Development of an SNP Marker Set for Marker-Assisted Backcrossing Using Genotyping-By-Sequencing in Tetraploid Perilla

Yun-Joo Kang

SEEDERS Inc

Bo-Mi Lee

SEEDERS Inc

Jangmi Kim

SEEDERS Inc

Moon Nam

SEEDERS Inc

Myoung-Hee Lee

Rural Development Administration

Keunpyo Lee

Rural Development Administration

Tae-Ho Kim

Rural Development Administration

Sunghwan Jo (✉ shjo@seeders.co.kr)

Seeders INC

Jeong-Hee Lee

Seeders INC

Research Article

Keywords: perilla, allotetraploid, genotyping-by-sequencing, marker-assisted backcrossing, crop breeding

Posted Date: December 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1177284/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

High-quality molecular markers are essential for marker-assisted selection to accelerate breeding progress. Compared with diploid species, recently diverged polyploid crop species tend to have highly similar homeologous subgenomes, which is expected to limit the development of broadly applicable locus-specific single-nucleotide polymorphism (SNP) assays. Furthermore, it is particularly challenging to make genome-wide marker sets for species that lack a reference genome. Here, we report the development of a genome-wide set of kompetitive allele specific PCR (KASP) markers for marker-assisted recurrent selection (MARS) in the tetraploid minor crop perilla. To find locus-specific SNP markers across the perilla genome, we used genotyping-by-sequencing (GBS) to construct linkage maps of two F₂ populations. The two resulting high-resolution linkage maps comprised 2,326 and 2,454 SNP markers that spanned a total genetic distance of 2,133 cM across 16 linkage groups and 2,169 cM across 21 linkage groups, respectively. We then obtained a final genetic map consisting of 22 linkage groups with 1,123 common markers from the two genetic maps. We selected 96 genome-wide markers for MARS and confirmed the accuracy of markers in the two F₂ populations using a high-throughput Fluidigm system. We confirmed that 91.8% of the SNP genotyping results from the Fluidigm assay were the same as the results obtained through GBS. These results provide a foundation for marker-assisted backcrossing and the development of new varieties of perilla.

Key Message

To make a genome-wide marker set for marker-assisted recurrent selection in tetraploid perilla, we identified markers covering all perilla chromosomes based on genotyping-by-sequencing genetic maps of two F₂ populations. We converted 96 selected markers to KASP markers for marker-assisted breeding and verified their accuracy in the F₂ populations with the Fluidigm system.

Introduction

Molecular markers enable efficient crop breeding through a variety of marker-assisted selection (MAS) approaches, including marker-assisted backcrossing (MABC), marker-assisted gene pyramiding, and marker-assisted recurrent selection (MARS) (Rossetto and Henry 2014). Linkage mapping studies of family-based populations, genome-wide association studies using natural diversity panels, and joint linkage-association mapping of both biparental and natural populations are the most widely used strategies to identify the underlying genetics of crop species. Genotyping-by-sequencing (GBS) can be used to create genetic maps and identify trait-associated markers, which, once validated, can be applied in MAS to combine genomic regions of interest (Rasheed et al. 2017). Marker-assisted pyramiding of quantitative trait loci (QTLs) that individually explain a small portion of the variance in a complex trait has been largely ineffective; however, phenotypic recurrent selection can be used to improve complex traits, although the long selection cycles and difficulty in identifying unique genes or alleles restrict its practicability (Ye and Smith 2008).

MARS is a strategy that allows targeted recombination of gene effects through recurrent selection. When MARS is used to introduce a desired target trait from a donor, it is important to accurately check the recovery rate of the background genome as well as the introduction of the target to shorten the time and increase the accuracy of inbred line development. When introducing a target trait from a wild species, it is necessary to minimize the unintentional introduction of additional wild traits by ensuring that only the locus related to the target trait is affected. Furthermore, in the case of polyploidy, it is necessary to confirm the degree of recovery by the subgenome and individual chromosomes after the new genetic material is introduced. Accordingly, a molecular marker set that can check the entire genome can be used to secure an inbred line with a high recovery rate in a relatively short time.

In order to develop MARS markers, genome sequence information must be assembled in a chromosome-level genome. Linkage maps can then be generated for segregated groups using GBS based on the assembled reference genome. These linkage maps can then be applied to select markers capable of identifying the highly similar structures of homeologous subgenomes. Polyploidy generally leads to high sequence conservation between homeologous genes, increased gene and genome dosage, and large genome size, which cause different layers of complexity for gene discovery and ultimately affect the efficient use of markers during breeding (Rasheed et al. 2017). Therefore, when developing markers in polyploid crops, it is essential to consider the redundancy present in the genome.

