RESEARCH ARTICLE

A Glimpse of Panax ginseng Genome Structure Revealed from Ten BAC Clone Sequences Obtained by SMRT Sequencing Platform

Woojong Jang¹, Nam-Hoon Kim¹, Junki Lee¹, Nomar Espinosa Waminal¹, Sang-Choon Lee¹, Murukarthick Jayakodi¹, Hong-Il Choi², Jee Young Park¹, Jong-Eun Lee³, Tae-Jin Yang¹,⁴*

¹Department of Plant Science, Plant Genomics and Breeding Institute, and Research Institute of Agriculture and Life Sciences, College of Agriculture and Life Sciences, Seoul National University, Seoul 08826, Korea
²Advanced Radiation Technology Institute, Korea Atomic Energy Research Institute, Jeongeup 56212, Korea
³DNA Link, Inc. Seoul 03759, Korea
⁴Crop Biotechnology Institute/GreenBio Science and Technology, Seoul National University, Pyeongchang 25354, Korea

ABSTRACT  Korean ginseng (Panax ginseng) is a well-known valuable medicinal plant with excellent therapeutic effects, however its complex genome structure has not been elucidated yet. To understand its genome structure, we obtained ten ginseng bacterial artificial chromosome (BAC) clone sequences by single-molecule real-time (SMRT) sequencing platform using a pooled DNA of the BAC clones. Out of the ten BAC clones, nine were completely assembled without any gap and one remained a single gap. The total length of BAC clone sequences was 1,163,364 bp. Sophisticated sequence analysis revealed that the 89.7% of the sequences are high copy repeat regions and the remaining 10.3% are non-repeat regions. Eleven protein-coding genes were identified in the non-repeat regions. Most of the repeat regions show more than 1,000 copies and complex structure of various repetitive elements. Ty3/Gypsy family long terminal repeat retrotransposons (LTR-RTs) are predominant repeats occupying 46.9% of the 1,163-kbp sequence. We identified six novel LTR-RTs and their insertion time. Fluorescence in situ hybridization (FISH) analysis demonstrated that PgDel2 and PgDel5 elements had a subgenome-biased distribution. Collectively, our analysis reveals that ginseng genome has very complex genome structure with abundant repetitive elements and rare gene frequency.

Keywords  Panax ginseng, Bacterial artificial chromosome, SMRT sequencing, Long terminal repeat retrotransposon

INTRODUCTION

Korean ginseng (Asian ginseng, ginseng; Panax ginseng C.A. Meyer), which belongs to the Araliaceae family, is a slow-growing perennial herbal plant distributed mainly in Northeastern Asia (Yun 2001). It has been cultivated as an important medicinal crop for hundreds of years (Park et al. 2012a). Various types of ginsenosides biosynthesized in ginseng show numerous therapeutic effects on human, such as anti-wrinkle (So et al. 2008), anti-stress (Kumar et al. 2016), boosts the immune system (Quan et al. 2007), controls symptoms of diabetes (Xie et al. 2005), Alzheimer’s disease (Lee et al. 2007) and cancer (Wong et al. 2015). Despite its excellent pharmacological effects and many studies on efficacy of their medicinal components, breeding or genetic studies are still limited in this plant species. With increasing interests in the genome in recent years, various efforts have been made for marker development (Choi et al. 2011; Kim et al. 2012; Kim et al. 2013; Jung et al. 2014), DNA library construction (Bang et al. 2010; Hong et al. 2004) and complete assembly of 45S nuclear ribosomal DNA and chloroplast genome (Kim et al. 2009).
P. ginseng has an estimated genome size of over 3.5 Gbp for haploid genome equivalent (Hong et al. 2004; Waminal et al. 2012) and is regarded as an allotetraploid \(2n=4x=48\) (Choi et al. 2014). A previous study illustrated that P. ginseng genome had experienced two rounds of whole genome duplication (WGD) (Choi et al. 2013), and recent WGD occurring 2-3 million years ago (MYA), which made its genome much bigger than those of other diploid Panax species (Yi et al. 2004; Choi et al. 2013). Subsequently, studies were also carried out using simple sequence repeats derived from duplicated genes (Kim et al. 2014) and three repeat-rich bacterial artificial chromosome (BAC) clone sequences (Choi et al. 2014) to understand the genome structure of ginseng. However, characterization of more diverse long genome sequences is necessary for understanding of P. ginseng genome structure.

