Diversity and authentication of Rubus accessions revealed by complete plastid genome and rDNA sequences

Young Sang Park, Jee Young Park, Jung Hwa Kang, Wan Hee Lee, and Tae-Jin Yang

ABSTRACT
Complete plastid genome (plastome) and ribosomal DNA (rDNA) sequences of three Rubus accessions (two Rubus longisepalus and one R. hirsutus) were newly assembled using Illumina whole-genome sequences. Rubus longisepalus Nakai and R. longisepalus var. tozawai, described as different varieties, have identical plastomes and rDNA sequences. The plastomes are 155,957 bp and 156,005 bp and the 45S rDNA transcription unit sizes are 5809 bp and 5811 bp in R. longisepalus and R. hirsutus, respectively. The 5S rDNA transcription unit is an identical 121 bp in three Rubus accessions. We developed three DNA markers to authenticate R. longisepalus and R. hirsutus based on plastome diversity. Phylogenomic analysis revealed that the Rubus species classified as two clades and R. longisepalus, R. hirsutus, and R. chingii are the most closely related species in clade 1.

Introduction
The genus Rubus consists of about 500 species, for which the taxonomy remains unclear due to frequent hybridizations, polyploidization, and asexual reproduction (Alice and Campbell 1999; Wang et al. 2016; Hytönen et al. 2018). The genus has been divided into 12 subgenera (Focke 1910, 1914). However, this classification is not unanimously supported, and each subgenus has been reported to be nonmonophyletic (Alice and Campbell 1999; Yang et al. 2012; Wang et al. 2016; Hummer et al. 2019). Even though previous studies contributed to current phylogenetic outline, short barcode regions such as internal transcribed spacer (ITS) and universal barcoding loci in the plastid genomes (plastome) have its own limitations (Li et al. 2015). Recently, nuclear genome and whole plastomes were used to analyze phylogenetic relationships among members of the genus Rubus and the chromosome scale genome assembly was released for R. occidentalis (VanBuren et al. 2016; Jibran et al. 2018; VanBuren et al. 2018; Hummer et al. 2019; Yang et al. 2021).

A super-barcoding approach using whole plastomes offers a solution to the limitations of using short barcoding regions to clearly distinguish inter- and intra-species diversity (Hollingsworth et al. 2009; Li et al. 2015). Since the plastome is inherited maternally in many plants, the absence of recombination preserves genome size, number of genes, and gene order in most plants (Palmer 1985; Wicke et al. 2011). However, sufficient variations are accumulated between species to allow estimation of their evolutionary path (Wolfe et al. 1987).

Nuclear ribosomal DNA (rDNA) exists in the plant nuclear genome in the form of thousands of tandem repeat arrays (Roa and Guerra 2012). Despite being part of the nuclear genome, its sequences are very conserved (Malinska et al. 2010). However, the internal transcribed sequences (ITS1 and ITS2) separating subunits of 45S rDNA (18S, 5.8S, and 28S) possess a meaningful level of variation among species (Álvarez and Wendel 2003). Whole-genome sequences produced by second- and third-generation sequencing platforms allow complete plastome and rDNA sequences to be assembled simultaneously in a time- and cost-effective manner (Kim et al. 2015a; Kim et al. 2015b). Comparison of plastomes and rDNA sequences have proved very useful for phylogenetic analysis and development of barcoding markers (Kim et al. 2017; Lee et al. 2019; Nguyen et al. 2020; Lee et al. 2021).

Rubus longisepalus Nakai, R. longisepalus var. tozawai (Nakai) T.B.Lee, are endemic to the Southern coasts and islands of the Korean Peninsula while R. hirsutus Thunb are distributed widely in Eastern Asia. R. longisepalus Nakai and R. longisepalus var. tozawai are regarded as distinct varieties with the common names ‘Macdo’ and ‘Geoje,’ respectively. R. hirsutus has a similar habitat and morphology as the two R. longisepalus varieties. Therefore, clear taxonomic identification and development of molecular markers are necessary for distinguishing these edible plant resources on the Korean Peninsula.
Material and methods

Plant materials and genome sequencing

Leaf samples of three Rubus accessions were provided from the Hantaek Botanical Garden, Gyeonggi-do, Republic of Korea. Each sample was ground into powder form using liquid nitrogen, and DNA was extracted using an Exgene Plant SV Midi Kit (Geneall Biotechnology, Seoul) following the manufacturer’s protocol. The extracted DNA was sequenced on the Illumina Miseq platform by Phyzen (www.phyzen.com, Seongnam, Gyeonggi-do). Approximately 1.3 Gbp paired-end sequence data were obtained for each of the three accessions.