MARS markers are available in major crops including rice (Ahmadikhah et al. 2015), pepper (Kim et al. 2017), and tomatoes (Park et al. 2018), but not in the tetraploid minor crop perilla. Furthermore, there has been no report on the development of markers for MABC in tetraploid perilla. Perilla is an attractive crop used for oil or as a leafy vegetable because of its unique aroma and composition (Ahmed 2018; Asif 2011a; Oh 2018). Perilla has become increasingly popular in recent years as an economical source of omega-3 fatty acids, flavor ingredients, and flavonoids containing anthocyanins (Dhyani et al. 2019; Oh et al. 2018). There is increasing demand for molecular breeding to efficiently enhance these desirable traits in perilla varieties; however, as perilla is cultivated for food only in northeastern Asia, research on its genome and genetics is limited (Asif 2011b; Sa et al. 2018). In addition, cultivars of perilla are reported to be polyploid, so it is expected that there will be difficulties in genomic research and molecular marker development (Lee et al. 2018a; Nitta et al. 2005). GBS is considered an effective and widely applicable method for genotyping various crops [26-33]. Recently, a large number of single-nucleotide polymorphisms (SNPs) were successfully identified by GBS of interspecies cross-breeding groups of the diploid wild species *Perilla citriodora* (Kang et al. 2019).

In this report, we identified genome-wide SNPs by GBS of two F₂ populations of tetraploid perilla. We then constructed a universal map composed of 22 linkage groups and selected SNP markers for each individual chromosome. Next, we selected 96 markers for MARS, converted them to allelic-specific kompetitive allele specific PCR (KASP) markers, and verified their accuracy in the two F₂ populations using the Fluidigm BioMark HD genotyping system.

Materials And Methods

DNA extraction and library preparation for GBS

In order to secure sufficient candidate SNP marker loci, we selected two F₂ segregating populations by crossing *P. frutescens* var. *crispa* (Chajogi #107-1 or Pureunchajogi #115-2) and *P. frutescens* var. *frutescens* (Namcheon). One population, #973109, was obtained by crossing Chajogi #107-1 × Namcheon, and the other population, #973116, was obtained by crossing Pureunchajogi #115-2 × Namcheon. After self-fertilization of the F₁ generation of each population, F₂ lines were cultivated at the National Institute of Crop Science, RDA (Miryang, Korea) (Lee et al. 2018b).

We collected 96 fresh young leaves from each F₂ line and extracted genomic DNA using a modified cetyltrimethylammonium bromide-based method (Allen et al. 2006). We then used the extracted genomic DNA to prepare libraries for GBS and whole-genome sequencing. The DNA was quantified using a Thermo Scientific Nanodrop 8000 spectrophotometer (Fisher Scientific; Hampton, NH, USA). To prepare the GBS libraries, we digested the DNA using the restriction enzymes *Pst*I (CTGCAG) and *Msp*I (CCGG) and ligated the DNA from each F₂ line to a set of 96 barcoded adapters using a modified version of a previously published protocol (Elshire et al. 2011). We then pooled the libraries from each F₂ line and sequenced them in 151-bp reads using TrueSeq Version 3.0 paired-end sequencing on the Illumina platform HiSeq X Ten. Each lane of sequencing contained DNA from all 96 samples of a single F₂ line.

