACTA1_1471* and *PgCACTA1_0502* favors their grouping to the TnpD-like group, while the relatively intact CDS for both genes of *PgCACTA1_1471* placed it at intermediate position between the two groups. *PgCACTA1_3595* and *PgCACTA1_BACH05J07* have truncated TnpD-like region but with conserved tnp2 transposase family domain (pfam13963), common to other TnpD-like proteins.

The Pg167TR satDNA is a major RE in ginseng genome

Assembly of 0.74x ginseng WGS reads through *de novo* assembly of low-coverage WGS (dnaLCW) method generated several high read mapping contigs. The top 30 contigs covered a total of 266 Mb with lengths that ranged from 203 to 633 bp. In terms of genomic representation (GR), the top 30 contigs covered from 4.2 to 15.6 Mb and copies from 19,721 to 43,005. Twenty-nine of the contigs, totaling 260 Mb, belong to the *Gypsy* family of the LTR retrotransposon superfamily, while Pg167TR, the only satDNA in the top 30 which totaled 6 Mb, was the 27th most abundant contig (Table 2-3).

Pg167TR conformation implies chromatin coiling

Often, major satDNA families in plant species have sequence variants that localize either distinctly or not in different chromosomal regions (Koo *et al.* 2005; Findley *et al.* 2010; Waminal *et al.* 2015; Waminal *et al.* 2016b). Accordingly, a genome-wide survey was carried out to investigate Pg167TR sequence diversity within the ginseng genome. In our previous report, we identified a 1,603-bp Pg167TR array in BAC_PgH005J07 with a 167-bp repeat unit arranged 9.6 times (Choi *et al.* 2014), which localized in distinct chromosomal regions that are not DAPI-rich (Fig. 3-5A). A total of 344 representative Pg167TR sequences were obtained from homology search against genomic scaffolds including those from BAC_PgH005J07. Although Pg167TR is characterized by a relatively balanced GC (~50%) content, supporting its distribution outside AT-rich DAPI bands, a biased distribution of GCs towards the 5' while AT towards the 3' end was observed (Fig. 3-5B). The periodic interval of high and low GC content along the Pg167TR array gives Pg167TR an alternate curvature propensity peaks. (Fig. 3-5C, D). The curvature-propensity plot, calculated with DNase I-based trinucleotide parameters, contains one peculiar peak between nt 120 and 150 which showed a curvature value of ~13. So we believe that this region may adopt a curved conformation (Vlahovicek *et al.* 2003), associated with chromatin coiling.

Table 2-5. Top 30 contigs from dnaLCW analysis corresponding to 1x ginseng gnome with description from customized ginseng repeat database.

No.	Contig ID	Contig length	GR (copies)	GR (kb)	Description
1	477	633	24,633	15,593	LTR#Ty3_Gypsy/PgDel1_5
2	5312	410	36,154	14,823	LTR#Ty3_Gypsy/PgDel1_2
3	103	350	41,220	14,427	LTR#Ty3_Gypsy/PgDel1_4
4	1018	358	33,301	11,922	LTR#Ty3_Gypsy/PgDel1_2
5	1300	355	32,715	11,614	LTR#Ty3_Gypsy/PgDel1_4
6	8810	380	29,491	11,206	LTR#Ty3_Gypsy/PgDel1_1
7	12564	244	43,005	10,493	LTR#Ty3_Gypsy/PgDel1_5
8	246	336	30,180	10,140	LTR#Ty3_Gypsy/PgDel1_1
9	584	342	29,379	10,048	LTR#Ty3_Gypsy/PgDel1_5
10	250	353	28,204	9,956	LTR#Ty3_Gypsy/PgDel1_3
11	2297	344	27,864	9,585	LTR#Ty3_Gypsy/PgDel1_3
12	519	352	26,875	9,460	LTR#Ty3_Gypsy/PgDel1_3
13	39338	286	32,993	9,436	LTR#Ty3_Gypsy/PgDel1_1
14	11009	273	32,397	8,844	LTR#Ty3_Gypsy/PgDel1_4
15	31930	338	25,934	8,766	LTR#Ty3_Gypsy/PgDel1_4
16	5643	285	30,430	8,673	LTR#Ty3_Gypsy/PgDel1_4
17	438	318	26,145	8,314	LTR#Ty3_Gypsy/PgDel1_4
18	17496	376	20,482	7,701	LTR#Ty3_Gypsy/PgDel1_3
19	6395	338	22,651	7,656	LTR#Ty3_Gypsy/PgDel1_1
20	19275	266	28,755	7,649	LTR#Ty3_Gypsy/PgDel1_3
21	9549	335	22,457	7,523	LTR#Ty3_Gypsy/PgDel1_4
22	7018	290	24,385	7,072	LTR#Ty3_Gypsy/PgDel1_5
23	372	280	22,864	6,402	LTR#Ty3_Gypsy/PgDel1_4
24	3633	295	21,351	6,299	LTR#Ty3_Gypsy/PgDel1_5
25	2202	267	23,577	6,295	LTR#Ty3_Gypsy/PgDel1_4
26	3230	303	20,730	6,281	LTR#Ty3_Gypsy/PgDel1_1
27	32327	313	19,721	6,173	satellite#satDNA/Pg167TR
28	10632	227	22,807	5,177	LTR#Ty3_Gypsy/PgDel1_1
29	9414	217	21,615	4,690	LTR#Ty3_Gypsy/PgDel1_3
30	8630	203	20,804	4,223	LTR, gypsy, "Fatima_107G22-2"; 3' truncated

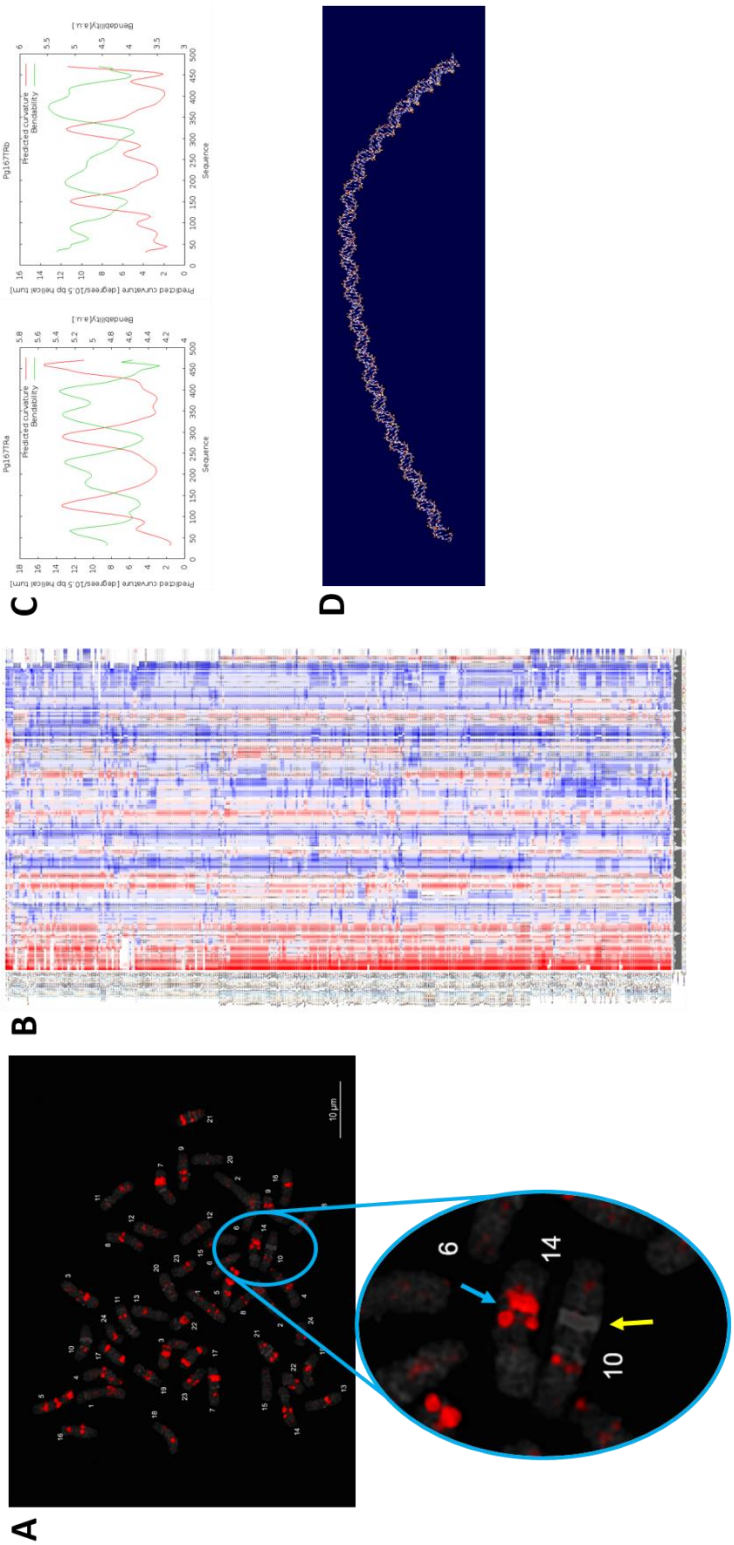

A
B
C
D
6
14
10
10 µm
Pg1977Ra
Pg1977Rb
Predicted curvature
Bendability
Sequence
Predicted curvature [degrees/10.5 bp helical turn]
Bendability [a.u.]

Fig. 3-5. Sequence characterization of Pg167TR revealed two major groups with high curvature propensity. A) Chromosomal distribution of Pg167TR (blue arrow in inset) along ginseng chromosomes showing distribution outside DAPI rich bands (yellow arrow). B) Multiple alignment of Pg167TR from representative *PgCACTA1* elements and top blast hits from scaffold sequences. Elements ranged from 140~168bp. Red and blue background denote high and low GC content, respectively. C) Curvature propensity and bendability peaks for three Pg167TRa and Pg167TRb units. Curvature propensity peak value (13) roughly corresponds to the value calculated for a highly curved motif in the *Leishmannia tarentolae* minicircle sequence. D) Simulation of Pg167TR unit curvature shows a curved conformation of Pg167TR.

Two subgroups of Pg167TR

Phylogenetic analysis of the 344 Pg167TR sequences resulted to two major groups we named Pg167TRa and Pg167TRb (Fig 2-2A). Sequence similarity of sequences within groups was greater than 90% while those between groups was relatively high at 86%. Although the sequences were relatively homogenized, polymorphism at nt 100~112 highlighted the variation between the two variants (Fig. 3-6B). Moreover, a 3-bp variant region at nt 110–112 has TCG for Pg167TRa and AGT for Pg167TRb (Fig. 3-6B). Pg167TRa was more abundant than Pg167TRb and was subdivided into three smaller groups (Fig. 3-6A, C). Exploiting this region to design oligonucleotide probes for FISH analysis supported the more abundance of Pg167TRa than Pg167TRb from *in silico* analysis (Fig. 3-7A-E). Pg167TRb hybridized to only 12 out of the 24 ginseng chromosome pairs (Fig. 3-7F). The Pg167TR locus in chromosome 3 showed a Pg167TRb-specific signal, and can be used as a chromosome 3-specific marker (Fig. 3-7E).

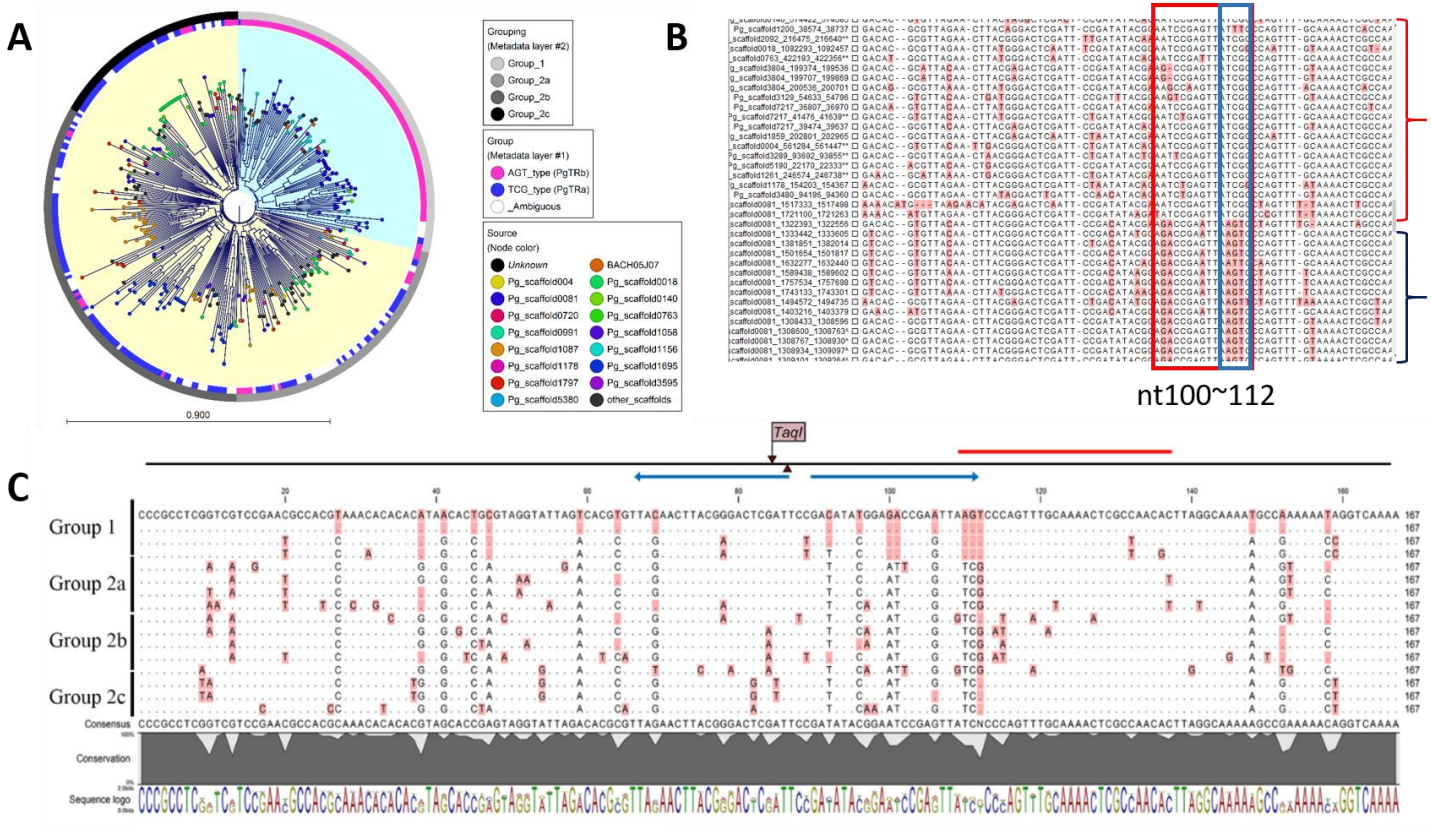

A
Grouping
(Metadata layer #2)
Group_1
Group_2a
Group_2b
Group_2c
Group
(Metadata layer #1)
AGT_type (PgTRb)
TCG_type (PgTRa)
_Ambiguous
Source
(Node color)
Unknown
BACH05J07
Pg_scaffold004
Pg_scaffold0018
Pg_scaffold0081
Pg_scaffold0140
Pg_scaffold0720
Pg_scaffold0763
Pg_scaffold0991
Pg_scaffold1058
Pg_scaffold1087
Pg_scaffold1156
Pg_scaffold1178
Pg_scaffold1695
Pg_scaffold1797
Pg_scaffold3595
Pg_scaffold5380
other_scaffolds
0.900
B
Pg167TRa
(>90%)
(~86%)
Pg167TRb
(>90%)
nt100~112
C
TaqI
Group 1
Group 2a
Group 2b
Group 2c
Consensus
Conservation
Sequence logo

Fig. 3-6. Analysis of the Pg167TR subgroups A) K-mer based phylogenetic tree of sequences from panel A showing two major groups (1 and 2), with three subgroups of group 2. B) Multiple sequence alignment Pg167TR sequences revealed the sequence polymorphism at nt 110~112 showing AGT and TCG major types. C) of five representative sequences of four Pg167TR groups. Group1 or Pg167TRb and three subgroups of Group2 or Pg167TRa. Blue bars depict PCR primers while the red bar indicate the oligoprobe to map Pg167TRa and Pg167TRb.