Recently, the genomes of various organisms have been analyzed rapidly and accurately with the advent of various high-throughput sequencing technologies (Pareek et al. 2011). Among those, the Pacific Biosciences (PacBio, Menio Park, CA, USA) platform has gained attention in various genome projects (Gordon et al. 2016; Hoshino et al. 2016). This single molecule real-time (SMRT) sequencing platform has a unique feature that produce ultra-long read sequences, which allows to uncover the structure of complex repetitive elements dispersed in centromeres or telomeres of a genome (Bennett et al. 2016; Wolfgruber et al. 2016). In addition, this platform has also been widely used in genome assemblies of various organisms ranging from bacteria (Liao et al. 2015; Tanizawa et al. 2015) to complex eukaryotes with large and highly repetitive genomes (VanBuren et al. 2015; Ming et al. 2015; Gordon et al. 2016; Hoshino et al. 2016).

BAC clones harbor around 100-kb long DNA fragments derived from genome of a specific organism (O’Connor et al. 1989). Investigation and characterization of BAC sequences provide an estimated genome structure for the target organism. In this study, ten ginseng BAC clones were randomly selected and sequenced using SMRT sequencing platform. From the comprehensive analysis of assembled BAC sequences, genomic structures and novel long terminal repeat retrotransposons (LTR-RTs) were characterized. This study will provide useful resources for further understanding of genome structure and evolution in P. ginseng.

**MATERIALS AND METHODS**

**BAC clone selection and DNA extraction**

Ten BAC clones were randomly selected from BAC library of P. ginseng cv. Chunpoong (Hong et al. 2004). The selected BAC clones were cultured in 500 mL 2xYT medium for 20 hours at 38°C and harvested by centrifugation. The BAC DNAs were isolated using the QIAGEN Plasmid Midi Kit (Qiagen, Hilden, Germany) according to a protocol suitable for very low-copy plasmid extraction. The extracted DNAs were quantified by NanoDrop ND-1000 (Thermo Scientific, USA).

**SMRT sequencing and assembly**

The same amount of each DNA sample of the ten BACs was pooled into a single tube and provided to DNA Link (Seoul, Korea) for the construction of a sequencing library. A 10-kb library was generated and sequenced using PacBio RS platform installed MagBead protocol. Two sequencing runs were performed using P4-C2 and P6-C4 chemistry versions. For production of long read, one SMRT cell was employed for each run. The generated reads were corrected and pre-assembled through the hierarchical genome assembly process (HGAP) method (Chin et al. 2013), and de novo assembly of BAC sequences was carried out using Celera assembler v8.2 (http://wgs-assembler.sourceforge.net/). Assembled sequences were polished by quiver program (https://github.com/PacificBiosciences/GenomicConsensus). Finally, the high-quality ten BAC sequences were obtained by removal of E.coli and vector sequences.

**BAC sequences annotation**

For preliminary prediction of genic regions, FGENESH program (Salamov and Solovyev 2000; http://www.softberry.com/) was used with tomato as reference model on default parameters. The gene structures were confirmed by BLASTP searches against the NCBI non-redundant (nr) protein sequence database and our customized trans-
criptome data. Repeat regions were predicted on the basis of NCBI Conserved Domain Search (CD-Search) (Marchler-Bauer et al. 2009; https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and CENSOR program (Kohany et al. 2006; http://www.girinst.org/censor/). Precise positions of repetitive elements in BAC sequences were finally determined by manual verification based on dot-plot analysis in which BAC sequences were compared with reported repeat elements of P. ginseng (Choi et al. 2014) using PIPMaker (Schwartz et al. 2000; http://pipmaker.bx.psu.edu/cgi-bin/pipmaker?advanced) and Tandem Repeat Finder (Benson 1999; http://tandem.bu.edu/trl/trl.html) programs with default parameters.

Repeat element analysis

To classify the repetitive elements, reverse transcriptase (RT) domain of each LTR-RT was extracted from BAC sequences and compared with core database of GyDB (Llorens et al. 2011; http://gydb.org/index.php/Main_Page) using BLASTX searches. Internal structure of LTR-RTs was confirmed by CD-Search analysis. To estimate insertion time of LTR-RTs, each LTR sequence from both sides were extracted and aligned using ClustalW (Thompson et al. 1994) with default parameters. The numbers of transition (Ts) and transversion (Tv) mutations were calculated using MEGA version 7.0 (Kumar et al. 2016). The pairwise distance between the LTR sequences from both side was calculated using the Kimura two-parameter model (Kimura 1980). A substitution rate value was adopted as 1.22×10^{-8} which was used in the previous expressed sequence tag analysis in ginseng (Choi et al. 2013).