Assembly and annotation of plastomes and rDNAs

Plastomes and 45S rDNA sequences were assembled using the de novo assembly of low-coverage whole-genome sequencing (dnaLCW) method (Kim et al. 2015b). To summarize, raw reads were trimmed using the trimming tool in CLC Assembly and then assembled de novo using the CLC novo assembly tool (CLC Inc, Denmark). Only contigs with similarity to the reference plastid genome (Rubus tridicus, NC_046585.1) were extracted using MUMmer (Kurtz et al. 2004). Contigs structurally identical to the reference plastome were then extracted, and assembly of the three Rubus plastomes was completed through manual curation. The complete plastomes were annotated using GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html), with manual curation using artemis (Carver et al. 2012; Tillich et al. 2017). Finally, a gene map was drawn using OGDRAW (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) (Greiner et al. 2019). The 45S rDNA sequences were assembled in the same way. Contigs similar to the reference (Sorbus commixta, MN215997.1) were selected and curated manually. After assembly, each subunit (18S, ITS1, 5.8S, ITS2, 28S) was determined using RNAmmer positions were modified. Intergenic spacer regions (IGS) in 45S rDNA and 5S rDNA were characterized by extending the end position of the rDNA unit through read mapping. Extension of the IGS proceeded until the IGS sequence met the start position of the next rDNA subunit. Manual curation was then conducted to obtain complete rDNA repeat sequences.

Polymorphism and marker development

The three completed chloroplast genomes and rDNA sequences were aligned using the MAFFT online version (Katoh et al. 2019). Plastome and rDNA variants were confirmed from the alignment results. Among the polymorphic regions, two single nucleotide polymorphisms (SNPs) and one insertion and deletion (InDel) region were selected for marker development. The two SNPs were developed into derived cleaved amplified polymorphic sequences (dCAPS) markers using dCAPS finder 2.0 (http://helix.wustl.edu/dcaps/) (Neff et al. 2002) and the InDel region was developed into a codominant marker. The three primer sets for these markers were validated in silico using NCBI primer blast (Ye et al. 2012) before adapting them to the three Rubus species (Table 1).

Phylogenetic analysis

A phylogenetic tree was reconstructed using coding sequences (CDSs) in the plastome. Sequences representing 11 additional species of the genus Rubus and three outgroup species also belonging to the family Rosaceae were obtained from NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Only 74 CDSs common to the 16 species were extracted by FeatureExtract (Wernersson 2005). These sequences were concatenated into one contig. The 16 CDS contigs were aligned using PRANK with the translate option (Löytynoja 2014), and a phylogenetic tree was reconstructed using the maximum-likelihood method in MegaX with 1000 bootstrap replicates (Kumar et al. 2018).

Results

Characteristics of complete plastomes

Assembled plastomes have distinct quadripartite structures consisting of one long single copy (LSC), one short single copy (SSC), and two inverted repeats (IRb and IRA). Rubus longisepalus Nakai and R. longisepalus var. tozawai have completely identical plastomes. Both have a total length of 155,957 bp, with 85,633 bp of LSC, 18,766 bp of SSC, and 25,779 bp of IR. The R. hirsutus plastome has a total length of 156,005 bp, with 85,745 bp of LSC, 18,734 bp of SSC, and 25,763 bp of IR. Both species have the same gene content and order: 85 CDSs, 37 tRNAs, and 8 rRNAs (Table 2; Figure 1). Analysis of nucleotide variations between R. longisepalus and R. hirsutus revealed 1882 SNPs and 325 InDels.