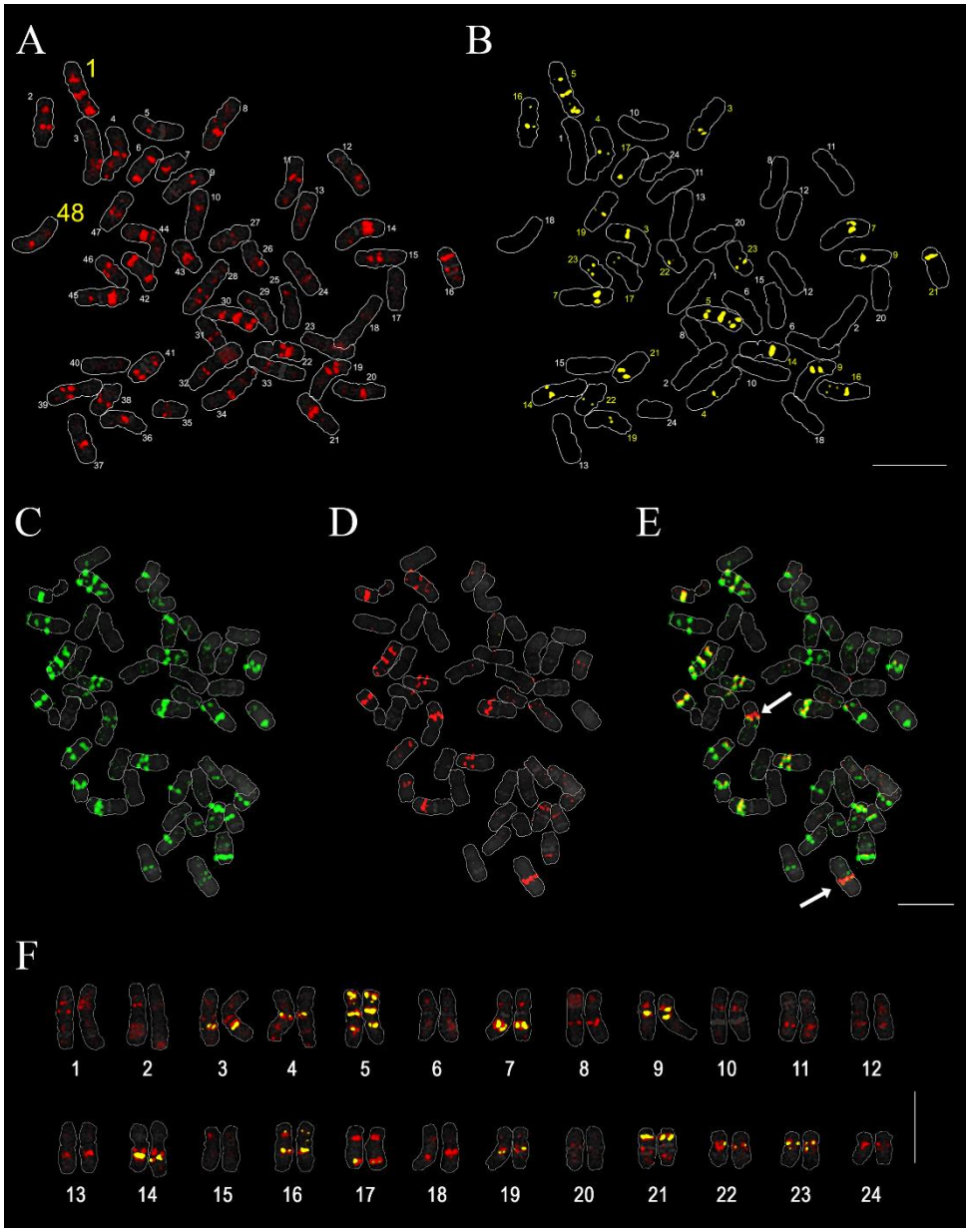

A
1
48
B
C
D
E
F
1 2 3 4 5 6 7 8 9 10 11 12
13 14 15 16 17 18 19 20 21 22 23 24

Fig. 3-7. Cytogenetic mapping of Pg167TR on *P. ginseng* chromosomes. A) Mapping of a 9.6 Pg167TR copies from BACH05J07 showing the distribution in ginseng chromosomes. Adapted from Waminal *et al.* (2016a). B) Distribution of Pg167TRb in the same chromosome spread as in A showing chromosome numbers. C) Distribution of Pg167TRa. D) Distribution of Pg167TRb. E) Overlay of Pg167TRa and Pg167TRb signals. F) Karyogram showing the distribution of Pg167TRb (yellow) over the BAC Pg167TR array (red). White, pink, and yellow arrows indicate chromosomes with relatively balanced signal intensity of the two sequence variants, more abundant Pg167TRa, and more abundant Pg167TRb, respectively. Blue arrowheads indicate Pg167TRa-specific chromosome loci. Bar = 10 µm.

Pg167TRa is more abundant than Pg167TRb in *Panax ginseng*, but less among species in related genera

To estimate the genomic abundance of each Pg167TR sequence variants between ginseng samples and related diploid and tetraploid species within and without the genus *Panax*, WGS reads of related species listed in Table 2-1 and 2-2 were mapped onto consensus sequences of Pg167TRa and Pg167TRb. Among ginseng samples, Pg167TR covered 22~52 Mb representing 1~2% of the ginseng genome (Fig _2-4). All ginseng samples had a mean Pg167TR content of 1.6%, with the hamyang sample having the lowest GR at 1.03% while the cultivar Sunhyang the highest at 2.23%. Pg167TR content among different ginseng samples was highly variable with a coefficient of variance of 22% (Fig 2-4).

Moreover, species within *Panax* generally had a higher Pg167TR GR compared with those outside the genus, which revealed low GR regardless of ploidy (Table 2-6). However, tetraploid species in *Panax* showed much more Pg167TR than diploid species. Ginseng (Chunpoong) had the highest of about 1.0% among all species compared, even more than those in *P. quinquefolius* (0.3%), whose genome size is almost 1 Gb more than that of *P. ginseng*.

Pg167TR variants showed differential amplification in all species, with Pg167TRa being more abundant within *Panax*, but less than Pg167TRb in species outside *Panax* (Fig. 3-9A-C). Moreover, despite considerable variation in copy number between the two sequence variants, a relatively conserved Pg167TRa-Pg167TRb ratio was observed in groups within and without *Panax* (Fig. 3-9B). Pg167TRa was about 80% of the total Pg167TR elements in *Panax*, while it was only about 20% in species outside the genus. Cytogenetic mapping of Pg167TR elements supported the *in silico* data showing the abundance of Pg167TR in ginseng than in *P. notoginseng* (Fig. 3-9C). FISH-based signal-to-chromosome area ratio quantified Pg167TR abundance to about 8% (~280 Mb) of the ginseng genome (Table 2-6), which is much more than what was estimated through WGS read mapping.

Table 2-6. Comparison between in silico and FISH estimation methods for genomic presence of Pg167TR.

Method	Copies	Genome coverage (kb)	GR
*In silico**	166,410	36,683	1.02%
FISH**	1,722,699	280,800	7.80%

*Total Pg167TR from WGS read mapping
**Signal area relative to chromosome area

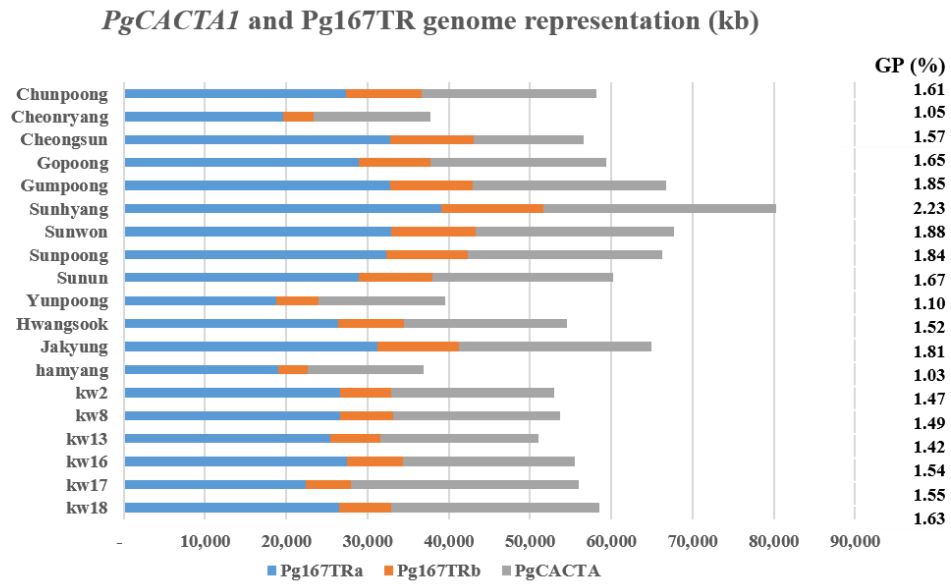


Fig. 3-8. Quantification of Pg167TR and *PgCACTA1* abundance among ginseng samples. Genome representation of Pg167TRa, Pg167TRb and *PgCACTA1* showing differential abundance of total *PgCACTA1* elements as well as Pg167TR sequences. The coefficient of variance of Pg167TR content was 22%, while for the total *PgCACTA1* abundance was 19%.

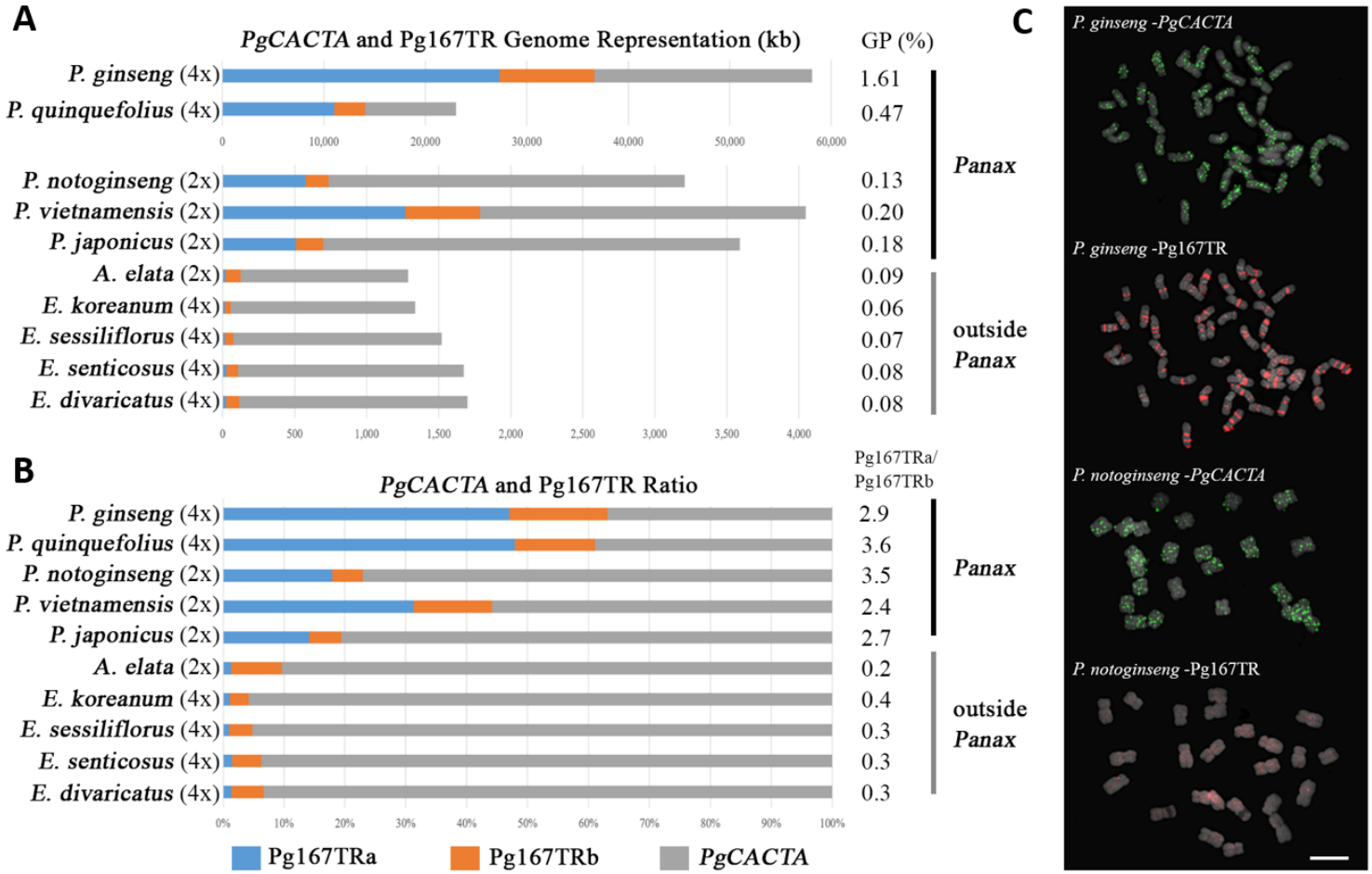
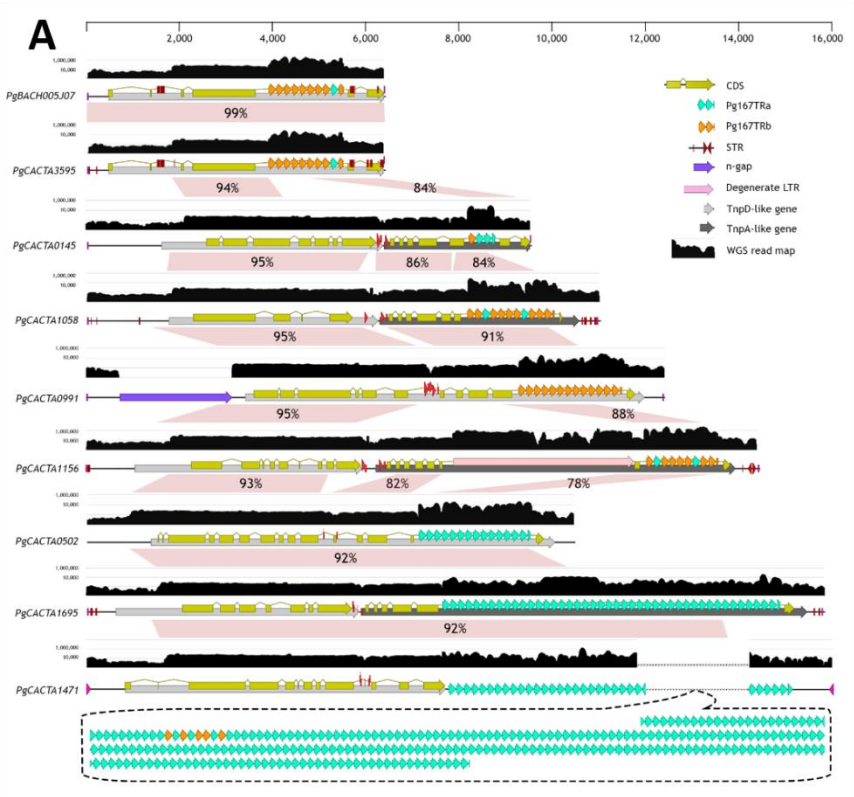

A
PgCACTA and Pg167TR Genome Representation (kb)
GP (%)
P. ginseng (4x)
P. quinquefolius (4x)
P. notoginseng (2x)
P. vietnamensis (2x)
P. japonicus (2x)
A. elata (2x)
E. koreanum (4x)
E. sessiliflorus (4x)
E. senticosus (4x)
E. divaricatus (4x)
1.61
0.47
0.13
0.20
0.18
0.09
0.06
0.07
0.08
0.08
Panax
outside Panax
0
10,000
20,000
30,000
40,000
50,000
60,000
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
B
PgCACTA and Pg167TR Ratio
Pg167TRa/ Pg167TRb
2.9
3.6
3.5
2.4
2.7
0.2
0.4
0.3
0.3
0.3
Panax
outside Panax
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Pg167TRa
Pg167TRb
PgCACTA
C
P. ginseng -PgCACTA
P. ginseng -Pg167TR
P. notoginseng -PgCACTA
P. notoginseng -Pg167TR

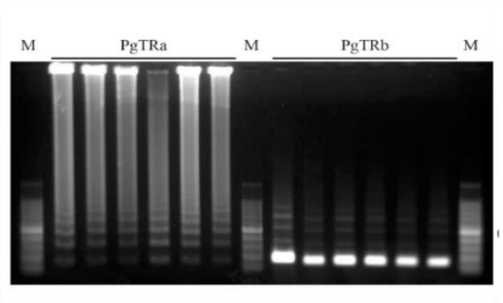
Fig. 3-9. Quantification of genomic Pg167TRa and Pg167TRb within *Panax* and outside *Panax*. A) Genome representation (GR) of each Pg167TR variant within and without the genus *Panax*, showing the higher and lower abundance in tetraploid and diploid *Panax* species, respectively, but related species outside the genus have highly reduced GR, regardless of ploidy. B) Ratio of total Pg167TR sequences with total *PgCACTA1* elements in species from panel A, showing a much lesser Pg167TR sequences in species outside the genus. C) Comparative FISH analysis with *PgCACTA1* and Pg167TR sequences between *P. ginseng* and *P. notoginseng* showed signals for PgCACTA*1* in both tetraploid and diploid species, respectively. While highly abundant and distinct Pg167TR signals were observed in *P. ginseng*, very faint signals were observed in *P, notoginseng*, supporting *in silico* data. Bar = 10 μm.

Pg167TRa was associated with Pg167TR expansion in *PgCACTA1*

Initial results showing lower abundance of Pg167TRb compared with Pg167TRa in phylogenetic (Fig. 3-5) and FISH (Fig. 3-9) analyses raised a question about the evolutionary dynamics of these two sequence variants. To explore the answer to this question, homology search was carried out using oligoprobes used in FISH analysis as query against each identified *PgCACTA1* elements in order to distinguish each sequence variant (Table 2-3). Pg167TRa was observed in *PgCACTA1* elements with either low Pg167TR copy number or expanded Pg167TR loci. However, Pg167TRb elements were detected only in elements with few copy numbers or in short fragments in elements with expanded Pg167TR repeats (Fig. 3-10A). PCR amplification supports this observation (Fig. 3-10B), corroborating phylogenetic and FISH analyses. Moreover, *PgCACTA1_0018* has a highly amplified Pg167TR region, all of which are of the Pg167TRa variant, and a lot of N-gaps within the Pg167TR array, indicating an even longer array not resolved by the assembly algorithm used (Fig. 3-11).