WGS read mapping and genome structure analysis

Whole genome sequence (WGS) reads from P. ginseng cv. Chunpoong were generated using HiSeq 2000 (Illumina, Inc., San Diego, CA, USA) as a part of P. ginseng genome sequencing project (data not shown). The 30 Gbp reads corresponding to approximately 10 times of ginseng genome size were randomly sampled and mapped onto assembled BAC sequences by CLC Assembly Cell version 4.21 software (https://www.qiagenbioinformatics.com/products/clc-assembly-cell/) with default parameters. The genome proportion of LTR-RTs was calculated by dividing the total mapping depth of each LTR-RT region by 30 Gbp. To estimate repetitive regions of the BAC sequence, RepeatMasker software (http://www.repeatmasker.org) was used with default parameters using a custom repeat database comprising both repeat sequences identified in our previous study (Choi et al. 2014) and newly in this study.

Fluorescence in situ hybridization (FISH) analysis

PgDel5 FISH probe was designed using its RT domain sequences (forward: 5’-TACCAATCTTCAGTGACTTACACGA-3’, reverse: 5’-TCGAGTTATAAATTGTGC-5AATGA-3’) and PgDel2 FISH probe which showed subgenome-specific pattern was used from a previous study (Choi et al. 2014) for FISH analysis. PCR amplicons were labeled through direct nick-translation with Alexa Fluor® 488-5-dUTP (Invitrogen, C11397) and Texas Red-5-dUTP (Perkin Elmer, NEL417001EA), respectively. FISH procedures were as described previously (Waminal et al. 2012). Briefly, FISH immediately followed fixation of slides with somatic metaphase chromosomes in 4% paraformaldehyde without pepsin and RNase pretreatment. The hybridization mixture contained 50% formamide, 10% dextran sulfate, 2×SSC, 5 ng/μL salmon sperm DNA and 20 ng/μL of each probe DNA and was adjusted with DNase- and RNase-free water (Sigma, USA, #W4502) to a total volume of 40 μL/slide. The hybridization mixture was denatured at 90°C for 10 minutes and immediately kept on ice for at least 5 minutes prior to mounting on slides. After covering with a glass coverslip, the chromosomes were denatured at 80°C for 3-5 minutes on a hot plate. The slides were then immediately transferred into a humid chamber preset at 37°C and incubated overnight (~16 hours). The following day, the slides were washed in 2× SSC (15 minutes at RT), 0.1× SSC (35 minutes at 42°C), and finally 2× SSC (30 minutes at room temperature). 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 were performed using Adobe Photoshop CC.
RESULTS

Sequencing and assembly of BAC clones

Ten BAC clones were randomly selected and pooled for sequencing using PacBio SMRT sequencing platform. Two independent reactions were conducted using P4-C2 and P6-C4 chemistry. Among two sequencing reactions, P4-C2 chemistry produced 44,736 reads with an average length of 3,605 bp, while P6-C4 chemistry produced 100,844 reads with an average length of 4,813 bp (Table 1). Approximately, 9.7% of reads were longer than 10 kb. Short and low quality reads (<50 bp and <0.75 respectively) were removed for optimal assembly. Initial assemblies using each read set from P4-C2 and P6-C4 chemistries produced good quality of contig sequences with only seven gaps and one gap, respectively. By combining of both data, we obtained complete sequences without gap for nine BACs (BAC IDs 8L14, 9P08, 6B09, 8G13, 5N01, 8P22, 8C22, 5B21 and 10M13) and with one gap for BAC ID 7P20. Overall, a total 1,163,364 bp of assembled sequences were generated (Table 2).