Table 1. Authentication markers and primers developed in this study.

| Primer          | Location | Product size (bp) | Recognition enzyme | Strand | Primer sequence                |
|-----------------|----------|------------------|--------------------|--------|-------------------------------|
| RubusdCAPS1     | psbA     | 125 (RL)          | Mbo                | F      | CCAAGGTAGCCGGCTTAA           |
| RubusdCAPS2     | atpI     | 148 (RH)          |                    | R      | GGCCTAGTTAGGTTATGGA          |
| RubusInDel1     | trnS–trnG| 162 (RL)          | Mbo                | R      | AGGATGGGTTTTTGTGAA          |
|                 |          | 137 (RH)          | Xho                | R      | GAAATCTACAGTACCTTCTCG        |
|                 |          | 226 (RL)          |                    | F      | GGGGCTTTTATGTTCGCC          |
|                 |          | 278 (RH)          |                    | R      | TGTGCACAGAAAACGACTTTCC       |

Mbo and Xho were used for the dCAPS markers. F and R, forward strand and reverse strand, respectively. RL and RH, R. longisepalus and R. hirsutus, respectively.
**Marker development**

We developed molecular markers based on the polymorphism between plastomes of *R. longisepalus* and *R. hirsutus*, and applied these to the three *Rubus* accessions. Sequence-based alignment of two dCAPS markers based on SNP regions and one codominant marker based on an InDel region confirmed their targets as polymorphic regions. All three markers could successfully distinguish *R. longisepalus* and *R. hirsutus* (Figure 2), validating the sequence assembly.

**Phylogenetic analysis**

To elucidate phylogenetic locations of *R. longisepalus* and *R. hirsutus*, plastomes of 11 additional species of the genus *Rubus* and three other species of the family Rosaceae were retrieved from NCBI GenBank. A total of 74 common CDSs were used to reconstruct and analyze a phylogenetic tree (Figure 3). Ten of the 13 *Rubus* species are classified into two subgenera in the

![Figure 1. Chloroplast gene map of *R. longisepalus* and *R. hirsutus*. The total length of plastomes ranges from 155,957 to 156,005 bp.](image)
GRIN database (https://npgsweb.ars-grin.gov/gringlobal/taxon/taxonomysearch): nine in the subgenus *Idaeobatus* and one in the subgenus *Malachobatus*. Meanwhile, our phylogenomic analysis classified the 13 *Rubus* species as two clades. Eight species including *R. longisepalus* and *R. hirsutus* fell into clade 1, with all species belonging to the monophyletic subgenus *Idaeobatus*, while the other five species belonged to clade 2, which is non-monophyletic and contains two subgenera. *R. lambertianus*, classified in subgenus *Malachobatus* based on GRIN database (https://npgsweb.ars-grin.gov/gringlobal/taxon/taxonomysearch), and three other *Rubus* species belonging to subgenus *Idaeobatus* were placed in clade 2. *R. longisepalus*, *R. hirsutus*, and *R. chinii* in clade 1 were the most closely related species among the 13 *Rubus* species studied.

**Nuclear rDNAs in *R. longisepalus* and *R. hirsutus***

We assembled complete rDNA units including transcription units and intergenic spaces (IGS) for all three *Rubus* accessions. The 45S rRNA and 5S rDNA units were assembled independently as repeated array forms. The 45S rDNA unit contains a transcription unit of 5809–5811 bp spanning 10,093 bp to 10,630 bp including IGS. The 5S rDNA transcription unit is 121 bp spanning 499 bp to 501 bp including IGS (Table 3; Figure 4). The transcription units in the 45S rDNA subunit are similar sizes in the two species, excluding ITS1 and ITS2, which are known to accumulate variations relatively fast. ITS1 and ITS2 of *R. hirsutus* are different from those of *R. longisepalus*. The 5S rDNA transcription unit
sequences are the same among all three accessions. The IGS of 45S rDNA are different among all three accessions, while the IGS of 5S rDNA are the same in the two R. longisepalus accessions but differ from those of R. hirsutus.