In addition, Pg167TRa was generally expressed in normal, stress-induced, and adventitious root cultures of ginseng (Fig. 3-10C), indicating transcription of Pg167TR sequences and a biased abundance towards Pg167TRa than Pg167TRb.


A
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
PgBACH005J07
99%
PgCACTA3595
94%
84%
PgCACTA0145
95%
86%
84%
PgCACTA1058
95%
91%
PgCACTA0991
95%
88%
PgCACTA1156
93%
82%
78%
PgCACTA0502
92%
PgCACTA1695
92%
PgCACTA1471
CDS
Pg167TRa
Pg167TRb
STR
n-gap
Degenerate LTR
TnpD-like gene
TnpA-like gene
WGS read map


B
M
PgTRa
M
PgTRb
M

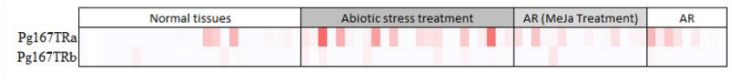

C
Normal tissues
Abiotic stress treatment
AR (MeJa Treatment)
AR
Pg167TRa
Pg167TRb

Fig. 3-10. Characterization of identified *PgCACTA1* elements. A) Diagrammatic representation of representative *PgCACTA1* elements. Black graphs at the upper portion of each element represent mapping depths of ~1x coverage WGS reads to each element at log 100 scale. The two transposase coding genes are shown in light and dark grey, respectively, corresponding to the highly conserved regions of the elements. Pg167TR units at the 3' regions show variable copy number. Some elements harbor only Pg167TRb units exclusively, while some harbor both Pg167TRa and Pg167TRb variants. Pg167TRa were often associated with expanded Pg167TR arrays. Some elements show n-gaps and LTR-insertion. The 75-bp direct repeats (see text) are shown as red arrows between two transposase genes. B) PCR amplification of Pg167TRa and Pg167TRb showing the amplification of high-molecular weight products in Pg167TRa but lower in Pg167TRb. C) Expression analysis of Pg167TR variants by mapping of total transcriptome reads to consensus sequences of Pg167TR variants showing a generally higher expression of Pg167TRa in normal and stress-induced tissues and adventitious root cultures.

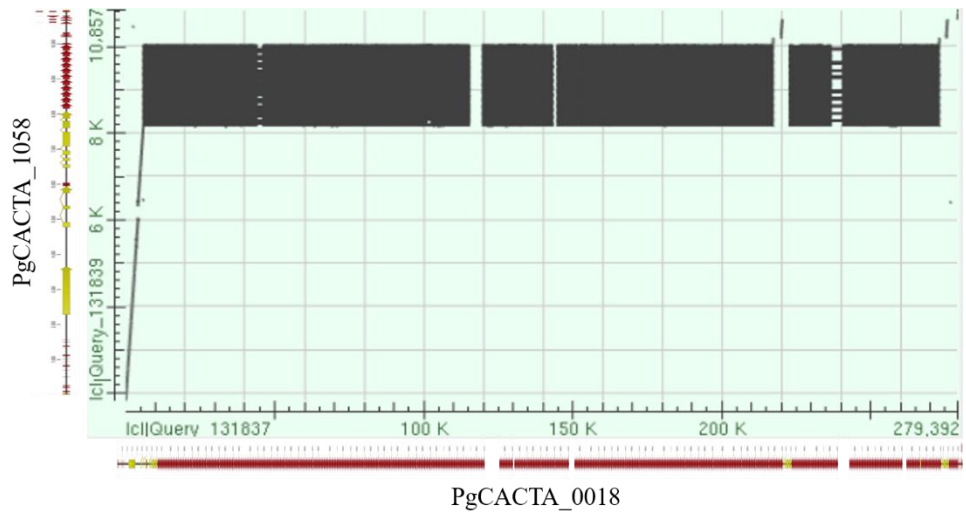


Fig. 3-11. Dot-plot of a Pg167TR-amplified region of Pg_scaffold0018. Sequence comparison between PgCACTA1_1058 and PgCACTA1_0018 (no. 20 in Table 2-2) revealed a high amplification of the Pg167TR region while having conserved TIR regions.

Putative *cis*-regulatory elements are encoded in Pg167TR sequences

Although once regarded as junk DNA, the abundance of satDNAs in most plant genomes imply a functional role. Indeed, several satDNAs have been implicated in gene regulation and heterochromatin formation (Ugarkovic 2005). To predict the regulatory potential of Pg167TR satDNA, homology search for *cis*-regulatory motifs imbedded in the Pg167TR units was conducted against plant regulatory element databases (see Materials and Methods).

Forty-five putative *cis*-regulatory element motifs associated with physiological, biotic and abiotic responses were found to be homologous with those reported in 15 plant species (Table 2-7). Out of the 45 motifs, 17 were involved as putative promoters for gene expression, followed by those involved in light response (Table 2-8).

Table 2-7. List of regulatory element motifs found in Pg167TR sequences.

No.	Site name	Source organism	Sequence	Function
1	A-box	*Petroselinum crispum*	CCGTCC	cis-acting regulatory element
2	ABRE	*Hordeum vulgare*	CCGCGTAGGC	cis-acting element involved in the abscisic acid responsiveness
3	ABRE	*Hordeum vulgare*	GCCACGTACA	cis-acting element involved in the abscisic acid responsiveness
4	ABRE	*Arabidopsis thaliana*	ACGTGGC	cis-acting element involved in the abscisic acid responsiveness
5	ABRE	*Arabidopsis thaliana*	CACGTG	cis-acting element involved in the abscisic acid responsiveness
6	ARE	*Zea mays*	TGGTTT	cis-acting regulatory element essential for the anaerobic induction
7	CAAT-box	*Arabidopsis thaliana*	gGCAAT	common cis-acting element in promoter and enhancer regions
8	CAAT-box	*Brassica rapa*	CAAAT	common cis-acting element in promoter and enhancer regions
9	CAAT-box	*Glycine max*	CAATT	common cis-acting element in promoter and enhancer regions
10	CAAT-box	*Hordeum vulgare*	CAAT	common cis-acting element in promoter and enhancer regions
11	CAAT-box	*Arabidopsis thaliana*	CCAAT	common cis-acting element in promoter and enhancer regions
12	CAAT-box	*Petunia hybrida*	TGCCAAC	common cis-acting element in promoter and enhancer regions
13	CATT-motif	*Zea mays*	GCATTC	part of a light responsive element
14	CE3	*Oryza sativa*	GACGCGTGTC	cis-acting element involved in ABA and VP1 responsiveness
15	CGTCA-motif	*Hordeum vulgare*	CGTCA	cis-acting regulatory element involved in the MeJA-responsiveness
16	chs-CMA1a	*Daucus carota*	TTACTTAA	part of a light responsive element
17	DOFCOREZM	*Zea mays*	AAAG	Core site required for binding of Dof proteins in maize (Z.m.)
18	G-box	*Daucus carota*	TACGTG	cis-acting regulatory element involved in light responsiveness
19	G-box	*Brassica napus*	CCACGTAA	cis-acting regulatory element involved in light responsiveness
20	G-Box	*Antirrhinum majus*	CACGTA	cis-acting regulatory element involved in light responsiveness
21	G-box	*Larix laricina*	GACACGTAGT	cis-acting regulatory element involved in light responsiveness
22	G-box	*Lycopersicon esculentum*	ACACGTGGCACC	cis-acting regulatory element involved in light responsiveness
23	G-box	*Brassica napus*	ACACGTGT	cis-acting regulatory element involved in light responsiveness
24	GC-motif	*Zea mays*	GCCCCGG	enhancer-like element involved in anoxic specific inducibility
25	GCN4_motif	*Oryza sativa*	TGTGTCA	cis-regulatory element involved in endosperm expression
26	LTR	*Hordeum vulgare*	CCGAAA	cis-acting element involved in low-temperature responsiveness
27	MBS	*Arabidopsis thaliana*	CAACTG	MYB binding site involved in drought-inducibility
28	MBS	*Zea mays*	CGGTCA	MYB Binding Site
29	MBS	*Arabidopsis thaliana*	TAACTG	MYB binding site involved in drought-inducibility
30	SEF4MOTIFGM7S	*Glycine max*	RTTTTTR	SEF4 binding site; Soybean (G.m.)

31	Skn-1_motif	*Oryza sativa*	GTCAT	cis-acting regulatory element required for endosperm expression
32	Sp1	*Oryza sativa*	GGGCGG	light responsive element
33	TATA-box	*Arabidopsis thaliana*	TATA	core promoter element around -30 of transcription start
34	TATA-box	*Daucus carota*	ccTATAAATT	core promoter element around -30 of transcription start
35	TATA-box	*Oryza sativa*	TATAAGAA	core promoter element around -30 of transcription start
36	TATA-box	*Arabidopsis thaliana*	TATAA	core promoter element around -30 of transcription start
37	TATA-box	*Lycopersicon esculentum*	TTTTA	core promoter element around -30 of transcription start
38	TATA-box	*Helianthus annuus*	TATACA	core promoter element around -30 of transcription start
39	TATA-box	*Arabidopsis thaliana*	tcTATATAtt	core promoter element around -30 of transcription start
40	TATA-box	*Brassica napus*	ATATAT	core promoter element around -30 of transcription start
41	TATA-box	*Glycine max*	TAATA	core promoter element around -30 of transcription start
42	TATA-box	*Pisum sativum*	TATATGT	core promoter element around -30 of transcription start
43	TCT-motif	*Arabidopsis thaliana*	TCTTAC	part of a light responsive element
44	TGACG-motif	*Hordeum vulgare*	TGACG	cis-acting regulatory element involved in the MeJA-responsiveness
45	W box	*Arabidopsis thaliana*	TTGACC	elicitation; wounding and pathogen responsievness. Binds WRKY type transcription factors

Table 2-8. Summary of putative *cis*-acting regulatory element motifs found in Pg167TR sequences.

Response	No. of Motifs	Percentage
Promoter	17	37.78%
Light	10	22.22%
Acid	4	8.89%
Myb	3	6.67%
Endosperm	3	6.67%
Drought	2	4.44%
Meja	2	4.44%
Others	2	4.44%
Temp	1	2.22%
Pathogen	1	2.22%
	45	100%

DISCUSSION

Several plant satDNAs have been discovered in many plant and animal species (Laurent *et al.* 1999; Mehrotra and Goyal 2014). While most of these satDNAs are often associated with Class I TEs or retrotransposons (Cheng and Murata 2003; Tek *et al.* 2005; Macas *et al.* 2009; Gong *et al.* 2012; Sharma *et al.* 2013), very limited report shows their origin from Class II TEs or DNA transposon (Dias *et al.* 2014). A recent study on *Drosophila virilis* has shown an origin of satDNA from a fold-back element, a Class II TE (Dias *et al.* 2014). And although a satDNA family of Triticeae, *Afa*, was associated with a *CACTA* transposon, there was no report on its amplification to long arrays (Nagaki *et al.* 1998; Wicker *et al.* 2003). Here, a novel satDNA amplification pathway of Pg167TR from *PgCACTA1*, a Class II TE, is presented.

Pg167TR predated the diversification of *Panax* and related Araliaceae species

Recent advances in genomics have shown that all angiosperms are either paleo- or neo-polyploids (Jiao *et al.* 2011; Project 2013; Tank *et al.* 2015). Accordingly, the genus *Panax* had experienced two rounds of whole genome duplications (WGD)—a paleo-polyploidization event (24.6–32.8 MYA) and a neo-polyploidization event (1.6–3.3 MYA) (Choi *et al.* 2014). A recent phylogenetic study based on chloroplast genome and nuclear rDNA sequences demonstrated the divergence of the genus *Panax* from *Aralia* around 11.2–12.1 MYA and from other genera, like *Kalopanax* and *Dendropanax*, around 15.7 MYA (Shi *et al.* 2015). This chronology supports the observed relationship between WGD and species diversification (Tank *et al.* 2015), suggesting that the paleo-WGD event promoted subsequent species diversification that eventually resulted to the extant species. Taking this relationship into account and the existence of Pg167TR sequences in species outside *Panax* imply the

existence of Pg167TR in the ancestral genome of the sampled Araliaceae species, which likely predated the paleo-WGD.

Accordingly, Pg167TR sequence variants have likely been present in the ancestral genomes of both *Panax* and related Araliaceae species in low copies much like in a satDNA library (Plohl *et al.* 2012), and after paleo-WGD and subsequent speciation, Pg167TR followed genome dynamics independent within each respective species; such that, Pg167TR dynamics within *Panax* favored amplification of Pg167TRa than Pg167TRb, while an opposite dynamics acted for species outside the genus. Eventually, the subsequent neo-WGD that predated the divergence of *P. ginseng* and *P. quinquefolius* further promoted the amplification of the Pg167TR sequences within the two tetraploids, albeit retaining the Pg167TRa to Pg167TRb ratio observed in the genus *Panax*. But more notably, Pg167TR showed a more efficient amplification or retention in *P. ginseng* than in *P. quinquefolius*, despite a much larger genome of the latter (Obae 2012).

FISH data supported *in silico* findings regarding the more abundance of Pg167TRa compared with Pg167TRb. Although both sequences showed rapid divergence (Ho and Leung 2002) and recombination, efficient homogenization could have been reduced through epigenetic "locking" of variants (Fedoroff 2012). Also, it is noteworthy that Pg167TRb hybridized to only 12 out of the 24 chromosome pairs (Choi *et al.* 2014). Although distinct chromosomal hybridization of sequence variants of the same satDNA family is not uncommon to plants (Lim *et al.* 2005; Han *et al.* 2008), a similar observation in Pg167TR and PgDel2 supports a more complex phylogenetic history of the ginseng genome (Yi *et al.* 2004; Shi *et al.* 2015). Tetraploids in the genus *Panax* (i.e. *P. ginseng*, and *P. quinquefolius*) are known to share a similar maternal genome, although paternal genomes are not yet fully elucidated (Shi *et al.* 2015). However, it is more likely that ginseng followed a more complex and reticulated evolutionary pathway, and an ancient basic chromosome number of six still cannot be ruled out (Yi *et al.* 2004).

The differential abundance of Pg167TR between these two genomes, like in other diploid and related species, can be attributed to the epigenetic control acting within each genome after WGD, and considering that different environmental filters may have influenced epigenetics during species evolution (Kalendar *et al.* 2000; Fedoroff 2012). Additionally, the favored amplification and transcription of Pg167TRa may suggest underlying functions that are more active in the genome compared with that of Pg167TRb, noting that satDNAs are involved in epigenetic mechanisms (Ugarkovic 2005), that repetitive elements are often transcribed to non-coding RNAs (Matylla-Kulinska *et al.* 2014), and that *cis*-regulatory elements reside in the Pg167TR sequences.

Pg167TR sequence implies role in heterochromatin packing and gene expression

Heterochromatin is categorized as either constitutive or facultative based on their chromosome location or characteristic state of condensation (Grewal and Jia 2007). Constitutive heterochromatins are condensed throughout the cell cycle and are often localized in centromeres, knobs, and telomeres (Grewal and Jia 2007; Vourc'h and Biamonti 2011). Facultative heterochromatins do not have specific chromosome loci and appear or disappear in response to cellular signals in different developmental stages (Grewal and Jia 2007; Puertas and Villasante 2013). Heterochromatins are necessary for efficient nuclear organization, chromosomal segregation, and interaction with *cis*-regulatory elements for control of gene expression (Jia *et al.* 2004; Grewal and Jia 2007).

Constitutive heterochromatins comprise different TE and TR families. Previous studies have demonstrated a crucial role of satDNAs in heterochromatin condensation, maintenance, and chromosome stability through RNAi-mediated mechanisms (Martienssen 2003; Grewal and Jia 2007; Fedoroff 2012; Pezer *et al.* 2012). Inherent structural DNA conformation, such as DNA curvature, also influence condensation of tandem repeats in heterochromatins (Radic *et al.* 1987;

Martienssen 2003; Grewal and Jia 2007). The periodic cycle of this GC content pattern along the Pg167TR array and the high curvature propensity of each repeat unit imply its association in heterochromatin formation (Radic *et al.* 1987; Vlahovicek *et al.* 2003).

Regions in Pg167TR units matching *cis*-acting regulatory elements from other plants imply an involvement of Pg167TRs in gene expression, as such DNA blocks are typically promoter regions and transcription factor binding sites (Brajkovic *et al.* 2012; Pezer *et al.* 2012; Belele *et al.* 2013; Mehrotra *et al.* 2014; Mehrotra and Goyal 2014). These imply cellular involvement of Pg167TR not just structurally but also functionally by acting as gene promoters and regulating mechanisms involved in stress responses, particularly light-related stress. Moreover, the observed transcription of satDNAs and their purported function as long non-coding RNA (Vourc'h and Biamonti 2011; Paço *et al.* 2014), provide an avenue for further research on Pg167TR gene control pathway, one that involves long non-coding RNAs. Further studies and manipulation of Pg167TR should test this hypothesis and provide insight to the roles of Pg167TR in *Panax* physiology considering that species in this genus, particularly *P. ginseng*, which also happen to have the most abundant Pg167TR content, are mostly sensitive to light and temperature than other Araliaceae species (Court 2000).