Table 1. Summary of sequence statistics.

| Read size | P4-C2 | P6-C4 |
|-----------|-------|-------|
|           | Total read length (bp) | Number of reads | Average read length (bp) | Total read length (bp) | Number of reads | Average read length (bp) |
| ~5 kb     | 7,107,7597 | 33,924 | 2,095 | 160,132,596 | 66,268 | 2,416 |
| 5 kb~10 kb| 58,026,397 | 8,263 | 7,022 | 164,293,017 | 22,992 | 7,145 |
| ~10 kb    | 32,183,323 | 2,549 | 12,625 | 161,018,656 | 11,584 | 13,900 |
| Total     | 161,287,317 | 44,736 | 3,605 | 485,444,269 | 100,844 | 4,814 |

Table 2. Assembly statistics of ten BAC clones.

| BAC ID | Assembled length (bp) | Number of contigs | Non-repetitive region (bp, <50 mapping depth) | GenBank Accession number |
|--------|------------------------|-------------------|-----------------------------------------------|-------------------------|
| 8L14   | 173,429                | 5                 | 11,843                                       | KY513616               |
| 9P08   | 139,247                | 1                 | 7,290                                        | KY513618               |
| 6B09   | 131,717                | 2                 | 25,544                                       | KY513612               |
| 8G13   | 123,350                | 1                 | 902                                          | KY513615               |
| 5N01   | 117,096                | 2                 | 90                                           | KY513611               |
| 8P22   | 109,113                | 1                 | 232                                          | KY513617               |
| 8C22   | 107,984                | 1                 | 61,448                                       | KY513614               |
| 5B21   | 100,349                | 1                 | 196                                          | KY513610               |
| 10M13   | 97,432                | 1                 | 2,611                                         | KY513619               |
| 7P20   | 63,647                 | 2                 | 9,707                                        | KY513613               |

Two BAC clones, 8P22 and 10M13, were not included in the second sequencing run using P6-P4 chemistry.

Sequence annotation

Annotations of ten BAC clone sequences allow us to predict complex genome structure of ginseng, which is mainly composed of various repetitive components with relatively small portion of genic regions (Fig. 1). The six BAC sequences (BAC IDs 8L14, 8G13, 5N01, 8P22, 8C22, and 5B21) harbor only repeat elements without any gene. Among repeat elements, Ty3/Gypsy type LTR-RTs were predominantly distributed throughout the entire sequences. Eleven genes were identified in four BAC sequences (BAC IDs 9P08, 6B09, 10M13 and 7P20) which have less repetitive elements (Fig. 1, Table 3). In addition, six novel LTR-RTs were identified in four BAC sequences (BAC IDs 8G13, 8P22, 8L14, 7P20 and 10M13) and one tandem repeat (TR) was identified in the middle of BAC ID 5N01 sequence (Table 4).
Fig. 1. Sequence analysis of ten BAC clones in ginseng. Horizontal black bars represent ten BAC sequences. Identified genome components were shown above the black bar and depicted according to their position on the BAC sequences. The description of each component is shown at the bottom right. Graphs under the black bars represent the depth distribution of 30 Gbp WGS read mapping (approximately 10x coverage). RD indicate the read depth plot for each nucleotide position using 10x WGS.
Table 3. Annotation summary of identified eleven genes in BAC sequences.

| Gene annotation based on BLASTP searches | BAC ID | Position (bp) | # of exon | Accession no. (E-value) |
|------------------------------------------|--------|--------------|-----------|------------------------|
| Cellulose synthase A catalytic subunit 3 [UDP-forming] | 9P08   | 215-9,773    | 14        | XP_01726278.1 (0.0)    |
| Calcium-dependent protein kinase 28      | 6B09   | 1-5,383      | 9         | XP_017251289.1 (0.0)   |
| Uncharacterized protein                  | 6B09   | 18,057-19,415| 3         | XP_017251291.1 (8e-115)|
| Transformation/transcription domain-associated protein | 6B09   | 21,538-59,489| 35        | XP_017271620.1 (0.0)   |
| Transformation/transcription domain-associated protein | 6B09   | 94,283-120,564| 35       | XP_017271620.1 (0.0)   |
| Acyl carrier protein 1                   | 10M13  | 1,127-3,315  | 4         | XP_017257036.1 (1e-55)|
| Superoxide dismutase                     | 10M13  | 33,367-36,315| 7         | O22668.1 (1e-100)     |
| Bifunctional 3-dehydroquinate dehydratase/shikimate dehydrogenase | 10M13  | 44,804-50,967| 10        | XP_017220070.1 (0.0)   |
| Uncharacterized protein                  | 10M13  | 55,540-58,599| 1         | XP_017218710.1 (0.0)   |
| Uncharacterized protein                  | 7P20   | 2,245-7,217  | 5         | XP_017238997.1 (0.0)   |
| Protein FAR1-related sequence 5-like     | 7P20   | 53,287-55,512| 2         | XP_015866013.1 (0.0)   |