Discussion
Completion of three newly assembled Rubus plastomes and rDNA sequences allowed us to identify their polymorphisms and phylogenetic relationships. Two accessions of R. longisepalus, known to represent the same species but classified as different varieties, have identical plastomes and rDNA sequences. Despite large variations between R. longisepalus and R. hirsutus, they are the most closely related species among the 13 species of the genus Rubus studied. The majority of species in the genus Rubus belong to the subgenus Idaeobatus, with only one species classified as subgenus Malachobatus. Since most of the branches reconstructed in this study correspond with those obtained in previous studies, we conclude that the overall topology of our phylogenetic tree is reliable (Yang and Pak2006; Yang et al.2012; Wang et al. 2020; Yang et al. 2021). The genome data and barcode markers developed in this study provide a basis for unveiling the phylogenetic relationships of species of the genus Rubus worldwide.

Disclosure statement
The authors declare that there are no competing interests.

Funding
This research was supported by a research grant from the KIST ORP program [BlueBell Research program 2E30650].

ORCID
Tae-Jin Yang http://orcid.org/0000-0002-9676-8801

Data availability statement
The data that support the findings in this study are available at NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Plastome accession number are R. longisepalus (MW436703), R. hirsutus (MW448480). Accession number of 45S rDNA with IGS are R. longisepalus Nakai (MW474728), R. longisepalus var. tozawai (MW474727) and R. hirsutus (MW474729). Accession number of 5S rDNA with IGS are R. longisepalus Nakai (MW474730), R. longisepalus var. tozawai (MW474730) and R. hirsutus (MW474731). SRA accession number are R. longisepalus Nakai (SRR14026745), R. longisepalus var. tozawai (SRR14027373) and R. hirsutus (SRR14038231) under BioProject accession (PRJNA716145).

References
Alice LA, Campbell CS. 1999. Phylogeny of Rubus (Rosaceae) based on nuclear ribosomal DNA internal transcribed spacer region sequences. Am J Bot. 86(1):81–97.
Álvarez I, Wendel JF. 2003. Ribosomal ITS sequences and plant phylogenetic inference. Mol Phylogen Evol. 29(3):417–434.
Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 28(4):464–469.
Focke WO. 1910. Species Ruborum: monographiae generis rubi prodro-mus. Stuttgart: E. Schweizerbart.
Focke WO. 1914. Species Ruborum: monographiae generis rubi prodro- mus. Stuttgart: E. Schweizerbart.

Greiner S, Lehwark P, Bock R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organ-ellar genomes. Nucleic Acids Res. 47(W1):W59–W64.

Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, et al. 2009. A DNA barcode for land plants. Proc Natl Acad Sci U S A. 106(31):12794–12797.

Hummer KE, Carter KA, Liston A, Bassil NV, Alice LA, Bushakra JM, Hyten T, Graham J, Harrison R. 2018. The genomes of Rosaceous berr-ies and their wild relatives. New York: Springer.

Jibran R, Dzierzon H, Bassil N, Bushakra JM, Sullivan S, Finn CE, Dossett M, Vining KJ, VanBuren R, et al. 2018. Chromosome-scale scaffolding of the black raspberry (Rubus occidentalis L) genome based on chromatin interaction data. Hortic Res. 5:8641.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 20(4):1160–1165.

Kim K, Nguyen VB, Dong J, Wang Y, Park JY, Lee S-C, Yang T-J. 2017. Evolution of the Aralaceae family based from complete chloroplast genomes and 45S nrDNAs of 10 Panax-related species. Sci Rep. 7(1): 1–9.

Kim K, Lee S-C, Lee J, Lee HO, Joh HJ, Kim N-H, Park H-S, Yang T-J. 2015. Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within Panax ginseng species. PloS One. 10(6):e0117159.

Kim K, Lee S-C, Lee J, Yu Y, Yang K, Choi B-S, Koh H-J, Waminal NE, Choi H-I, Kim N-H, et al. 2015. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of Oryza AA genome species. Sci Rep. 5:15655.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 35(6):1547–1549.

Kurtz S, Phillips A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5(2):R12.

Lagesen K, Hallin P, Redland EA, Staerfeldt HH, Rogne T, Usery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35(9):3100–3108.