PgCACTA1 spurred the amplification of long Pg167TR arrays

Among Class II TEs, members of the *CACTA* family are characterized by terminal inverted repeats (TIRs) of 10 to 100 bp that end with 5'-CACTA-3' motif (Wicker *et al.* 2003), and often harbor TRs of different sizes and orientation in the non-coding regions (Wicker *et al.* 2003). Additionally, *CACTA* elements are characterized by two important gene products, transposase (TnpD) and a regulatory transposase (TnpA), which are both necessary for element transposition (Masson *et al.* 1991; Fedoroff 2013b). The *PgCACTA1* element identified in BAC PgH005J07 has a truncated TnpD-like domain and lacks the TnpA-like domain (Fig. 3-2). Subterminal repeats (STRs) in the 5' and 3' untranslated regions of *CACTA* elements serve as binding sites of TnpA for transposition (Fedoroff 2013b). The incomplete transposase domains in *PgCACTA1_BACH05J07* indicates that it is a non-autonomous element, while other elements like *PgCACTA1_1058* and *PgCACTA1_0502* have conserved and transcribed TnpD- and TnpA-like regions suggesting autonomous transposability. Some elements like *PgCACTA1_0991* and *PgCACTA1_0502*, might have lost a stop codon in between the transposase domains, making only a single ORF coding for TnpD- and TnpA-like proteins. However, although complete elements are present in a genome, their transcriptional activity is often governed by epigenetic mechanisms, rendering most elements inactive (Fedoroff 2013a; Fedoroff 2013b).

In addition to STRs in *PgCACTA1*, a conserved 75-bp direct repeat often separating the two transposase genes may possibly play a role in *PgCACTA1* function. Further analysis of this repeat may provide information about the *PgCACTA* function. In addition, although there is no functional study to support the direct involvement of Pg167TR in the overall function of *PgCACTA1*, its insertion at the last intron of the TnpA-like gene suggests an STR-like role for transposition (Masson *et al.* 1991).

Several *PgCACTA1* elements carry variable copy numbers of Pg167TR, with some elements highly amplified that current algorithms in the assembler used can't

resolve these highly repetitive loci as manifested by abundant N-gaps within Pg167TR arrays (Table. 2-6). These data suggest the evolution of long Pg167TR chromosomal loci from the amplification of Pg167TR in functional *PgCACTA1* elements. FISH data supports the observation of *PgCACTA1* spurring the amplification of Pg167TR to long arrays.

In addition, the abundance of both *PgCACTA1* and Pg167TR elements in *Panax*, especially in the tetraploid species, also supports the origin of Pg167TR from *PgCACTA1*. A recent concomitant Pg167TR expansion most likely occurred during the recent WGD that activated the *PgCACTA1* transposition in response to genomic shock (Fedoroff 2012), resulting to a burst in *PgCACTA1* and Pg167TR in tetraploid *Panax* species. Differential epigenetic responses of species to environmental nuances after WGD often lead to species diversification (Tank *et al.* 2015), which most likely acted between *P. ginseng* and *P. quinquefolius*.

Mechanisms involved in novel satDNA formation or amplification often include sequence homology-driven unequal crossovers, rolling circle replication of extrachromosomal circular DNA, gene conversion, and segmental duplication (Ma and Jackson 2006; Sharma *et al.* 2013). More detailed analysis on the amplification of Pg167TR from *PgCACTA1* may shed more light on how this repeats impact ginseng physiology.

REFERENCES

Albert PS, Gao Z, Danilova TV, and Birchler JA. 2010. Diversity of chromosomal karyotypes in maize and its relatives. Cytogenet Genome Res 129: 6-16.

Alix K, Joets J, Ryder CD, Moore J, Barker GC, Bailey JP, *et al.* 2008. The cacta transposon bot1 played a major role in brassica genome divergence and gene proliferation. The Plant Journal 56: 1030-1044.

Ananiev EV, Phillips RL, and Rines HW. 1998. Complex structure of knob DNA on maize chromosome 9: Retrotransposon invasion into heterochromatin. Genetics 149: 2025-2037.

Belele CL, Sidorenko L, Stam M, Bader R, Arteaga-Vazquez MA, and Chandler VL. 2013. Specific tandem repeats are sufficient for paramutation-induced trans-generational silencing. PLoS Genetics 9: e1003773.

Bennetzen JL, and Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol 65: 505-30.

Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res 27: 573-80.

Brajkovic J, Feliciello I, Bruvo-Madaric B, and Ugarkovic D. 2012. Satellite DNA-like elements associated with genes within euchromatin of the beetle tribolium castaneum. G3 (Bethesda) 2: 931-41.

Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, and Parkhill J. 2005. Act: The artemis comparison tool. Bioinformatics 21: 3422-3423.

Charlesworth B, Sniegowski P, and Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371: 215-20.

Cheng Z-J, and Murata M. 2003. A centromeric tandem repeat family originating from a part of ty3/gypsy-retroelement in wheat and its relatives. Genetics 164: 665-672.

Choi HI, Waminal NE, Park HM, Kim NH, Choi BS, Park M, *et al.* 2014. Major repeat components covering one-third of the ginseng (panax ginseng c.A. Meyer) genome and evidence for allotetraploidy. Plant J 77: 906-916.

Cohen S, Yacobi K, and Segal D. 2003. Extrachromosomal circular DNA of tandemly repeated genomic sequences in drosophila. Genome Res 13: 1133-45.

Coluccia E, Pichiri G, Nieddu M, Coni P, Manconi S, Deiana AM, *et al.* 2011. Identification of two new repetitive elements and chromosomal mapping of repetitive DNA sequences in the fish gymnothorax unicolor (anguilliformes: Muraenidae). Eur J Histochem 55: e12.

Court WE. 2000. Ginseng: The genus *panax*, Amsterdam, Netherlands, Hardwood Academic Publishers.

Dias GB, Svartman M, Delprat A, Ruiz A, and Kuhn GC. 2014. Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of drosophila virilis. Genome Biol Evol 6: 1302-13.

Dover G. 1982. Molecular drive: A cohesive mode of species evolution. Nature 299: 111-117.

Fambrini M, Basile A, Salvini M, and Pugliesi C. 2014. Excisions of a defective transposable cacta element (tetu1) generate new alleles of a cycloidea-like gene of helianthus annuus. Gene 549: 198-207.

Fedoroff NV. 2012. Presidential address. Transposable elements, epigenetics, and genome evolution. Science 338: 758-67.

Fedoroff NV 2013a. The discovery of transposition. *Plant transposons and genome dynamics in evolution.* Wiley-Blackwell.

Fedoroff NV. 2013b. Molecular genetics and epigenetics of cacta elements. Methods Mol Biol 1057: 177-92.

Findley SD, Cannon S, Varala K, Du J, Ma J, Hudson ME, *et al.* 2010. A fluorescence in situ hybridization system for karyotyping soybean. Genetics 185: 727-44.

Gao D, Zhao D, Abernathy B, Iwata-Otsubo A, Herrera-Estrella A, Jiang N, *et al.* 2016. Dynamics of a novel highly repetitive cacta family in common bean (phaseolus vulgaris). G3: Genes|Genomes|Genetics 6: 2091-2101.

Gong Z, Wu Y, Koblížková A, Torres GA, Wang K, Iovene M, *et al.* 2012. Repeatless and repeat-based centromeres in potato: Implications for centromere evolution. The Plant Cell 24: 3559-3574.

Grewal SI, and Jia S. 2007. Heterochromatin revisited. Nat Rev Genet 8: 35-46.

Han YH, Zhang ZH, Liu JH, Lu JY, Huang SW, and Jin WW. 2008. Distribution of the tandem repeat sequences and karyotyping in cucumber (cucumis sativus l.) by fluorescence in situ hybridization. Cytogenet Genome Res 122: 80-8.

He ZH, Dong HT, Dong JX, Li DB, and Ronald PC. 2000. The rice rim2 transcript accumulates in response to magnaporthe grisea and its predicted protein product shares similarity with tnp2-like proteins encoded by cacta transposons. Mol Gen Genet 264: 2-10.

Higo K, Ugawa Y, Iwamoto M, and Korenaga T. 1999. Plant cis-acting regulatory DNA elements (place) database: 1999. Nucleic Acids Res 27: 297-300.

Ho IS, and Leung FC. 2002. Isolation and characterization of repetitive DNA sequences from *panax ginseng*. Moleclar Genetics and Genomics 266: 951-61.

Jayakodi M, Lee S-C, Park H-S, Jang W, Lee YS, Choi B-S, *et al.* 2014. Transcriptome profiling and comparative analysis of panax ginseng adventitious roots. Journal of Ginseng Research 38: 278-288.

Jia S, Yamada T, and Grewal SI. 2004. Heterochromatin regulates cell type-specific long-range chromatin interactions essential for directed recombination. Cell 119: 469-80.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473: 97-100.

Jo SH, Koo DH, Kim JF, Hur CG, Lee S, Yang TJ, *et al.* 2009. Evolution of ribosomal DNA-derived satellite repeat in tomato genome. BMC Plant Biol 9: 42.

Kalendar R, Tanskanen J, Immonen S, Nevo E, and Schulman AH. 2000. Genome evolution of wild barley (hordeum spontaneum) by bare-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci U S A 97: 6603-7.

Kelly LJ, Renny-Byfield S, Pellicer J, Macas J, Novak P, Neumann P, *et al.* 2015. Analysis of the giant genomes of fritillaria (liliaceae) indicates that a lack of

DNA removal characterizes extreme expansions in genome size. New Phytol 208: 596-607.

Kim JH, Kim MK, Wang H, Lee HN, Jin CG, Kwon WS, *et al.* 2016. Discrimination of korean ginseng (panax ginseng meyer) cultivar chunpoong and american ginseng (panax quinquefolius) using the auxin repressed protein gene. J Ginseng Res 40: 395-399.

Kim K, Lee S-C, Lee J, Yu Y, Yang K, Choi B-S, *et al.* 2015. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of oryza aa genome species. Scientific Reports 5: 15655.

Kim NH, Choi HI, Kim KH, Jang W, and Yang TJ. 2014. Evidence of genome duplication revealed by sequence analysis of multi-loci expressed sequence tag-simple sequence repeat bands in panax ginseng meyer. J Ginseng Res 38: 130-5.

Koo DH, Choi HW, Cho J, Hur Y, and Bang JW. 2005. A high-resolution karyotype of cucumber (cucumis sativus l. 'Winter long') revealed by c-banding, pachytene analysis, and rapd-aided fluorescence in situ hybridization. Genome 48: 534-40.

Koo DH, Hong CP, Batley J, Chung YS, Edwards D, Bang JW, *et al.* 2011. Rapid divergence of repetitive dnas in brassica relatives. Genomics 97: 173-85.

Kubis SE, Heslop-Harrison JS, Desel C, and Schmidt T. 1998. The genomic organization of non-ltr retrotransposons (lines) from three beta species and five other angiosperms. Plant Mol Biol 36: 821-31.

Laurent A-M, Puechberty J, and Roizès G. 1999. Hypothesis: For the worst and for the best, l1hs retrotransposons actively participate in the evolution of the human centromeric alphoid sequences. Chromosome Research 7: 305-317.

Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van De Peer Y, *et al.* 2002. Plantcare, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 30: 325-7.

Leung KW, and Wong AS. 2010. Pharmacology of ginsenosides: A literature review. Chinese Medicine 5: 20.

Lim KB, De Jong H, Yang TJ, Park JY, Kwon SJ, Kim JS, *et al.* 2005. Characterization of rdnas and tandem repeats in the heterochromatin of brassica rapa. Molecules and Cells 19: 436-44.

Lim KB, Yang TJ, Hwang YJ, Kim JS, Park JY, Kwon SJ, *et al.* 2007. Characterization of the centromere and peri-centromere retrotransposons in brassica rapa and their distribution in related brassica species. Plant J 49: 173-83.

Ma J, and Jackson SA. 2006. Retrotransposon accumulation and satellite amplification mediated by segmental duplication facilitate centromere expansion in rice. Genome Res 16: 251-9.

Macas J, Koblížková A, Navrátilová A, and Neumann P. 2009. Hypervariable 3′ utr region of plant ltr-retrotransposons as a source of novel satellite repeats. Gene.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, *et al.* 2015. Cdd: Ncbi's conserved domain database. Nucleic Acids Res 43: D222-6.

Martienssen RA. 2003. Maintenance of heterochromatin by rna interference of tandem repeats. Nat Genet 35: 213-4.

Masson P, Strem M, and Fedoroff N. 1991. The tnpa and tnpd gene products of the spm element are required for transposition in tobacco. Plant Cell 3: 73-85.

Matyasek R, Gazdova B, Fajkus J, and Bezdek M. 1997. Ntrs, a new family of highly repetitive dnas specific for the t1 chromosome of tobacco. Chromosoma 106: 369-79.

Matylla-Kulinska K, Tafer H, Weiss A, and Schroeder R. 2014. Functional repeat-derived rnas often originate from retrotransposon-propagated ncrnas. Wiley Interdisciplinary Reviews. RNA 5: 591-600.

Mehrotra S, Goel S, Raina SN, and Rajpal VR. 2014. Significance of satellite DNA revealed by conservation of a widespread repeat DNA sequence among angiosperms. Appl Biochem Biotechnol 173: 1790-801.

Mehrotra S, and Goyal V. 2014. Repetitive sequences in plant nuclear DNA: Types, distribution, evolution and function. Genomics, Proteomics & Bioinformatics 12: 164-171.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, *et al.* 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. Genome Biol 14: R10.

Mendes S, Moraes AP, Mirkov TE, and Pedrosa-Harand A. 2011. Chromosome homeologies and high variation in heterochromatin distribution between citrus l. And poncirus raf. As evidenced by comparative cytogenetic mapping. Chromosome Research 19: 521-530.

Michael TP, and Jackson S. 2013. The first 50 plant genomes. Plant Genome 6.

Mondin M, Santos-Serejo JA, Bertao MR, Laborda P, Pizzaia D, and Aguiar-Perecin ML. 2014. Karyotype variability in tropical maize sister inbred lines and hybrids compared with kys standard line. Front Plant Sci 5: 544.

Nagaki K, Tsujimoto H, and Sasakuma T. 1998. Dynamics of tandem repetitive afa-family sequences in triticeae, wheat-related species. J Mol Evol 47: 183-9.

Nosaka M, Ishiwata A, Shimizu-Sato S, Ono A, Ishimoto K, Noda Y, *et al.* 2013. The copy number of rice cacta DNA transposons carrying mir820 does not correlate with mir820 expression. Plant Signal Behav 8.

Obae SG. 2012. Nuclear DNA content and genome size of american ginseng.

Ouyang S, and Buell CR. 2004. The tigr plant repeat databases: A collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32: D360-3.

Paço A, Adega F, Meštrović N, Plohl M, and Chaves R. 2014. Evolutionary story of a satellite DNA from phodopus sungorus (rodentia, cricetidae). Genome Biology and Evolution 6: 2944-2955.

Palomeque T, and Lorite P. 2008. Satellite DNA in insects: A review. Heredity 100: 564-573.

Park HJ, Kim DH, Park SJ, Kim JM, and Ryu JH. 2012. Ginseng in traditional herbal prescriptions. J Ginseng Res 36: 225-41.

Pezer Z, Brajkovic J, Feliciello I, and Ugarkovc D. 2012. Satellite DNA-mediated effects on genome regulation. Genome Dyn 7: 153-69.

Pezer Z, Brajkovic J, Feliciello I, and Ugarkovic D. 2011. Transcription of satellite dnas in insects. Prog Mol Subcell Biol 51: 161-78.

Plohl M, Luchetti A, Mestrovic N, and Mantovani B. 2008. Satellite dnas between selfishness and functionality: Structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409: 72-82.

Plohl M, Mestrovic N, and Mravinac B. 2012. Satellite DNA evolution. Genome Dyn 7: 126-52.

Project AG. 2013. The amborella genome and the evolution of flowering plants. Science 342.

Puertas MJ, and Villasante A 2013. Is the heterochromatin of meiotic neocentromeres a remnant of the early evolution of the primitive centromere? *Plant centromere biology*. Wiley-Blackwell.

Radic MZ, Lundgren K, and Hamkalo BA. 1987. Curvature of mouse satellite DNA and condensation of heterochromatin. Cell 50: 1101-8.

Rozen S, and Skaletsky HJ. 1998. Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Sharma A, Wolfgruber TK, and Presting GG. 2013. Tandem repeats derived from centromeric retrotransposons. BMC Genomics 14: 142.

Shi F-X, Li M-R, Li Y-L, Jiang P, Zhang C, Pan Y-Z, *et al.* 2015. The impacts of polyploidy, geographic and ecological isolations on the diversification of panax (araliaceae). BMC Plant Biology 15: 297.

Solovyev V, Kosarev P, Seledsov I, and Vorobyev D. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol 7 Suppl 1: S10 1-12.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, *et al.* 2015. Nested radiations and the pulse of angiosperm diversification: Increased diversification rates often follow whole genome duplications. New Phytol 207: 454-67.

Tek AL, Song J, Macas J, and Jiang J. 2005. Sobo, a recently amplified satellite repeat of potato, and its implications for the origin of tandemly repeated sequences. Genetics 170: 1231-1238.

Ugarkovic D. 2005. Functional elements residing within satellite dnas. EMBO Rep 6: 1035-9.

Vlahovicek K, Kajan L, and Pongor S. 2003. DNA analysis servers: Plot.It, bend.It, model.It and is. Nucleic Acids Res 31: 3686-7.

Vourc'h C, and Biamonti G. 2011. Transcription of satellite dnas in mammals. Prog Mol Subcell Biol 51: 95-118.

Walsh JB. 1987. Persistence of tandem arrays: Implications for satellite and simple-sequence dnas. Genetics 115: 553-67.