Table 4. List of novel LTR retrotransposons identified in BAC sequences.

| Type          | BAC ID | Position (bp)       | TSD$^a$ | Length / LTR length (bp) | Ts$^b$, Tv$^c$ | K$^w$ | Insertion time$^x$ (MYA) |
|---------------|--------|---------------------|---------|--------------------------|---------------|-------|------------------------|
| PgdDel4       | 8G13   | 75,443-76,688, 77,213-87,016 | GCAAC   | 11,050       | 21, 15 | 1.4    | 0.014 0.57 |
| PgdDel5       | 8P22   | 1-10,587            | CAAGC   | 10,587       | 5, 8  | 0.63   | 0.010 0.41 |
| PgdDel6       | 8P22   | 20,535-32,786       | GCGCT   | 12,252       | 60, 53 | 1.13   | 0.037 1.51 |
| PgTork2       | 8L14   | 140,265-148,881     | GCAAC   | 8,615        | 13, 6  | 2.17   | 0.012 0.49 |
| PgSire1       | 7P20   | 32,485-35,143, 39,282-46,836 | AAAGG   | 10,214       | 41, 14 | 2.93   | 0.037 1.51 |
| PgSire2       | 10M13  | 65,051-65,495, 65,765-71,239 | CCAGT   | 5,920        | 8, 3   | 2.67   | 0.045 1.81 |
$^a$Target site duplication.
$^b$Number of transition mutations.
$^c$Number of transversion mutations.
$^w$Kimura's distance.
$^x$Insertion times were estimated by adopting the substitution rate of $1.22 \times 10^{-8}$.

Classification and characterization of LTR-RTs

Among the six novel LTR-RTs, three are classified to Del family, which belong to Ty3/Gypsy type LTR-RTs, and named as PgdDel4, PgdDel5, and PgdDel6 subfamilies. The remaining three elements had RT domain similar to Ty1/Copia type LTR-RTs. Two belonged to Sire family, named as PgSire1 and PgSire2, and another one belong to Tork family, named as PgTork2. The overall structure of each element was confirmed by CD-Search (Fig. 2). Unlike the previously reported PgdDel subfamilies (PgdDel1, 2, 3), a zinc-knuckle (accession number cl15298 in NCBI CDD) and a chromodomain (accession number cl15261 in NCBI CDD) were not found in the PgdDel4 and PgdDel6. On the other hand, PgTork2 contain an unusual zinc-knuckle domain, which is a distinct feature from other Ty1/Copia family members. PgSire1 seemed to experience internal domain losses caused by the insertion of other LTR-RT
**Fig. 2.** Structure of newly identified long terminal repeat retrotransposons (LTR-RTs) in ginseng BAC sequences. Vertical bars on both ends represent the target site duplication (TSD). The mark (+ or −) next to the transposable elements ID indicates the direction on the BAC sequence. Identified internal domain, AP (aspartic protease), CH (chromodomain), GAG (capsid protein), INT (integrase), RH (RNase H), RT (reverse transcriptase) and Zn (zinc knuckle), were depicted according to their position in the internal regions. The dotted box region of *PgDel5* are estimated structure because it was truncated in the BAC clone sequence. The omitted region of *PgSire1* indicates area lost by external factors.

members belonged to *Ty3/Gypsy* family.

To estimate insertion time of the six novel LTR-RTs, the ratios of transition to transversion (Ts/Tv) was calculated between both flanking LTR sequences (Table 4). The Ts/Tv ranged from 0.63 to 2.93 with an average value of 1.82. We measured nucleotide substitution rate between both LTRs of each element to estimate insertion time of each of six LTR-RTs. Kimura’s nucleotide substitution rates ranged from 0.010 to 0.045 and the insertion time of each element was estimated to 0.41-1.81 MYA.

**Estimation of ginseng genome structure**

Mapping of 10x WGS reads revealed that approximately 10% of BAC sequences were non-repetitive and the remaining 90% regions were repeat-replete regions (Table 2, Fig. 1). The non-repetitive regions showed less than 50x coverage mapping depth, while repeat-replete regions showed high mapping depth with more than 10,000x coverage. The repetitive regions were mainly occupied by various LTR-RTs with complex and nested insertion patterns. For calculation of actual repeat proportion in the ten BACs, we conducted repeat masking with previously reported and newly identified LTR-RTs. Homology-based search revealed that over 60% of the BAC sequences comprised various LTR-RTs (Table 5). Most of the regions were occupied by *Ty3/Gypsy* type LTR-RTs family members (46.88%), unknown repeats (8.57%) and *Ty1/Copia* family members (5.09%). *PgDel* elements are the most common in the BAC sequences, of which, *PgDel1* was the most abundant (24.17%).