Lee HO, Joh HJ, Kim K, Lee S-C, Kim N-H, Park JY, Park H-S, Park M-S, Kim S, Kwak M, et al. 2019. Dynamic chloroplast genome rearrange-ment and DNA barcoding for three Apiaceae species known as the medicinal herb "Bang-Poong". UMS. 20(9):2196.

Lee YS, Kim J, Woo S, Park JY, Park H-S, Shim H, Choi H-I, Kang JH, Lee TJ, Sung SH, et al. 2021. Assessing the genetic and chemical diversity of Taraxacum species in the Korean Peninsula. Phytochemistry. 181:112576.

Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. 2015. Plant DNA barcoding: from gene to genome. Biol Rev Camb Philos Soc. 90(1):157–166.

Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. In: Russell DJ, editor. Multiple sequence alignment methods. New York: Springer, p. 155–170.

Malinska H, Tate JA, Matyasek R, Leitch AR, Soltis DE, Soltis PS, Kvarnik A. 2010. Similar patterns of rDNA evolution in synthetic and recently formed natural populations of Tragopogon (Asteraceae) allotetra-ploids. BMC Evol Biol. 10:291.

Neff MM, Turk E, Kalishman M. 2002. Web-based primer design for single nucleotide polymorphism analysis. Trends Genet. 18(12):613–615.

Nguyen VB, Linh Gia VN, Waminal NE, Park HS, Kim NH, Jang W, Lee J, Yang TJ. 2020. Comprehensive comparative analysis of chloroplast genomes from seven Panax species and development of an authenti-caction system based on species-unique single nucleotide polymorphism markers. J Ginseng Res. 44(1):135–144.

Palmer JD. 1985. Comparative organization of chloroplast genomes. Annu Rev Genet. 19(1):325–354.

Roa F, Guerra M. 2012. Distribution of 45S rDNA sites in chromosomes of plants: structural and evolutionary implications. BMC Evol Biol. 12(1): 225.

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017. GeSeq - versatile and accurate annotation of organ-elle genomes. Nucleic Acids Res. 45(W1):W6–W11.

vanBuren R, Bryant D, Bushakra JM, Vining KJ, Edger PP, Rowley ER, Priest HD, Michael TP, Lyons E, Filichkin SA, et al. 2016. The genome of black raspberry (Rubus occidentalis). Plant J. 87(6):535–547.

VanBuren R, Wai CM, Colle M, Wang J, Sullivan S, Bushakra JM, Liachko I, Vining KJ, Dossett M, Finn CE, et al. 2018. A near complete, chromo-some-scale assembly of the black raspberry (Rubus occidentalis) genome. Gigascience. 7(8):giy094.

Wang Q, Yu S, Gao C, Ge Y, Cheng R. 2020. The complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant Rubus chinii Hu. Mitochondrial DNA B. 5(2):1307–1308.

Wang Y, Chen Q, Chen T, Tang H, Liu L, Wang X. 2016. Phylogenetic insights into Chinese Rubus (Rosaceae) from multiple chloroplast and nuclear DNAs. Front Plant Sci. 7:968.

Wernersson R. 2005. FeatureExtract-extraction of sequence annotation made easy. Nucleic Acids Res. 33(Web Server issue):W567–W569.

Wicke S, Schneeeweiss GM, Depamphilis CW, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 76(3-5):273–297.

Wolfe KH, Li W-H, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A. 84(24):9054–9058.

Yang J, Chiang Y-C, Hsu T-W, Kim S-H, Pak J-H, Kim S-C. 2021. Characterization and comparative analysis among plastome sequences of eight endemic Rubus (Rosaceae) species in Taiwan. Sci Rep. 11(1): 12.

Yang J, Yoon H-S, Pak J-H. 2012. Phylogeny of Korean Rubus (Rosaceae) based on the second intron of the LEAFY gene. Can J Plant Sci. 92(3): 461–472.

Yang JY, Pak J-H. 2006. Phylogeny of Korean Rubus (Rosaceae) based on ITS (nrDNA) and trnL/F intergenic region (cpDNA). J Plant Biol. 49(1): 44–54.