Waminal N, Park HM, Ryu KB, Kim JH, Yang TJ, and Kim HH. 2012. Karyotype analysis of panax ginseng c.A.Meyer, 1843 (araliaceae) based on rdna loci and dapi band distribution. Comparative Cytogenetics 6: 425-441.

Waminal NE, Choi H-I, Kim N-H, Jang W, Lee J, Park JY, *et al.* 2016a. A refined panax ginseng karyotype based on an ultra-high copy 167-bp tandem repeat and ribosomal dnas. Journal of Ginseng Research.

Waminal NE, Perumal S, Lee J, Kim HH, and Yang T-J. 2016b. Repeat evolution in brassica rapa (aa), b. Oleracea (cc), and b. Napus (aacc) genomes. Plant Breeding and Biotechnology 4: 107-122.

Waminal NE, Perumal S, Lim K-B, Park B-S, Kim HH, and Yang T-J 2015. Genomic survey of the hidden components of the b. Rapa genome. *The brassica rapa genome.* Springer.

Wicker T, Guyot R, Yahiaoui N, and Keller B. 2003. Cacta transposons in triticeae. A diverse family of high-copy repetitive elements. Plant Physiol 132: 52-63.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, *et al.* 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973-982.

Xiong Z, and Pires JC. 2011. Karyotype and identification of all homoeologous chromosomes of allopolyploid brassica napus and its diploid progenitors. Genetics 187: 37-49.

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

Zabala G, and Vodkin L. 2007. Novel exon combinations generated by alternative splicing of gene fragments mobilized by a cacta transposon in glycine max. BMC Plant Biol 7: 38.

Zabala G, and Vodkin LO. 2014. Methylation affects transposition and splicing of a large cacta transposon from a myb transcription factor regulating anthocyanin synthase genes in soybean seed coats. PLoS One 9: e111959.

CHAPTER III

Panax ginseng genome structure and evolution revealed by cytogenomics of major TEs and genic blocks

ABSTRACT

Transposable elements (TEs) make up a considerable proportion of plant genomes. Whole genome duplications (WGD) often drive bursts and accumulation of different TE families in different species as well as duplication of single-copy genic blocks, which usually complicates genome assemblies of polyploid species. Two rounds of WGDs are known to have shaped the extant *Panax ginseng* (ginseng) genome. Here, I conducted a molecular cytogenetic mapping of major TE families and selected genic blocks of ginseng in order to understand the chromosomal distribution of different ginseng TE families as well as to chromosomally investigate the recent WGD. Different subfamilies of ginseng *Ty3/Gypsy* family showed differential chromosome hybridization. *PgDel1* hybridized to the entire chromosome lengths, while *PgDel2* and *PgDel5* showed subgenomic pericentromeric distribution in both ginseng, and *PgTat* was preferentially localized at subtelomeric regions. On the other hand, *PgTork* of the *Ty1/Copia*, family hybridized on the pericentromeric regions. Amplification of genic regions from putative adjacent assembly scaffolds, Pg_scaffold0266 and Pg_scaffold2259, with corresponding putative paralogous scaffolds, Pg_scaffold0762 and Pg_scaffold0978, implicated from the current version of the ginseng genome assembly, revealed one to four gel bands. FISH analysis also revealed duplicated genic blocks evidence of WGD, and while supporting the proper assembly of two adjacent scaffolds, Pg_scaffold0266 and Pg_scaffold2259, also indicated disjunct location of Pg_scaffold0762 and Pg_scaffold0978 from each other despite being paralogous to Pg_scaffold0266 and Pg_scaffold2259, respectively. This cytogenetic information supports the allotetraploid origin of the ginseng genome as well as provided a cytogenetic support on the assembly status of the current ginseng genome.

INTRODUCTION

Whole genome duplications (WGDs) are ubiquitous in flowering plants, promoting species diversification (Tank *et al.* 2015b). Repetitive elements (REs) are important players in genome reorganization and stabilization during and after WGD events that disrupt nuclear homeostasis (Fedoroff 2012b). Subsequent genomic rearrangements to stabilize the genome often results in elimination of large DNA segments, sometimes biased towards one parental genome (Renny-Byfield *et al.* 2012). Comparative measurement of the abundance of different TE families among related species often reveals TE dynamics and provide understanding of history of a genome.

REs constitute a considerable genomic proportions in most angiosperm species, even reaching up to 85% of a genome (Michael and Jackson 2013), and genome size variations in most organisms are often attributed to REs (Michael and VanBuren 2015), which influence genome architecture, diversity and evolution via homologous recombination and chromosome rearrangements such as duplication, deletion, inversion, and translocation (Parisod *et al.* 2010; Lisch 2013; Choi *et al.* 2014; Sigman and Slotkin 2016). In addition, different chromosomal distribution of different TE families have been observed in different species (Higashiyama *et al.* 1997; Kubis *et al.* 1998; Lim *et al.* 2007; Ma *et al.* 2007; Wolfgruber *et al.* 2009; Qi *et al.* 2013) which is often associated with heterochromatin function and TE dynamics.

Based on their transposition mechanisms, TEs are classified into two major classes*:* I, retrotransposons, and II, DNA transposons (Wicker *et al.* 2007). Retrotransposons, especially those belonging to the *Gypsy* and *Copia* families, occupy a major fraction of most plant genomes (Tenaillon *et al.* 2010; Macas *et al.* 2015). In some cases, a major proportion of the genome is made up of only a few retrotransposon families; for example, *Del* subfamily of the

Ty3/Gypsy family retrotransposons occupy about 30% of the 3.6 Gb ginseng genome (Choi et al. 2014). Different *Del* subfamilies were distributed in different chromosomal regions and subgenome, supporting the allotetraploid origin of the ginseng genome.

Single-copy DNA segments are powerful tools in chromosomally mapping a gene locus (Khrustaleva and Kik 2001), identifying individual chromosomes (Lou *et al.* 2014), as well as verifying a genome assembly (Chamala *et al.* 2013). These are often difficult to achieve with REs due to numerous homologous regions in a genome.

Here, I cytogenetically mapped the major TE families identified in the ginseng genome, as well as TE-deficient single-copy genomic regions, in order to have a cytogenetic perspective about the allotetraploid origin of the ginseng genome and to demonstrate the usefulness of single-copy genic blocks to validate genome scaffold assembly. Information here will be beneficial to the holistic understanding of the genome history of ginseng, as well as facilitate genome scaffold assembly.

MATERIALS AND METHODS

Root sample preparation

Stratified seeds of three ginseng cultivars 'Sunun' were provided by the Korea Ginseng Corporation (KGC) Natural Resources Research Institute (Daejeon, Korea). Stratified seeds were allowed to germinate in petri dishes with wet filter papers at 10-15ºC. The root meristems were then excised (about 2 cm from the root tips), pretreated with 0.002M 8-hydroxyquinoline for 5 hours at 18ºC, fixed in 90% acetic acid for 15 min at room temperature (RT, ~24ºC), and then stored in 70% ethanol until use.

PCR amplification of probes

TE probes were amplified according to primers from Choi *et al.* (2014). Genic probes from Pg_scaffold0266 and Pg_scaffold2259 were designed by identifying RE-deficient regions according to the current ginseng genome annotation visualized using JBrowse (Skinner *et al.* 2009). Primers used for the genic block amplification are listed in Table 3-1. Amplifications were done using either *ExTaq* (Takara Bio Inc, RR001A) or Quick Taq HS Dye Mix (Toyobo, DTM-101) according to annealing temperatures of primers.

Fluorescence *in situ* hybridization (FISH) analysis

Root mitotic chromosome spreads were prepared from stratified seeds provided by the Korea Ginseng Corporation Natural Resources Research Institute (Daejeon, Korea), according to the methods of Waminal *et al.* (2012). *PgDel1*, *PgDel2*, *PgTat1*, and *PgTork* amplicons ethanol purified and were labeled with either Alexa Fluor 488-5-dUTP (Invitrogen, C11397) or Diethyl amino coumarin-5-dUTP (Perkin Elmer, NEL455001EA), and *PgDel5* amplicons were labeled with Texas Red-5-dUTP (Perkin Elmer, NEL417001EA). PCR amplicons form genic regions of

Pg_scaffold0266 and Pg_scaffold2259 were pooled and labeled with Alexa Fluor 488-5-dUTP and Texas Red-5-dUTP, respectively. All probes were labeled through direct nick translation. FISH procedures were done according to Waminal *et al.* (2012). The Pg167TR was labeled with Texas Red-5-dUTP (Perkin Elmer, NEL417001EA). For directly labeled probes and oligoprobes, slides were immediately used for FISH after fixation with 4% paraformaldehyde, without subsequent pepsin and RNase pretreatment. Images were captured with an Olympus BX53 fluorescence microscope equipped with a Leica DFC365 FS CCD camera, and processed using Cytovision ver. 7.2 (Leica Microsystems, Germany). We performed further image enhancements and FISH-based estimation of Pg167TR genomic proportion in Adobe Photoshop CC.

Table 3-1. List of primers used to amplify genic blocks from two contiguous scaffolds Pg_scaffold2259 and Pg_scaffold0266.

Gene information						Primer information						
gene	source	start	end	length	strand	N0.	seq	mer	seq	mer	Tm	size
1	Pg_scaffold2259	2199	8176	5978	-	1	AGATGCCAGTTTTTCCTTTCTGC	23	ATTTGACTTGCACAACCCTTAGC	23	60	2607
2	Pg_scaffold2259	29241	33975	4735	+	2	CGCAAATTAAAAATCCGAGCTGC	23	TGTGCATCCTATATCCTGTGACG	23	60	3041
3	Pg_scaffold2259	35980	37943	1964	-	3	ACCAAGGGTTACCGCTACAAG	21	CTGGTTTATCAGGGCAGCAGA	21	60	1471
4	Pg_scaffold2259	76342	85183	8842	-	4	GCTATCGGTAAATCAGCTCCTGA	23	AAGAATGACAGCAAAGTTCGGTG	23	60	3979
5	Pg_scaffold2259	89215	92793	3579	-	5	GTCAGGGAGCTCAAAGACGAA	21	AGCATTGTAGCGACCCTTTGA	21	60	3061
6	Pg_scaffold2259	108649	123882	15234	+	6	ACATGCATTGTACCGGTCAGT	21	ACGGCTCTTTTTGGTCATTGC	21	60	3302
						7	GTTGTCTGGCGAAACCTTTCC	21	GCAGCTTCAAATATGCGAGGG	21	60	3129
7	Pg_scaffold2259	208283	211837	3555	-	8	CAGCCAACATGACCAACATCG	21	CTACACTTGATGGGCACTGGT	21	60	3155
8	Pg_scaffold2259	235573	241373	5801	+	9	GGTCTTCGGACTGCTTCTTCA	21	TTTGCTGTGTGGCGATTGAAG	21	60	3905
9	Pg_scaffold2259	262980	264515	1536	+	10	TCGTTATGGGTGTCAGTGGTG	21	ATGCTTTCCTTCTGCTGCTCT	21	60	1206
10	Pg_scaffold2259	282734	284949	2216	+	11	TGGCTACTGATCTCAAAGGTAAT	23	AATGGCCTGAGGACTTTGTGT	21	58	1505
11	Pg_scaffold2259	306027	308723	2697	+	12	ATGGCCGCTAATTCGTACAGT	21	CCTGCAGCAGAATTTTCCTCG	21	60	2600
12	Pg_scaffold2259	322009	325292	3284	-	13	GCAGCAATGGTTGATTTGGGT	21	AAAATGTTGCTGTTCGAGGGC	21	60	2882
13	Pg_scaffold2259	356149	357911	1763	+	14	TGAAAGTTAGGGCTGGCAGAG	21	GACTTGAGTTGCTGTTGCCAG	21	60	1246
14	Pg_scaffold2259	360776	365377	4602	+	15	GGTGTCGCTACTTCCTTGGAA	21	CCAACCGACTCTGTAGTTGCT	21	60	3621
15	Pg_scaffold2259	397924	401997	4074	+	16	ATGCTGCTTCTTTCCTTCCGA	21	CTGCCAACTTTTCTCACCTGC	21	60	3480
16	Pg_scaffold2259	403893	406389	2497	-	17	GCTTCTGGTGTGTGTGTGAAC	21	TCTCCTCCACACCAGCAATTC	21	60	2336
17	Pg_scaffold2259	407760	412439	4680	-	18	CCACGTTATTCCCACCACAGA	21	TTGCTGACCAATTGAGTTGCG	21	60	3982
18	Pg_scaffold2259	430735	434948	4214	-	19	GGGACGGAAAGCAAGTAGTGA	21	TCAGATAGGGCAGGTCCTGAA	21	60	3128
			432749	81251					49275			53636
1	Pg_scaffold0226	48370	52372	4003	-	1	AGGTACCTGGGCTAGCTGTAA	21	GTGTTCAGCTCCCGAAGAAGA	21	60	3878
2	Pg_scaffold0226	61440	72161	10722	-	2	TGGACATCTGGAAGGGAAACG	21	TTGCCGAGTGAGATTGAACGA	21	60	3582
3	Pg_scaffold0226	271395	278413	7019	-	3	CTTTCTGGGTTCTGCAAGTGC	21	GTTGGAGCAAGTGAAGCACTG	21	60	3521
						4	GATGCATTGGTTCATCTGCCC	21	GGCTACACGATTCCAATTCGC	21	60	3709
4	Pg_scaffold0226	661951	666917	4967	+	5	CCAAATATCCCTTGGCATGCG	21	AGTTGCTGCAGGCTTGATTTG	21	60	3709
5	Pg_scaffold0226	703685	708191	4507	-	6	GACTTGCAGCTTCCTTGTGTG	21	AACAGCAGCTAGAGGGTTCAC	21	60	4000
6	Pg_scaffold0226	892358	894193	1836	-	7	GGTTAAATGACCAGGCCCTCA	21	ACTTGGTTTCATGGGCATTGC	21	60	1493
7	Pg_scaffold0226	1043589	1048937	5349	+	8	CTGTCGAGTAGCAGCGATCAT	21	CATCAAGGCCCAAGGGTACTT	21	60	3386
8	Pg_scaffold0226	1053113	1057725	4613	+	9	GCTCCTCCAGATCATTCTCCG	21	ATTGCAAATGACACCAGCACC	21	60	3116
9	Pg_scaffold0226	1166070	1169628	3559	-	10	TATTGCAGTTCTGGGTTGCCT	21	CCTCGTTGAAGTAATGGGGCT	21	60	3055
10	Pg_scaffold0226	1178834	1182400	3567	+	11	CATCACCAGCAGCTCCTAGAG	21	GGTCCATCCATGTCAAGCAGA	21	60	3276
Total			1134030	50142					29900			36725

RESULTS

Different TE families localized in different chromosomal regions

Cytogenetic mapping of major *P. ginseng* TEs revealed hybridization of different repeat families to different chromosomal niches (Figs. 3-1, 3-2, 3-3). *PgDel1* hybridized to the entire chromosomes, supporting their predominant abundance in the ginseng genome (Choi *et al.* 2014). However, *PgDel2* hybridized to the pericentromeric area of only 24 out of the 48 chromosomes, showing subgenome preference. Comparative FISH analysis with *P. quinquefolius* revealed a similar subgenome hybridization (Fig. 4-3). *PgTat* elements preferred to localize in the subtelomeric regions, and *PgTork* mostly hybridized at pericentromeric regions. *PgDel5* hybridized to pericentromeric regions of chromosomes with weak *PgDel2* signals, indicating a subgenome distribution, but to the opposite subgenome (Fig 3-4). Taking into account the distribution of the satDNA Pg167TRb and the major ginseng TEs, it can be observed that half of the *PgDel2*-rich chromosomes and half of the *PgDel2*-poor chromosomes hybridize with Pg167TRb. In addition, *PgDel2*-rich chromosomes are generally shorter than those that are *PgDel2*-poor (Figs. 3-1, 3-2, 3-5).

PgDel1
PgDel2
PgDel5
PgTork
PgTat1
PgTat2

Fig. 4-1. Chromosomal distribution of major *P. ginseng* REs in *P. ginseng* chromosomes. Different repeat families hybridized to different chromosomal regions. *PgDel1* localized to all chromosomal regions, *PgDel2* in half of the subgenome, *PgDel5* in regions with low *PgDel2* abundance, *PgTork* in interstitial regions, *PgTat1* in subtelomeric regions, and *PgTat2* in pericentromeric regions. Bar = 10 μm.