We estimated the proportion of LTR-RTs in whole ginseng genome by mapping 30 Gbp of WGS reads onto each element. Approximately 36% of the ginseng genome was estimated to be occupied by 14 LTR-RTs including eight reported and six novel subfamilies. The estimated genome proportion of LTR-RTs is lower than the
Table 5. Proportion of LTR-RTs in the ginseng genome calculated from repeat masking and WGS read mapping.

| Type         | Proportion in BACs (%) | Expected proportion in genome (%) |
|--------------|------------------------|-----------------------------------|
| Ty3/Gypsy    | 46.88                  | 28.66                             |
| PgDel        | 34.53                  | 24.04                             |
| PgDel1       | 24.17                  | 20.0                              |
| PgDel2       | 0.72                   | 0.93                              |
| PgDel3       | 4.07                   | 1.17                              |
| PgDel4       | 1.08                   | 0.37                              |
| PgDel5       | 1.61                   | 0.51                              |
| PgDel6       | 2.88                   | 1.06                              |
| PgTat        | 11.13                  | 4.1                               |
| PgTat1       | 9.83                   | 3.85                              |
| PgTat2       | 1.3                    | 0.25                              |
| PgAthila     | 1.22                   | 0.52                              |
| Ty1/Copia    | 5.09                   | 5.1                               |
| PgTork       | 2.32                   | 1.28                              |
| PgTork1      | 0.9                    | 0.72                              |
| PgTork2      | 1.12                   | 0.56                              |
| PgSire       | 2.9                    | 1.2                               |
| PgSire1      | 2.25                   | 0.92                              |
| PgSire2      | 0.65                   | 0.28                              |
| PgOryco      | 0.17                   | 0.04                              |
| Degenerated LTR-RTs | 8.57            | 2.58                              |
| Total        | 60.54                  | 36.34                             |

proportion found in the BAC sequences. However, both results show similar distribution pattern. For example, both result show that PgDel1 members are most abundant although the estimated value for genome proportion (20.0%) is lower than the proportion in the 1,163 Kb sequences (24.2%) (Table 5).

Our previous FISH analysis revealed that PgDel1 was present in all chromosomes and PgDel2 was in 12 of 24 chromosomes (Choi et al. 2014). To understand the chromosomal distribution of repeat elements, two LTR-RTs, PgDel2 and PgDel5 showing uneven distribution were selected for cytogenetic analysis. Simultaneous FISH analysis with PgDel2 and PgDel5 displayed biased signal between twelve opposite chromosome pairs (Fig. 3). PgDel2 (green signal) was present in only twelve chromosome pairs as consistent with previous study (Choi et al. 2014), while PgDel5 (red signal) showed intense signal in other twelve PgDel2-poor chromosome pairs.

DISCUSSION

Sequencing of pooled BAC clones using SMRT sequencing

The sensational development of sequencing technology enabled assembly of nearly complete genomes in various species (Pareek et al. 2011). SMRT sequencing produce long contiguous sequences, which are essential for comprehensive assembly of repeat-rich genomes. It is also well suited for assembly of relatively small genomes such as bacterial and organelar genomes or BAC sequences (Frank et al. 2015). With this scientific blessing, we successfully assembled ten ginseng BAC clones with abundant repetitive elements. In this study, we conducted single-cell SMRT sequencing for a ten BAC clone pool using two different chemistries, P4-C2 and P6-C4 chemistries. Although we report ten BAC sequences finished by combining of both reactions, it should be noted that each single reaction produced almost complete assembly which can be utilized for further genome study.

Overview of a ginseng genome structure

The ten randomly selected BAC clone sequences showed brief overview of a ginseng genome. Among the
ten BACs, two BACs, 6B09 and 10M13, contained four genes with less repetitive elements, two BACs, 7P20 and 9P08, contained two and one gene, respectively, with moderate repetitive elements, whereas the other six BACs were entirely composed of only repetitive elements without any gene. In particular, many LTR-RTs occupied the largest portion of the BAC sequences with complex nested insertion patterns (Table 5, Fig. 1). The results suggest that LTR-RTs might play a major role in the increase of the genome size such as many other plants with large genomes size (SanMiguel et al. 1996; Park et al. 2012b).