Sequencing and SNP filtration

After sequencing, we removed the reads containing only the common adapter using the Cutadapt software (Martin 2011). We then de-multiplexed the raw reads using a Python script, which filtered out reads containing ambiguous bases in the barcodes and split the raw Illumina FASTQ file into 96 separate FASTQ files based on the barcode sequences (Eun et al. 2016). Next, we trimmed the de-multiplexed reads using the Solexa QA package v.1.13 (Cox et al. 2010). Poor-quality reads and reads shorter than 25 bases were discarded. We applied the Burrows–Wheeler Aligner program (BWA; 0.6.1-r104) (Li and Durbin 2009) to align the clean reads to the draft genome sequence of *Perilla frutescens* (unpublished data) for SNP calling. The high mapping quality ensured reliable mapping of the reads, which is important for variant calling. Next, we extracted the mapped reads from the resulting BAM file for further analyses using SAMtools v.0.1.16 (Li et al. 2009). Using the SAM tools varFilter command, SNPs were called only for variable positions with a minimal mapping quality (i.e., -Q) of 30. The minimum and maximum read depths were set to 3 and 100, respectively. We used an in-house script that considered biallelic loci to select the significant sites among the called SNP positions (Kang et al. 2019; Kim et al. 2014; Kim et al. 2018). A raw SNP matrix was constructed by collecting the raw SNP positions obtained from comparison of each sample with the draft genome. Then, the blank region (non-SNP loci) was filled from the consensus sequence of the sample, and a final SNP matrix was created by filtering out mis-called SNP loci. Based on the ratio of the mapped reads to each SNP locus, we classified the SNPs into three types: Homozygous SNPs, read depth ≥ 90%; Heterozygous SNPs, 40% ≤ read depth ≤ 60%; and Other SNPs, which did not meet the criteria for Homozygous or Heterozygous. We then selected SNPs that had minor allele frequency (MAF) > 5% and missing data in less than 30% of the samples.

Linkage map construction

We converted the filtered SNPs in the F₂ populations to a genotype matrix consisting of four characters [maternal type: a; paternal type: b; heterozygous type: h; unknown: n]. We then constructed a genetic linkage map of *P. frutescens* var. *frutescens* separately using Joinmap 4 (Van Ooijen 2006). Also, we filtered out genetically identical markers and individuals using the genetic mapping program. The SNP markers were grouped into linkage groups (LGs) with logarithm of the odds threshold value ≥ 4. The genetic distances of the SNP markers, which were based on the recombination rates, were then converted using Kosambi's mapping function (Kosambi 2016). The final genetic linkage map was drawn using MapChart ver2.3 (Voorrips 2002).

SNP selection and validation by Fluidigm genotyping assay

SNP markers were selected from the markers constituting the two genetic maps. To design Fluidigm SNP genotyping assays, 60–150 bp sequences flanking the selected SNPs on either side were founded in the perilla draft genome sequence and checked for specificity using BLAST. Finally, the selected SNPs and flanking sequences were uploaded to the D3 Assay Design (<https://d3.fluidigm.com/>) website. After confirming the results, the designed assays were ordered. One Fluidigm SNP assay contains allele-specific primer 1 (ASP1), ASP2, locus-specific primer (LSP), and specific target amplification (STA) primer.

The SNP genotyping assays were performed in series using the BioMark™ HD system (Fluidigm, San Francisco, CA, USA) at SNP genetics (Seoul, Korea) according to the manufacturer's instructions using the following thermal cycling conditions: 95°C for 15 s, 64°C for 45 s, and 72°C for 15 s, with a touchdown of –1°C per cycle from 64°C to 61°C, followed by 34 cycles of 95°C for 15 s, 60°C for 45 s, and 72°C for 15 s. The genotyping results were acquired using the Fluidigm SNP Genotyping Analysis software. All base calls were manually verified, and any errors in homozygous or heterozygous clusters were curated.

Results

Identification of genome-wide SNPs using GBS

The summary statistics of the sequencing and GBS analysis are shown in **Table S1**. Paired-end sequencing produced approximately 105.4 Gbp (698,076,320 reads) from #973109 and 109.8 Gbp (727,703,068 reads) from #973116. Following removal of the barcodes and *Pst*I and *Msp*I overhang sequences, 684,566,618 and 715,551,854 short reads from #973109 and #973116, respectively, were de-multiplexed. After elimination of adapters and ambiguous nucleotides, 629,718,268 and 660,395,486 high-quality reads with Phred quality score ≥ 20 were obtained for #973109 and #973116, respectively. A total of 587,492,927 reads and 622,599,718 reads from #973109 and #973116, respectively, were mapped to the *P. frutescens* var. *frutescens* draft genome, which accounted for about 93.2% of the trimmed reads from #973109 and 94.2% of the trimmed reads from #973116. The average depth of the mapped reads in the GBS data sets was 108× and 103× for #973109 and #973116, respectively.