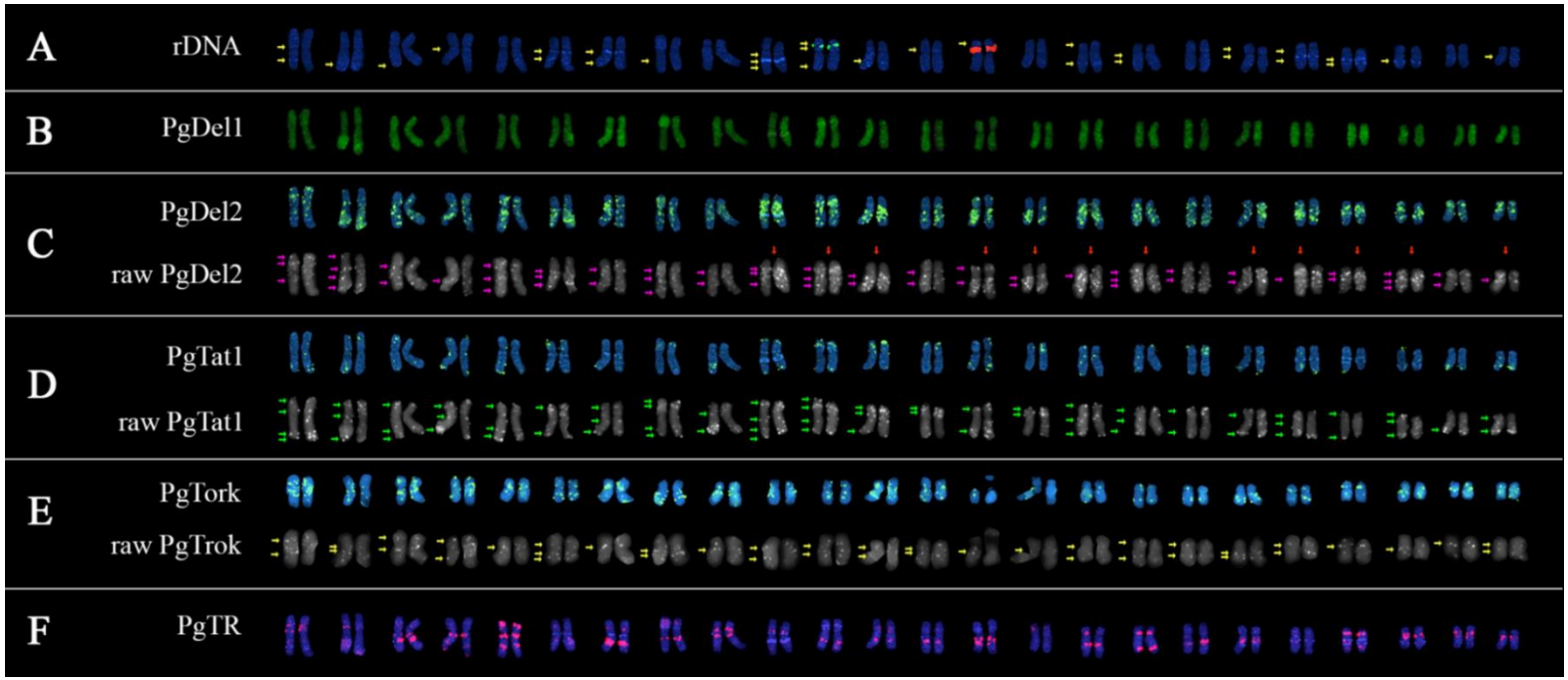

A
rDNA
B
PgDel1
C
PgDel2
raw PgDel2
D
PgTat1
raw PgTat1
E
PgTork
raw PgTrok
F
PgTR

Fig. 4-2. Karyogram of *P. ginseng* with chromosomal distribution of major ginseng REs. A) *P. ginseng* karyogram based on the distribution of the rDNA, and LTR retrotransposon probes (B, PgDel2; C, PgTat1; and D, PgTork), and E) Pg167TR tandem repeat along the *P. ginseng* genome. The PgDel2 localized mostly at the intercalary and the pericentric area (B, pink arrows), while the PgTat1 localized mostly at the subtelomeric regions (green arrows). PgTork, although doesn't have chromodomains is shown to hybridize at the centromeric regions in addition to the euchromatic regions. Pink arrows in panel d indicate chromosomes with centromeric probe hybridization while yellow arrows indicate the other observed PgTork signals. Red arrows indicate the 12 pairs with relatively more intense PgDel2 signals than the other pair. Scale bar, 5 μm.

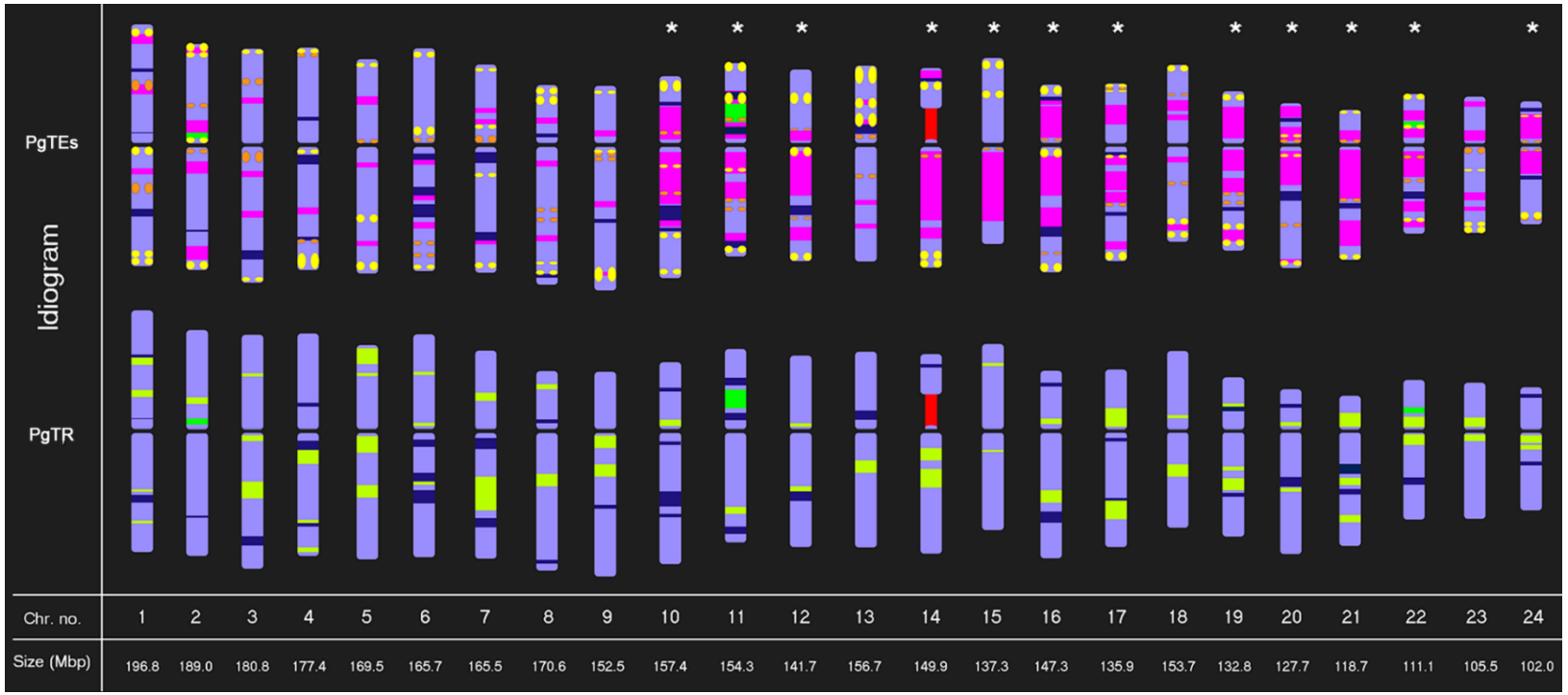


Fig. 4-3. *P. ginseng* FISH karyotype idiogram based on LTR retrotransposons (upper panel) and PgTR (lower panel). 5S and 45S rDNA are indicated by green and red bars, respectively, DAPI bands by dark blue bars. **(Upper panel)** PgDel2: pink, PgTat1: yellow ovals, PgTork: orange ovals. **(Lower panel)** PgTR distribution: mint green. Chromosome sizes are based on the 3.6 Gb genome size of ginseng.

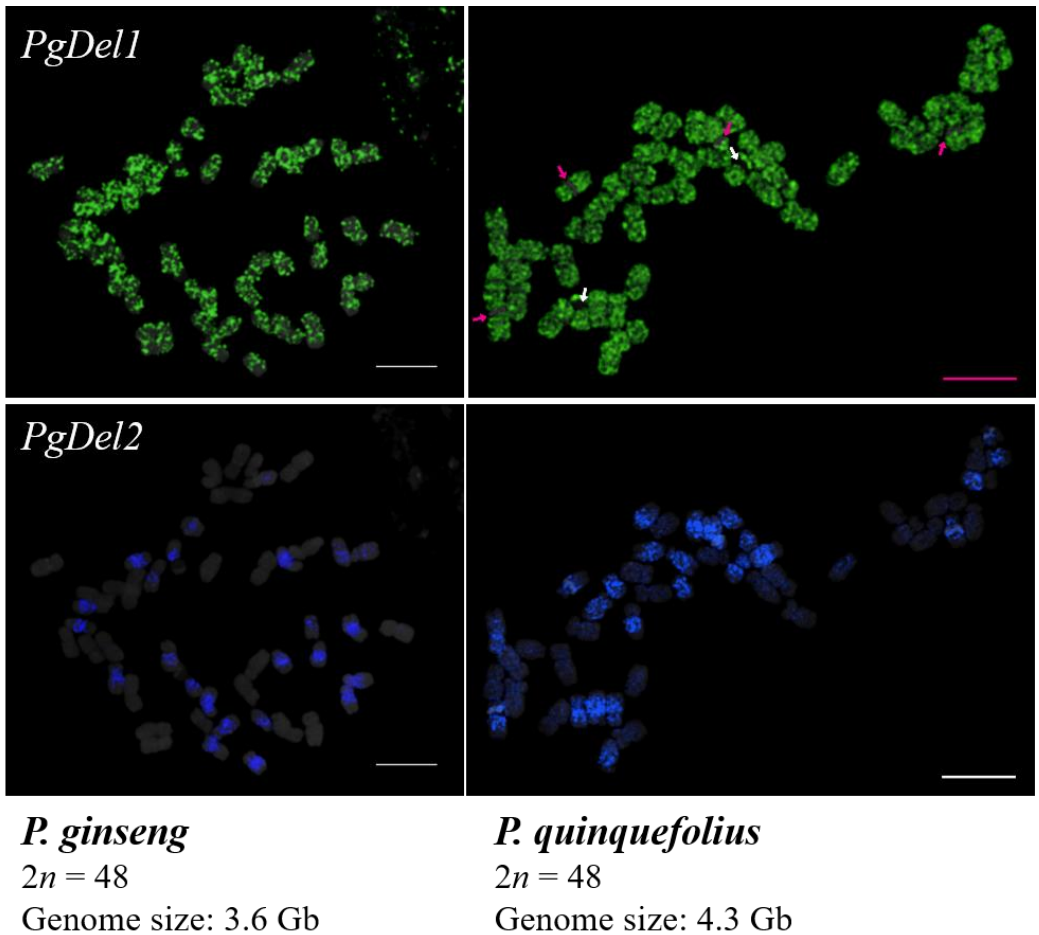


Fig. 4-4. Comparative cytogenetic mapping of *PgDel1* and *PgDel2* between *P. ginseng* and *P. quinquefolius*. *PgDel1* was distributed across all chromosomes, but not in 45S rDNA region (white arrows) and DAPI bands (pink arrows) in *P. quinquefolius*. *PgDel2* hybridized to 12 out of 24 homologous chromosome pairs. Bar = 10 μm.

Idiogram
Chr. no.
Size (Mbp)
196.8 189.0 180.8 177.4 169.5 165.7 165.5 170.6 152.5 157.4 154.3 141.7 156.7 149.9 137.3 147.3 135.9 153.7 132.8 127.7 118.7 111.1 105.5 102.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Reorganized
13 18 23
10 11 12 14 15 16 17 19 20 21 22 24

Fig. 4-5. Karyotype idiogram of *P. ginseng* showing repetitive elements previously described as well as the Pg167TR elements. Blue, green, red, and yellow bars indicate DAPI, 5S rDNA, 45S rDNA, and Pg167TR bands. Pg167TR bands with red borders indicate Pg167TRb. Purple and stroked chromosomes represent Pg167TR and *PgDel2*-rich (Choi et al, 2014) chromosomes. Pg167TRb loci localized in six out of 12 *PgDel2*-rich ginseng chromosomes. This brings the possibility that the ginseng genome was derived from ancient genome with six as basic chromosome number. Lower panel: After rearranging the chromosome based on the presence or absence of *PgDel2*, it is more apparent that chromosomes bearing the *PgDel2* LTR retrotransposons are generally shorter than those without *PgDel2*. Bars= 10 μm .

Contiguous scaffolds have paralogous sequences in disjunct chromosomal regions

A total of ten and 18 genic regions were targeted and 11 and 19 PCR amplicons were generated from Pg_scaffold2259 and Pg_scaffold0266, respectively because some genes were amplified with two sets of primers (Table 3-1, Fig. 4-6). Due to abundant repetitive elements distributed in Pg_scaffold0266, only ten genes were identified which were loosely distributed along a span of 1,134 kb. On the other hand, the fewer repeat elements in Pg_scaffold2259 allowed identification of 18 genes that were densely distributed in a short span of 433 kb (Fig. 4-6A). Out of the 11 and 19 amplicons, two elements amplicons that showed intense bands of unexpected sizes were not used in the pooling of sequences for FISH probe. Consequently, only nine and 17 amplicons were pooled totaling 30 and 49 kb for Pg_scaffold0266 and Pg_scaffold2259, respectively (Fig 3-6B). These genic regions were designed from paralogous scaffolds that were identified through a zigzag assembly approach (Kim, 2015, dissertation), so I expected to observe two co-localized green and red loci (representing Pg_scaffold0266 and Pg_scaffold2259, respectively, see Materials and Methods) if the assembly is correct, assuming contiguous Pg_scaffold0266 and Pg_scaffold2259 assembly as well as the paralogous scaffolds of Pg_scaffold0762 and Pg_scaffold0978 (Fig. 4-6C).

FISH analysis of pooled genic blocks revealed two important things. First, the detection of regions paralogous to Pg_scaffold0266 and Pg_scaffold2259, and second, while Pg_scaffold0266 and Pg_scaffold2259 could be two contiguous scaffolds as the assembly predicts, Pg_scaffold0762 and Pg_scaffold0978 may not be contiguous, instead localized in different chromosomal regions.

Strong signals were observed for probes derived from Pg_scaffold2259 compared with those from Pg_scaffold0266, reflecting the quantity of amplicons included in each respective pools of probes (Fig. 4-7, 3-8). While very little background signals were observed in Pg_scaffold2259, there were more in

Pg_scaffold0266. This suggests amplification of RE fragments inserted in some of the amplified regions that were not masked in the JBrowse annotation.

In addition to the two more intense signals of Pg_scaffold0266, a very weak signal was observed at the locus matching the location of one Pg_scaffold2259, indicating contiguous portion of Pg_scaffold0266 and Pg_scaffold2259 that is paralogous to Pg_scaffold0978, breaking at the 3' region of Pg_scaffold0266 (Fig. 4-6).

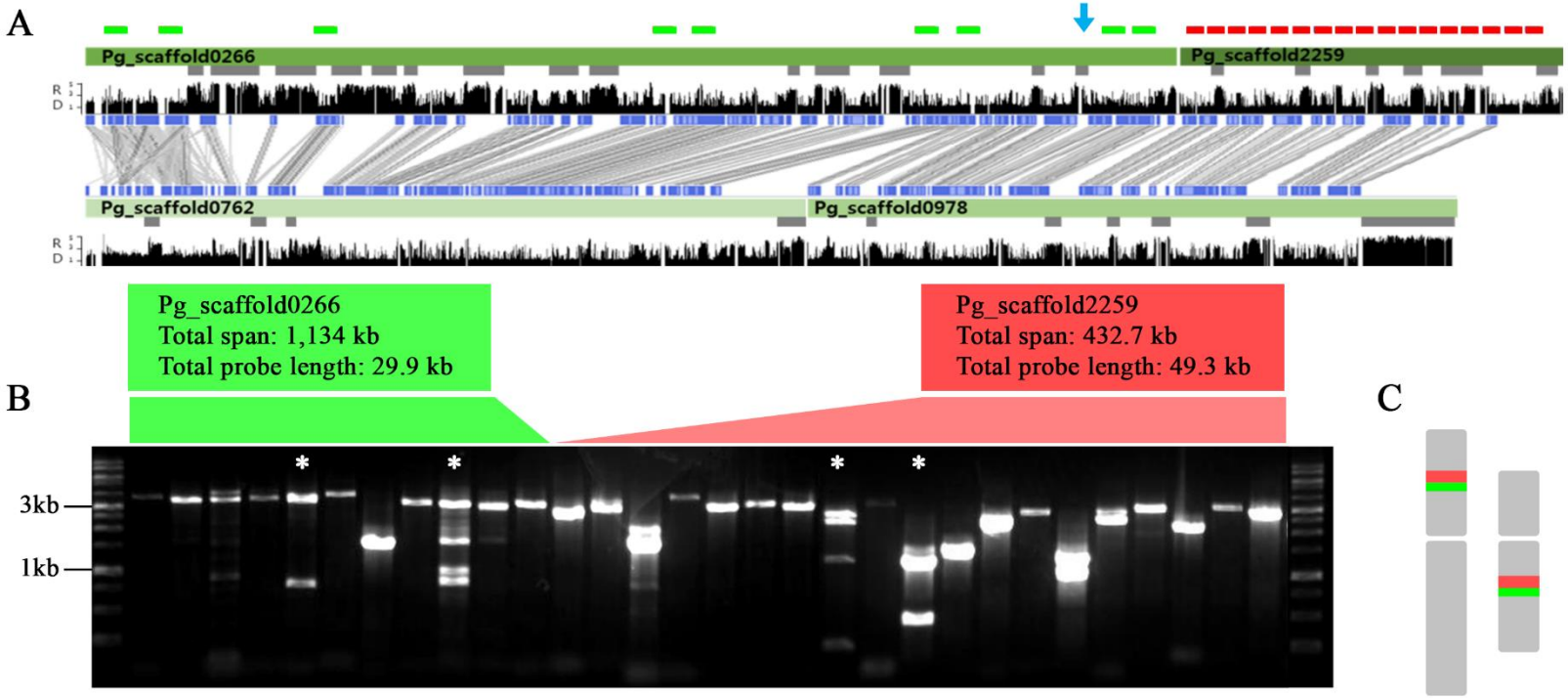

A
Pg_scaffold0266
Pg_scaffold2259
Pg_scaffold0762
Pg_scaffold0978
Pg_scaffold0266
Total span: 1,134 kb
Total probe length: 29.9 kb
Pg_scaffold2259
Total span: 432.7 kb
Total probe length: 49.3 kb
B
3kb
1kb
C

Fig. 4-6. Amplification of genic regions from two adjacent assembled scaffolds. A) Assembly of recent and ancient duplicated blocks with corresponding assembly scaffold names. Red and blue bars designate arbitrary PCR amplification regions. Blue arrow indicates the possible scaffold mis-assembly breakpoint. B) amplification of 11 and 19 genic regions from Pg_scaffold0266 and Pg_scaffold2259, respectively. Several targets showing multiple bands, may indicate paralogous loci with several indel variations, or RE fragment insertion. C) Chromosome diagram showing the predicted FISH signals if the scaffold assembly in the current version is correct.