Our previous study identified eight LTR-RTs which occupied a third of the ginseng genome (Choi et al. 2014). Here, we characterized six novel LTR-RTs. Overall, a total of 14 LTR-RTs were characterized from 13 ginseng BAC sequences. More than 60% of the whole BAC sequences were characterized as components of 14 LTR-RTs by repeat masking. Although the 14 LTR-RTs are major repeats occupying more than 60% of ginseng genome, other uncharacterized LTR-RT members, transposable elements and repetitive elements will be found in the ginseng genome.

We estimated the copy numbers of all the components in the 1,163 kb sequences by mapping of 10x WGS reads on the homologous regions based on more than 80% sequence similarity. Approximately 10% of BAC sequences were non-repetitive, while, the remaining 90% show repeat regions with more than 10,000x coverage from the 10x WGS. Overall, 90% of the 1,163 kb are composed of various repetitive elements with over 1,000 copies in the ginseng genome. Repeat masking tool revealed that 60% of 1,163 kb sequences are components of 14 LTR-RTs. However, 36.34% of WGS reads were mapped on the 14 LTR-RTs that is much lower than the proportion of LTR-RTs (60%) occupying in the 1,163 kb sequences (Table 5) that is similar phenomena found in our previous research (Choi et al. 2014). More diverse forms of LTR-RT members will be in the actual ginseng genome that make biased genome proportion estimation values. However, both analysis show the relative abundance for each family member. The most abundant LTR-RTs are PgDel and PgTat families. Genome proportion of PgDel and PgTat families are estimated to be 34.5-24.0% and 11.1-4.1%, respectively (Table 5).

**Amplification of LTR-RTs during ginseng genome evolution**

We proposed that recent WGD in ginseng genome was caused by an allotetraploidization event based on finding of subgenome-specific distribution of transposable elements, PgDel2 (Choi et al. 2013, 2014). The subgenome-specific LTR-RTs distribution was also discovered in other plants such as allopolyploid wheat (Sabot et al. 2006; Salina et al. 2011). In this study, we found another clue for allopolyploidization event in the ginseng genome. FISH analysis revealed that rich signal of PgDel5 are located on 12 PgDel2-poor chromosome pairs. We estimated that allotetraploidization event in ginseng is originated from the hybridization of two related ancestral species, as occurred in other allotetraploid plant species. The ancestral species is unknown for P. ginseng up to now. Discovery of these sub-genome unique LTR-RTs may contribute to unveil the evolutionary story by further comparative analysis against related Panax species.

Insertion time of the LTR-RT has been calculated by estimating the sequence divergence between two LTRs that independently accumulate point mutations at each LTR sequences (Dangel et al. 1995; SanMiguel et al. 1996; SanMiguel et al. 1998). The insertion time of six novel LTR-RTs was estimated to be 0.41-1.81 MYA (Table 4). These results also supported a hypothesis that speciation of ginseng from related Panax species might be accelerated by uneven amplification of various transposable elements after WGD.

The structure and characteristic of the 1,163 kb ginseng genome sequences analyzed in this study will expand our understanding about the complex genome structure of ginseng. Furthermore, our data provide valuable resources for understanding of genome structure and evolution, as well as for breeding and related researches in the genus Panax.

**ACKNOWLEDGEMENTS**

This research was supported by “Cooperative Research
Program for Agriculture Science & Technology Development (Project No. PJ01100801)”, Rural Development Administration, Republic of Korea.

REFERENCES

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

Bennett HW, Liu N, Hu Y, King MC. 2016. Insights into telomerase action from high-throughput sequencing of S. pombe telomeres. bioRxiv: 047837.

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

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods. 10: 563-569.

Choi HI, Kim NH, Lee J, Choi BS, Do Kim K, Park JY, et al. 2013. Evolutionary relationship of Panax ginseng and P. quinquefolius inferred from sequencing and comparative analysis of expressed sequence tags. Genet. Resour. Crop Ev. 60: 1377-1387.

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. J. Ginseng. Res. 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.

Dangel AW, Baker BJ, Mendoza AR, Yu CY. 1995. Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K (C4) are a molecular clock of evolution. Immunogenetics 42: 41-52.