SNP filtering by linkage mapping

To secure biparental polymorphic SNPs in each population, we aligned the whole-genome sequencing data of Chajogi #107-1, Pureunchajogi #115-2, and Namcheon to the *P. frutescens* var. *frutescens* draft genome and identified genome-wide SNPs. In total, we identified 3,531,724 SNPs between Chajogi #107-1

and Namcheon and 3,659,790 SNPs between Pureunchajogi #115-2 and Namcheon. (Table 1). Subsequent analysis using our in-house GBS analysis pipeline considering read depth, mapping quality, and biallelic loci (Eun et al. 2016; Kim et al. 2014; Lee et al. 2018b) identified a total of 38,156 and 44,219 raw SNPs in #973109 and #973116, respectively (Table 1). These SNPs were filtered to identify putative markers using the criteria of 30% missing values across the genotyped individual and $MAF \geq 0.05$, which yielded a total of 5,052 and 5,462 SNPs for #973109 and #973116, respectively. Next, considering the biparental polymorphic SNPs in each population, we identified 2,840 markers in #973109 and 3,147 markers in #973116 (Table 1). After additional filtering using the mapping program with criteria such as chi-square-test and removal of duplicated markers and individuals, 2,783 and 3,072 markers were converted separately to genotyping matrixes for mapping of each GBS dataset.

Table 1
Single-nucleotide polymorphism (SNP) selection statistics

Filtration criteria	Number of SNPs	
	#973109	#973116
No. of polymorphic SNPs between parents (from resequencing)	3,531,724	3,659,790
No. of total SNPs from F2 populations by GBS	38,156	44,219
MAF (minor allele frequency) > 5%	12,744	15,086
Missing data < 30%	5,052	5,462
Polymorphic SNPs between parents	2,840	3,147
No. of SNP for genotyping	2,783	3,072

The genetic map of #973109 formed 16 LGs using 2,326 out of the 2,783 selected SNP markers (Table S2). The number of SNP markers constituting each group ranged from 25 (LG15) to 323 (LG16). The genetic distance (cM) of the LGs ranged from 47.9 cM for LG15 to 220.5 cM for LG08. The total genetic distance was 2,133.9 cM, and the average genetic distance between markers was 0.9 cM. The genetic map of #973116 formed 21 LGs using 2,454 out of the 3,072 selected markers (Table S3). The number of markers constituting each group ranged from 30 (LG10) to 244 (LG16). The genetic distance (cM) of the LGs ranged from 40.8 cM for LG10 to 167.0 cM for LG19. The total genetic distance was 2,169.3 cM, and the average genetic distance between markers was 1.0 cM. When we integrated the two genetic maps with the commonly used markers, nine paired LGs were matched one-to-one between the two genetic maps; however, for the remaining 19 LGs, different numbers of LGs were matched between the two maps (Table 2; Fig. 1). Several of the 19 LGs may be the result of the inability to distinguish between sub-genomes. However, by comparing the two genetic maps, the sub-genomes can be more clearly distinguished

Table 2
Statistics of single-nucleotide polymorphism (SNP) selection for MARS

#973109	#973116	No. of common SNP	No. of SNP by Fluidigm
LG02	LG07	74	7
LG01	LG05	12	3
LG01	LG20	54	3
LG14	LG21	27	5
LG08	LG01	100	3
LG08	LG02	65	4
LG03	LG15	45	5
LG13	LG12	65	5
LG12	LG18	49	6
LG12	LG17	48	4
LG06	LG13	49	4
LG05	LG14	57	4
LG16	LG16	143	5
LG16	LG11	78	5
LG03	LG09	42	3
LG04	LG10	19	4
LG09	LG06	27	4
LG11	LG06	29	2
LG10	LG06	57	5
LG07	LG19	34	9
LG05	LG04	31	4
LG05	LG08	18	2
sum.		1123	96

Marker selection for MARS

To select markers for MARS, we compared the 2,454 SNPs in #973116 to the 2,326 SNPs in #973109 and identified the SNPs that were common to both linkage maps. A total of 1,123 SNPs were represented in both F_2 populations (Table 2; Fig. 1). In addition, we identified 556 SNP markers by comparing the genetic maps of #973116 and #973109 with a genetic map of the recombinant inbred line developed from a cross between Daesil and Ipdeulkae 01 (unpublished data). We selected target loci from among the candidate SNPs by choosing those with no other SNPs or indels within 300 bp of the target region. In addition, we ensured that the selected SNP markers occurred at regular intervals along the genome so that the whole region of each LG was covered. The largest and smallest LGs had nine and two SNP markers, respectively. Thus, 96 SNP markers were appropriately assigned to individual LGs corresponding to the number of chromosomes.