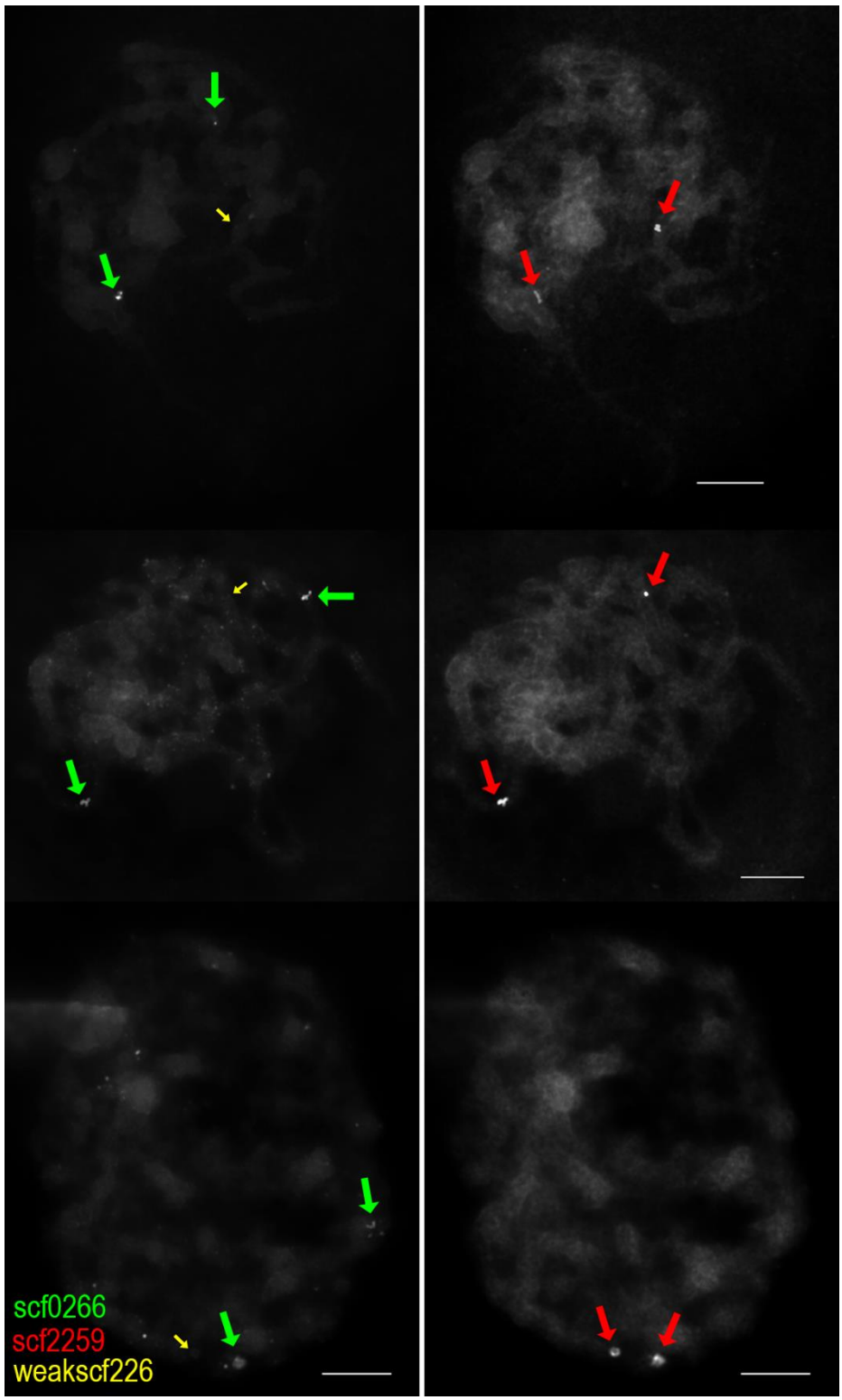

scf0266
scf2259
weakscf226

Fig. 4-7. Chromosomal mapping of genic regions from two adjacent contiguous scaffolds. Two intense clustered signals were observed for each pooled genic probes from Pg_scaffold0266 and Pg_scaffold2259, indicating duplicated genic blocks from each scaffold. In addition, a very weak Pg_scaffold0226 signal linked with Pg_scaffold2259 locus (yellow arrow), supports their linkage in one paralogous site.

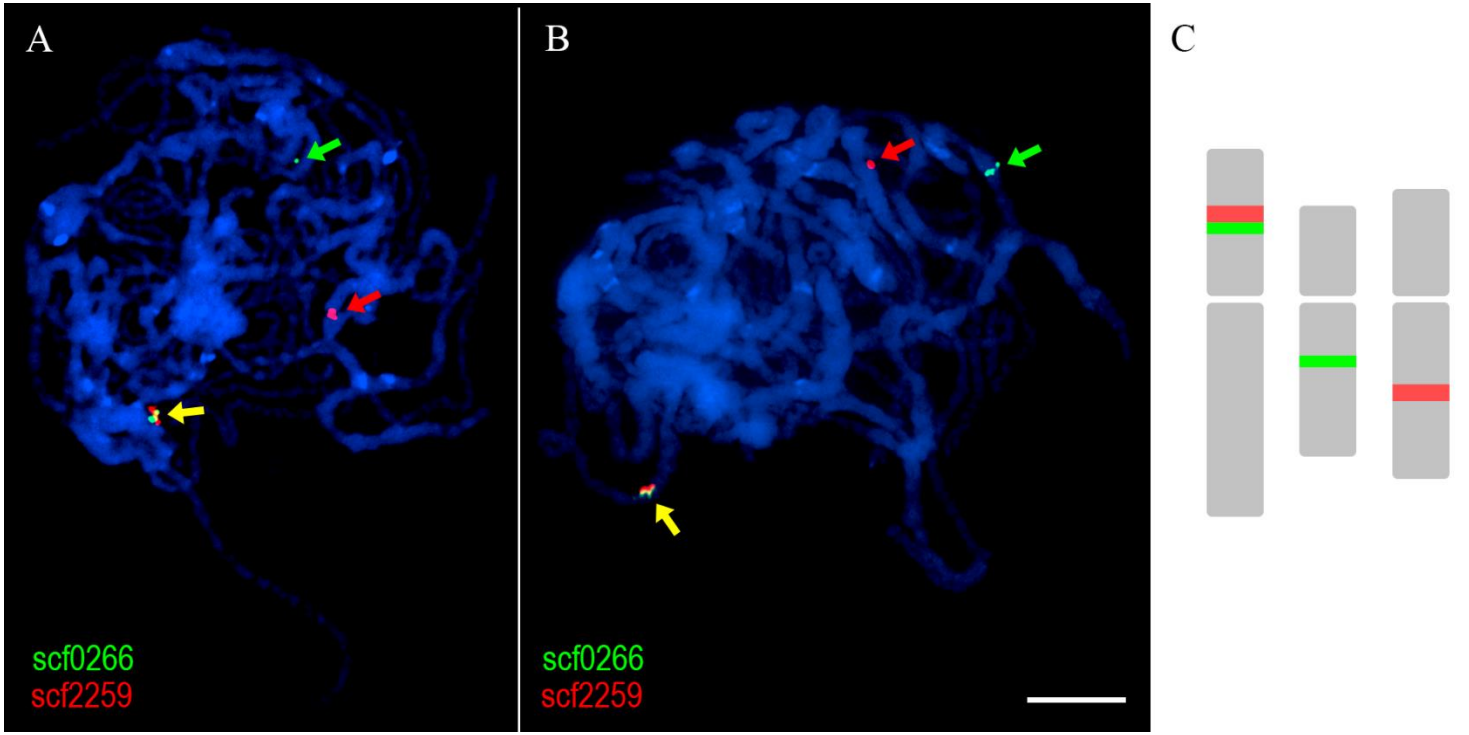


Fig. 4-8. FISH mapping of paralogous genic blocks. Pseudo-colored pachytene images from Fig. 4-7, showing separate and colocalization of pooled genic probes in different chromosomal regions. Bar = 10 µm.

DISCUSSION

Cytogenetic mapping of major ginseng TE supports allopolyploid origin of the ginseng genome

The usefulness of FISH for localizing TEs and measuring genomic abundance of REs has been demonstrated is some crops (Macas *et al.* 2007; Bilinski *et al.* 2014). Preferential hybridization of TEs to different chromosomal regions are often directed by epigenetic mechanisms (Neumann *et al.* 2011). *PgDel* elements contain chromodomain motifs (Choi *et al.* 2014) that have chromatin-targeting functions (Gao *et al.* 2008; Neumann *et al.* 2011) which could help explain their preferential targeting at heterochromatin regions like pericentromeres (Choi *et al.* 2014).

Like many other plant species that revealed a single TE family being predominantly amplified within their genomes (Kelly *et al.* 2015; Macas *et al.* 2015), ginseng showed high abundance of the *PgDel* subfamily. More importantly, the distribution of the *PgDel2* subfamily to a subgenome that corresponds to half of the ginseng chromosome complement suggests an allotetraploid origin of the ginseng genome. A similar observation of a *Ty1/Copia* and *CACTA* elements in the allotetraploid *Brassica napus* enabled identification of subgenomes derived from *B. oleracea* (Waminal *et al.* 2016). Furthermore, this shared feature of subgenome distribution of *PgDel2* with *P. quinquefolius* suggests a common allotetraploidization event prior to their geographic isolation and subsequent speciation. TEs play an important role in adaptation of species to environmental changes, even in microenvironments (Kalendar *et al.* 2000; Fedoroff 2012a). This nuances in TE content between the two allotetraploids may have facilitated their adaptation to their present environments.

WGDs and subsequent diploidization often results to diversification of species (Tank *et al.* 2015a). The 1st (ancient) and 2nd (recent) rounds of WGDs undergone by ginseng's ancestral genome have shaped the extant ginseng genome. Analysis of the

dynamic of its REs provide insights about the evolutionary route it has navigated. While chromosome numbers and TE hybridization pattern taken individually could suggest an ancestral genome with 24 chromosomes, a combination of cytogenetic markers could provide more hints about the more distant events in the evolutionary pathway of ginseng. Taking into account the distribution of Pg167TRb and *PgDel2*, as well as the conserved chromosome number within the genus *Panax*, provide a support for a more ancient diploid ancestral genome with a nuclear chromosome number of $2n = 12$, supporting an ancient base chromosome number $x = 6$ (Yi *et al.* 2004).

This scenario suggests an ancestral genome with $2n = 12$ and low-copy *PgCACTA1* and Pg167TR elements in a library of RE [plohl2012]. After the 1st WGD and subsequent diploidization, chromosome numbers doubled to $2n = 24$ while *PgCACTA1* elements remained in low copy but Pg167TR have diverged into Pg167TRa and Pg167TRb but in low copy (Fig. 4-9). Divergence of related species outside the genus *Panax*, such as *Aralia* species retained low copies of *PgCACTA1* and Pg167TR but favored homogenization to Pg167TRb variant over Pg167TRa. Meanwhile, species preceding *Panax* favored amplification of Pg167TRa but in low copies, are observed in diploid *Panax* species. Eventually, the 2nd WGD which is most likely an allotetraploidization event between related genomes, one replete with *PgDel2* elements and another with *PgDel5* but in lesser abundance, spurred the amplification of *PgCACTA1* elements with concomitant proportional amplification of Pg167TR sequence variants.

Chromosomes carrying *PgDel2* retrotransposon were shown to be generally shorter than those without *PgDel2*. This suggests biased genomic elimination between related but incompatible genomic segments, similar to those observed in *Nicotiana tabacum* (Kovarik *et al.* 2012; Renny-Byfield *et al.* 2012; Renny-Byfield *et al.* 2013). This differential elimination of parental genome is often observed in allotetraploids, further corroborating the allotetraploid origin of the ginseng genome.

Cytogenetic mapping of pooled gene blocks supported a tetraploid ginseng genome and validated the contiguity of two assembly scaffolds

Pooled genic regions used as FISH probes has more advantages in identifying chromosomes and chromosomal segments compared with TE probes, (Huang *et al.* 2009; Lou *et al.* 2014). Furthermore, genic regions are more reliable in validating scaffold assemblies than TEs since TEs contain highly homologous sequences from other genomic regions, making it non-specific (Chamala *et al.* 2013).

Pooling of genic regions have facilitated identification of individual chromosomes in *Cucumis sativus* (Lou *et al.* 2014) and BAC probes for identification and validation of genome assembly of *Amborella* (Chamala *et al.* 2013).

The pooled genic regions from Pg_scaffold0266 and Pg_scaffold2259 revealed two distinct signals indicating duplicated genic blocks. Their relatively intact signals suggest their origin from the recent WGD. Moreover, this analysis demonstrated the limitations of *in silico* analysis in assembling WGS reads (Nielsen *et al.* 2010; Alkan *et al.* 2011), while demonstrating the power of synergism between *in silico* and FISH analyses. This success of genic probes in validating two contiguous ginseng scaffolds can be further utilized in other more cryptic assembly.

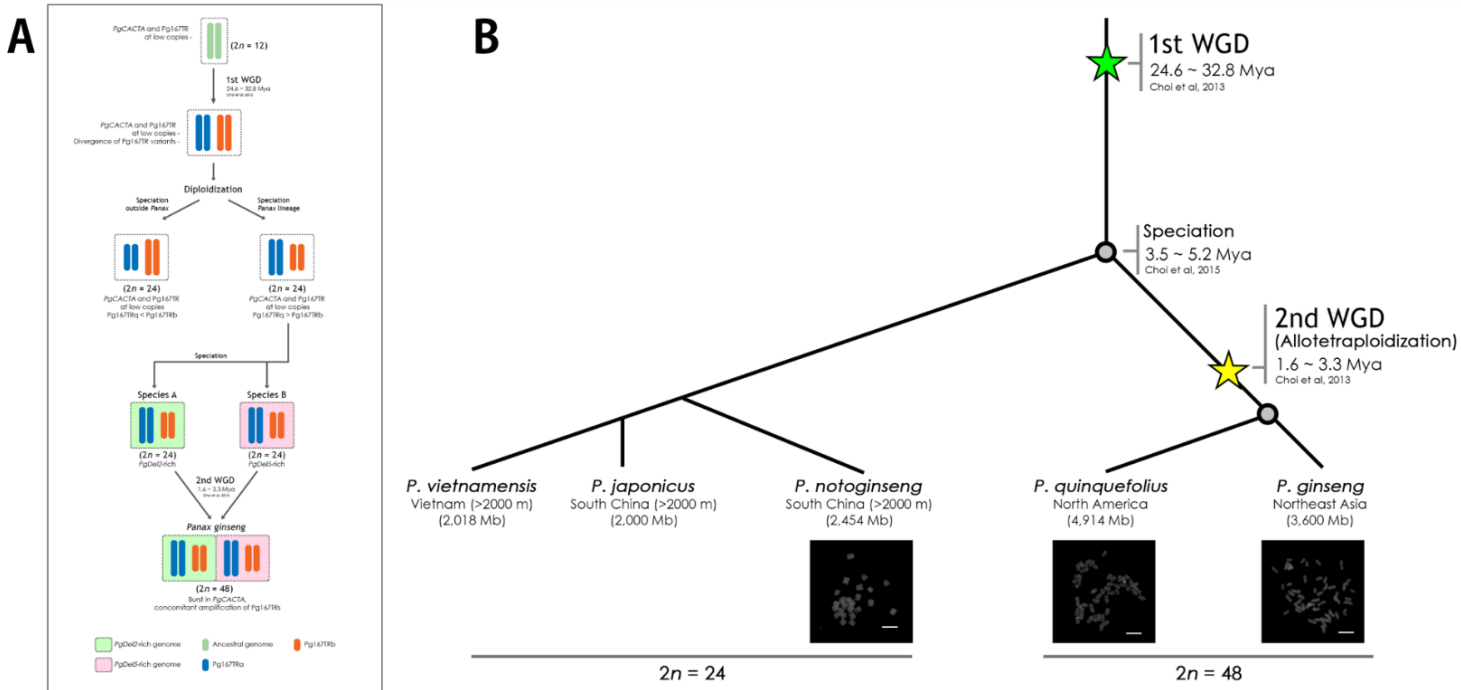

A
PgCACTA and Pg167TR at low copies -
(2n = 12)
1st WGD
24.6 ~ 32.8 Mya
PgCACTA and Pg167TR at low copies - Divergence of Pg167TR variants -
Diploidization
Speciation outside Panax
Speciation Panax lineage
(2n = 24)
PgCACTA and Pg167TR at low copies
Pg167TRq < Pg167TRb
(2n = 24)
PgCACTA and Pg167TR at low copies
Pg167TRq > Pg167TRb
Speciation
Species A
Species B
(2n = 24)
PgDel2-rich
(2n = 24)
PgDel3-rich
2nd WGD
1.6 ~ 3.3 Mya
Panax ginseng
(2n = 48)
Burst in PgCACTA, concomitant amplification of Pg167TRs
PgDel2-rich genome
Ancestral genome
Pg167TRb
PgDel3-rich genome
Pg167TRa
B
1st WGD
24.6 ~ 32.8 Mya
Choi et al, 2013
Speciation
3.5 ~ 5.2 Mya
Choi et al, 2015
2nd WGD
(Allotetraploidization)
1.6 ~ 3.3 Mya
Choi et al, 2013
P. vietnamensis
Vietnam (>2000 m)
(2,018 Mb)
P. japonicus
South China (>2000 m)
(2,000 Mb)
P. notoginseng
South China (>2000 m)
(2,454 Mb)
P. quinquefolius
North America
(4,914 Mb)
P. ginseng
Northeast Asia
(3,600 Mb)
2n = 24
2n = 48

Fig. 4-9. Cytogenetic-aided evolutionary model of *P. ginseng* genome. A) Schematic model for genome evolution of the *P. ginseng* genome taking into account cytogenetic information and previously reported divergence time (Choi *et al.* 2013 and Kim *et al.* 2014). B) Evolutionary model between five related *Panax* species. Divergence times adapted from Choi *et al*. 2014. Mya = Million years ago.