Frank J, Dingemanse C, Schmitz AM, Vossen RH, van Ommen GJ, den Dunnen JT, et al. 2015. The complete genome sequence of the murine pathobiont Helicobacter typhlonius. Front. Microbiol. 6: 1549.

Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM, et al. 2016. Long-read sequence assembly of the gorilla genome. Science 352: aae0344.

Hong CP, Lee SJ, Park JY, Plaha P, Park YS, Lee YK, et al. 2004. Construction of a BAC library of Korean ginseng and initial analysis of BAC-end sequences. Mol. Genet. Genomics 271: 709-716.

Hoshino A, Jayakumar V, Nitasaka E, Toyoda A, Noguchi H, Itoh T, et al. 2016. Genome sequence and analysis of the Japanese morning glory Ipomoea nil. Nat. Commun. 7: 13295.

Jung J, Kim KH, Yang K, Bang KH, Yang TJ. 2014. Practical application of DNA markers for high-throughput authentication of Panax ginseng and Panax quinquefolius from commercial ginseng products. J. Ginseng Res. 38: 123-129.

Kim JH, Jung JY, Choi HI, Kim NH, Park JY, Lee Y, et al. 2013. Diversity and evolution of major Panax species revealed by scanning the entire chloroplast intergenic spacer sequences. Genet. Resour. Crop Ev. 60: 413-425.

Kim K, Lee SC, Lee J, Lee HO, Joh HJ, Kim NH, et al. 2015. Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within Panax ginseng species. PLoS One 10: e0117159.

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

Kim NH, Choi HI, Kim KH, Jang W, 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: 130-135.

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

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

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33: 1870-1874.

Lee ST, Chu K, Kim JM, Park HJ, Kim MH. 2007. Cognitive improvement by ginseng in Alzheimer's disease. J. Ginseng Res. 31: 51-53.

Liao YC, Lin SH, Lin HH. 2015. Completing bacterial genome
assemblies: strategy and performance comparisons. Sci. Rep. 5: 8747.

Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, Tamarit D, et al. 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucl. Acids Res. 39: D70-74.

Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, et al. 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucl. Acids Res. 37: D205-210.

Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, et al. 2015. The pineapple genome and the evolution of CAM photosynthesis. Nat. Genet. 47: 1435-1442.

O’Connor M, Peifer M, Bender W. 1989. Construction of large DNA segments in Escherichia coli. Science 244: 1307-1312.

Pareek CS, Smoczynski R, Tretyn A. 2011. Sequencing technologies and genome sequencing. J. Appl. Genet. 52: 413-435.

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

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. Plant J. 69: 1018-1029.

Quan FS, Compans RW, Cho Y-K, Kang S-M. 2007. Ginseng and Salviae herbs play a role as immune activators and modulate immune responses during influenza virus infection. Vaccine 25: 272-282.

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

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, et al. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274: 765-768.

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, et al. 2000. PipMaker—a web server for aligning two genomic DNA sequences. Genome Res. 10: 577-586.

So SH, Lee SK, Hwang EI, Koo BS, Han GH, Chung JH, et al. 2008. Mechanisms of Korean red ginseng and herb extracts (KTNG0345) for anti-wrinkle activity. J. Ginseng Res. 32: 39-47.

Tanizawa Y, Tohno M, Kaminuma E, Nakamura Y, Arita M. 2015. Complete genome sequence and analysis of Lactobacillus hokkaidonensis LOOC260(T), a psychrotrophic lactic acid bacterium isolated from silage. BMC Genomics 16: 240.

VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, et al. 2015. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. Nature 527: 508-511.

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

Wolfgruber TK, Nakashima MM, Schneider KL, Sharma A, Xie Z, Albert PS, et al. 2016. High quality maize centromere 10 sequence reveals evidence of frequent recombination events. Front. Plant Sci. 7: 308.

Wong AS, Che C-M, Leung K-W. 2015. Recent advances in ginseng as cancer therapeutics: a functional and mechanistic overview. Nat. Prod. Rep. 32: 256-272.

Xie JT, Mehendale SR, Li X, Quigg R, Wang X, Wang CZ, et al. 2005. Anti-diabetic effect of ginsenoside Re in ob/ob mice. Biochim. Biophys. Acta. 1740: 319-325.

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

Yun TK. 2001. Brief introduction of Panax ginseng CA Meyer. J. Korean Med. Sci. 16: S3.