REFERENCES

Alkan C, Sajjadian S, and Eichler EE. 2011. Limitations of next-generation genome sequence assembly. Nat Methods 8: 61-5.

Bilinski P, Distor K, Gutierrez-Lopez J, Mendoza GM, Shi J, Dawe RK, *et al.* 2014. Diversity and evolution of centromere repeats in the maize genome. Chromosoma.

Chamala S, Chanderbali AS, Der JP, Lan T, Walts B, Albert VA, *et al.* 2013. Assembly and validation of the genome of the nonmodel basal angiosperm amborella. Science 342: 1516-1517.

Choi HI, Kim NH, Kim JH, Choi BS, Ahn IO, Lee JS, *et al.* 2011. Development of reproducible est-derived ssr markers and assessment of genetic diversity in panax ginseng cultivars and related species. Journal of Ginseng Research 35: 399-412.

Choi HI, Kim NH, Lee J, Choi BS, Do Kim K, and Park JY. 2013. Evolutionary relationship of panax ginseng and p. Quinquefolius inferred from sequencing and comparative analysis of expressed sequence tags. Genet Resour Crop Evol 60.

Choi HI, Waminal NE, Park HM, Kim NH, Choi BS, Park M, *et al.* 2014. Major repeat components covering one-third of the ginseng (panax ginseng c.A. Meyer) genome and evidence for allotetraploidy. Plant J 77: 906-916.

Fedoroff NV. 2012a. Presidential address. Transposable elements, epigenetics, and genome evolution. Science 338: 758-67.

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

Gao X, Hou Y, Ebina H, Levin HL, and Voytas DF. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res 18: 359-69.

Higashiyama T, Noutoshi Y, Fujie M, and Yamada T. 1997. Zepp, a line‐like retrotransposon accumulated in the chlorella telomeric region. The EMBO Journal 16: 3715-3723.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, *et al.* 2009. The genome of the cucumber, *cucumis sativus* l. Nature Genetics 41: 1275-81.

Kalendar R, Tanskanen J, Immonen S, Nevo E, and Schulman AH. 2000. Genome evolution of wild barley (hordeum spontaneum) by bare-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci U S A 97: 6603-7.

Kelly LJ, Renny-Byfield S, Pellicer J, Macas J, Novak P, Neumann P, *et al.* 2015. Analysis of the giant genomes of fritillaria (liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytol 208: 596-607.

Khrustaleva LI, and Kik C. 2001. Localization of single-copy t-DNA insertion in transgenic shallots (allium cepa) by using ultra-sensitive fish with tyramide signal amplification. Plant J 25: 699-707.

Kim N-H, Lee S-C, Choi H-I, Kim K, Choi B-S, Jang W, *et al.* Genome sequences and evolution of *panax ginseng*. *In:* KIM, S.-K. & RHEE, D.-K., eds. Evidence-based Efficacy of Ginseng 2014, Proceedings of the 11th International symposium on Ginseng, 2014a Seoul, Korea. The Korean Society of Ginseng, 125-146.

Kim NH, Choi HI, Ahn IO, and Yang TJ. 2012. Est-ssr marker sets for practical authentication of all nine registered korean ginseng cultivars. Journal of Ginseng Research 36.

Kim NH, Choi HI, Kim KH, Jang W, and Yang TJ. 2014b. Evidence of genome duplication revealed by sequence analysis of multi-loci expressed sequence tag-simple sequence repeat bands in panax ginseng meyer. J Ginseng Res 38: 130-5.

Kovarik A, Renny-Byfield S, Grandbastien M-A, and Leitch A 2012. Evolutionary implications of genome and karyotype restructuring in nicotiana tabacum l. *In:* SOLTIS, P. S. & SOLTIS, D. E. (eds.) *Polyploidy and genome evolution.* Springer Berlin Heidelberg.

Kubis SE, Heslop-Harrison JS, Desel C, and Schmidt T. 1998. The genomic organization of non-ltr retrotransposons (lines) from three beta species and five other angiosperms. Plant Mol Biol 36: 821-31.

Lim KB, Yang TJ, Hwang YJ, Kim JS, Park JY, Kwon SJ, *et al.* 2007. Characterization of the centromere and peri-centromere retrotransposons in

brassica rapa and their distribution in related brassica species. Plant J 49: 173-83.

Lisch D. 2013. How important are transposons for plant evolution? Nat Rev Genet 14: 49-61.

Lou Q, Zhang Y, He Y, Li J, Jia L, Cheng C, *et al.* 2014. Single-copy gene-based chromosome painting in cucumber and its application for chromosome rearrangement analysis in cucumis. Plant J 78: 169-79.

Ma J, Wing RA, Bennetzen JL, and Jackson SA. 2007. Plant centromere organization: A dynamic structure with conserved functions. Trends Genet 23: 134-9.

Macas J, Neumann P, and Navratilova A. 2007. Repetitive DNA in the pea (pisum sativum l.) genome: Comprehensive characterization using 454 sequencing and comparison to soybean and medicago truncatula. BMC Genomics 8: 427.

Macas J, Novak P, Pellicer J, Cizkova J, Koblizkova A, Neumann P, *et al.* 2015. In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe fabeae. Plos One 10.

Michael TP, and Jackson S. 2013. The first 50 plant genomes. Plant Genome 6.

Michael TP, and Vanburen R. 2015. Progress, challenges and the future of crop genomes. Current opinion in plant biology 24: 71-81.

Neumann P, Navratilova A, Koblizkova A, Kejnovsky E, Hribova E, Hobza R, *et al.* 2011. Plant centromeric retrotransposons: A structural and cytogenetic perspective. Mob DNA 2: 4.

Nielsen CB, Cantor M, Dubchak I, Gordon D, and Wang T. 2010. Visualizing genomes: Techniques and challenges. Nat Methods 7: S5-S15.

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, *et al.* 2010. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytologist 186: 37-45.

Qi LL, Wu JJ, Friebe B, Qian C, Gu YQ, Fu DL, *et al.* 2013. Sequence organization and evolutionary dynamics of brachypodium-specific centromere retrotransposons. Chromosome Res 21: 507-21.

Renny-Byfield S, Kovarik A, Chester M, Nichols RA, Macas J, Novak P, *et al.* 2012. Independent, rapid and targeted loss of highly repetitive DNA in natural and synthetic allopolyploids of nicotiana tabacum. Plos One 7.

Renny-Byfield S, Kovarik A, Kelly LJ, Macas J, Novak P, Chase MW, *et al.* 2013. Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. Plant Journal 74: 829-839.

Sigman MJ, and Slotkin RK. 2016. The first rule of plant transposable element silencing: Location, location, location. The Plant Cell 28: 304-313.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, and Holmes IH. 2009. Jbrowse: A next-generation genome browser. Genome Research 19: 1630-1638.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, *et al.* 2015a. Nested radiations and the pulse of angiosperm diversification: Increased diversification rates often follow whole genome duplications. New Phytol 207: 454-67.

Tank DC, Eastman JM, Pennell MW, Soltis PS, Soltis DE, Hinchliff CE, *et al.* 2015b. Nested radiations and the pulse of angiosperm diversification: Increased diversification rates often follow whole genome duplications. New Phytologist 207: 454-467.

Tenaillon MI, Hollister JD, and Gaut BS. 2010. A triptych of the evolution of plant transposable elements. Trends in Plant Science 15: 471-478.

Waminal N, Park HM, Ryu KB, Kim JH, Yang TJ, and Kim HH. 2012. Karyotype analysis of panax ginseng c.A.Meyer, 1843 (araliaceae) based on rdna loci and dapi band distribution. Comparative Cytogenetics 6: 425-441.

Waminal NE, Perumal S, Lee J, Kim HH, and Yang T-J. 2016. Repeat evolution in brassica rapa (aa), b. Oleracea (cc), and b. Napus (aacc) genomes. Plant Breeding and Biotechnology 4: 107-122.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, *et al.* 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973-982.

Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo D-H, Shi J, *et al.* 2009. Maize centromere structure and evolution: Sequence analysis of

centromeres 2 and 5 reveals dynamic loci shaped primarily by retrotransposons. PLoS Genet 5: e1000743.

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

CONCLUSION

Genomic studies on *Panax ginseng* has gained pace in recent years, providing data for better understanding of its structure and evolution (Choi *et al.* 2011; Kim *et al.* 2012; Choi *et al.* 2013; Choi *et al.* 2014; Kim *et al.* 2014a; Kim *et al.* 2014b). Most of this information is derived from molecular and *in silico* data, and this study provides the first cytogenomic data in understanding the *P. ginseng* genome. Through this study, individual chromosomes of *P. ginseng* was characterized. A high-copy Pg167TR satDNA was vital in achieving this objective. It was also revealed that amplification of *PgCACTA1* in the genus *Panax* was influential in Pg167TR expansion, which was more profound among allotetraploid species and even more intense in *P. ginseng*. This study also opens a research opportunity towards functional studies of *PgCACTA1* in *P. ginseng* genome, particularly in analyzing whether or not it has close association with significant genes that define the medicinal properties of *P. ginseng*. These endeavors should eventually allow us to exploit these data for crop improvement, for engineering cell systems for targeted production of specific types of ginsenosides, and other desired outcome. This study also allowed the visualization of the chromosomal location of different types of major REs in the *P. ginseng* genome, as well as demonstrated the use of pooled genes from assembly annotation to validate integrity of the current version of *P. ginseng* genome assembly. Cytogenomics approach will be useful in understanding genome structure and evolution of other species that are, or will be, undergoing genome sequencing, especially those with putative allopolyploid origin.

ABSTRACT IN KOREAN

인삼 (*Panax ginseng* C.A Meyer)은 다양한 약리작용을 가지고 있어 전통적으로 동아시아 지역을 중심으로 오랫동안 중요 약용식물의 하나로 이용되어 왔다. 오래전부터 인삼의 약리작용에 관해서 활발히 연구가 진행되어 왔으며 많은 연구 결과가 보고되어 있고, 최근들어 유전학 및 유전체학 연구를 통해 인삼 유전체의 구조와 특징 및 일련의 진화적인 사건들에 대해 밝혀지면서 유전체 연구에 있어 많은 진전을 이뤄냈지만, 이를 뒷받침하는 분자세포유전학 연구는 아직 부족한 실정이다. 따라서 본 연구에서는 인삼 유전체의 구조 및 진화 과정을 보다 확실한 이해를 위해, 분자세포유전학 기술과 유전체 내의 반복서열 및 유전자밀집지역 분석을 접목한 종합적인 세포유전학적 접근 방법을 이용하였다. 이는 염색체 수준에서의 인삼 유전체 분석, 개별 염색체의 판별, 새로운 satellite DNA 인 Pg167TR 의 기원 판별, 주요 인삼 반복서열들의 세포유전학적 mapping, 그리고 scaffold assembly 의 검증에 관한 내용들을 포함하고 있다.

첫 번째 챕터에서는 형광동소보합법 (fluorescence *in situ* hybridization)을 활용하여 인삼 염색체들에서 관찰되는 DAPI 밴드 및 3 개의 분자탐침(5S 리보솜 RNA 와 45s 리보솜 RNA, Pg167TR)을 통해 염색체의 구성과 핵형을 분석하였다. 인삼의 정확한 염색체 개수는 2n=48 로 분석되었으며 두릅나무과의 기본 염색체 수가 x=12 임을 고려할 때 4 배체인 것으로 추정하였다. 본 연구에서는 염색체 크기, 염색체 팔의 비율과 함께 4 개의 세포유전학기반 마커들의 조합을 통해 각각의 개별 염색체들을 효과적으로 구별할 수 있었다. 특히 Pg167TR 은 24 쌍의 상동염색체를 구별할 수 있을 뿐 아니라, 최초로 FISH 기반 인삼 핵형 분석을 가능하게 한 효과적인 세포유전학 마커였다.

두 번째 챕터에서는 인삼 유전체 내에서의 Pg167TR 의 분포와 특징을 분석하였고, 몇몇의 (다른 염색체 지역에 있는) 긴 반복배열좌위(long tandem array loci)를 유발한 잠재적인 증폭 경로에 대해 연구하였다. Pg167TR 은 인삼속 내의 다른 종들과 비교하였을 때 인삼 유전체에 특이적으로 많이 증폭되어 있는 것으로 확인되었다. 또한 Pg167TR 내에는 두 종류의 서열 변이가 있는 것을 확인하였고 이를 각각 Pg167TRa 와 Pg167TRb 로 명명하였다. 이 중 Pg167TRa 는 Pg167TRb 보다 유전체 내에 더 많이 존재하며 다양성이 높고, Pg167TR 의 인삼 유전체 내에 증폭에 관여한 주요 인자라 할 수 있었다. 인삼속 내의 다른 종들에서는 Pg167TRa 가 Pg167TRb 보다 유전체 내의 비율이 더 높게 나왔으나, 인삼속 외의 다른 두릅나무과 식물들에서는 반대의 양상이 관찰되었다. Pg167TR 내에서 생물학적 및 비생물학적 반응과 관련이 있는 잠재적인 *cis*-regulatory elementary 들을 발견하였으며 이는 Pg167TR 이 인삼의 생리적 반응과 연관이 있을 것임을 시사한다. 추가적인 *in silico* 분석에서 몇몇 CACTA 전이인자들의 3’ 지역에서 다양한 Pg167TR 단위 반복수가 관찰되었으며, 어떤 것들은 1000 개 이상이 반복되기도 하였다. CACTA 전이인자의 전이효소 도메인과 Pg167TR 지역 각각을 개별적인 탐침으로 이용한 FISH 분석은, Pg167TR 배열들의 증폭이 CACTA 전이인자로부터 유래되었음을 뒷받침하며, 이는 satDNA 가 Class II 전이인자에 속하는 CACTA DNA 전이인자로부터 또 다른 경로를 통해 진화하였음을 의미한다.

세 번째 챕터에서는 세포유전학적 맵핑을 통해 기존에 밝혀져 있던 주요 인삼 반복서열들의 분포 양상에 대해 종합적으로 분석하였다. 더불어 인삼 스캐폴드 내의 유전자 구역을 PCR 로 증폭 및 이를 탐침으로 이용하여, 인삼 유전체의 주요 특징 중 하나인 최근 전장유전체배가 현상(Recent

whole-genome duplication)에 대해 조사하였다. 주요 인삼 반복서열의 세포유전학적인 맵핑을 통해서는 각 반복서열 그룹들이 각 염색체의 부위별로 다양하게 분포하고 있음을 알 수 있었다. 몇몇의 반복인자들은 동원체 인접지역(pericentromeric region)에 주로 분포되어 있는가 하면, 어떤 반복인자들은 아종말체(subtelomere)지역에 주로 분포되어 있는 것을 확인하였다. 더불어 최근 전장염색체배가 현상에서 유래한 paralog 를 이용하여 인접한 스캐폴드를 예상할 수 있는데, 이런 2 개의 파랄로그 관계인 스캐폴드 그룹을 대상으로 FISH 분석을 수행하여 이들 2 개의 paralogous block 에 대한 검증도 동시에 진행하였다.

본 연구는 최초의 FISH 기반 인삼 핵형 분석기법을 완성하였으며, satDNA 가 CACTA 전이인자로부터 진화한 대체 경로를 제시하였고, 각 전이인자 그룹 별로 다른 양상의 염색체 분포를 밝혔다. 이들을 종합하여 인삼 유전체의 구조와 진화 역사를 이해할 수 있을 것으로 생각된다. 또한 이는 향후 인삼 및 관련 종들의 세포유전학 분석을 위한 기반이 될 것이며, 인삼 유전체 어셈블리의 검증과 함께 유전지도와 세포유전학 지도를 통합하는데 도움을 주어, 인삼의 작물로서의 가치 향상에 기여할 것이다.