High-throughput genotyping using Fluidigm

We checked the accuracy of the 96 selected SNP markers by genotyping the 192 samples of the F_2 lines #973116 and #973109 using the Fluidigm BioMark HD system. The primers used for the Fluidigm SNP assay are summarized in Table 3. Ninety-one (94.8%) of the 96 markers could successfully differentiate the ancestral alleles in the F_2 generation, although three of those had a call rate of 100 but a confidence value of 0 (Fig. 2). Out of a total of 18,432 high-throughput genotyping reactions, 17,403 reactions were successful, whereas 1,029 reactions were unsuccessful. In order to confirm whether the Fluidigm SNP genotyping results were the same as the GBS results, we identified 15,935 experiments present in both sets of genotyping results (Table 4). Of these common experiments, 14,658 (91.8%) gave the same result with both genotyping methods, whereas 1,277 (9.1%) gave different results. When we investigated the results in both F_2 lines, 191 reactions (15.2%) were Heterozygous-type markers in the GBS results but Homozygous-type markers in the Fluidigm results. Conversely, 1,071 reactions (83.9%) were Homozygous-type markers in the GBS results but heterozygous-type markers in the Fluidigm results (Table 4).

Table 3
Primer sequences used for 96 SNP assay

SNP_NAME	ASP1_SEQ	ASP2_SEQ	LSP_SEQ
PfSNP01	GGTGTCTCTGAGTGTCTTTGCTT	GGTGTCTCTGAGTGTCTTTGCTC	TGGTAGAACAAATGGGCGCTCT
PfSNP02	CTTTTGGATAAATGACAATCACTCAGTTTTTC	CCTTTTGGATAAATGACAATCACTCAGTTTTTA	CACCCTGCACTTCAGCATCAA
PfSNP03	CTACTGCTGATTATCACGGTGATCATA	CTGCTGATTATCACGGTGATCATG	ACAACGACACCACGGAAGGA
PfSNP04	TGTAGTTCGCCTCCATTACACA	TGTAGTTCGCCTCCATTACAG	CATGACAGCCAACTGCAGCTA
PfSNP05	GCTGAGAGAATCAAGGAGTATGACA	GCTGAGAGAATCAAGGAGTATGACC	GCCCTTTGCATAAACCGATATATGC
PfSNP06	TCTTTCCATCAAGAGTGGATCCAC	TCTTTCCATCAAGAGTGGATCCAG	GCTCATGCGGCTGAAGATGT
PfSNP07	AGCTATCTCCATTTGCTAAGATACACC	AGCTATCTCCATTTGCTAAGATACACA	ACATAAAGTTGAAAAATTAAGTGCA
PfSNP08	CGGCCAAGGACATTGGAGTA	GGCCAAGGACATTGGAGTG	TGAAGAGGCTGCAGTTACCA
PfSNP09	CACCTAGCAACCCCATCCC	CACCTAGCAACCCCATCCA	ACTGGCGGTGCATCTGGT
PfSNP10	GGATATCCAGTTCGTGTTCACCAT	GGATATCCAGTTCGTGTTACCAA	CTCCGGTGAGTGCAATGTGAT
PfSNP11	CGGAATCCCCAGCTGCTA	CGGAATCCCCAGCTGCTC	CAAGGCCGCTTTCTTCGCTA
PfSNP12	TTAACAAGTAGATCAACACTATCAACCCT	ACAAGTAGATCAACACTATCAACCCG	TCAGATTTCTACAGTTCAGTACTA
PfSNP13	TTTGAGGATGAAAGAAGCCGGA	TTGAGGATGAAAGAAGCCGGT	CCACATCGTGCAGCACAA
PfSNP14	GCTGTTACTGCATTTGTATTCCAACCTT	GCTGTTACTGCATTTGTATTCCAACCTC	GCTCCAGATTTATGTAAAGTGTTTT
PfSNP15	GCTTTGTTTGGACTTGAGGCA	GCTTTGTTTGGACTTGAGGCG	CGAGAGCAGCGCAAAGCA
PfSNP16	CCTAAGGAGTTTGAATAACCGGACG	CCTAAGGAGTTTGAATAACCGGACC	AGACAAGACGCACTCTGCCT
PfSNP17	GGAATCCTCTCTCTTGCTTCG	GGAATCCTCTCTCTTGCTTCC	TGGTCATGGAGACGAGGCATT
PfSNP18	CCTAGGACCCAGATAAACAAACATAGAT	CCTAGGACCCAGATAAACAAACATAGAA	CAGCATGCAGCATGCAACAA
PfSNP19	CACCTTTCAAACAGTTCCCATATTT	CACCTTTCAAACAGTTCCCATATTC	TCTTGATGTTCTGGTGAAGGAAGA
PfSNP20	GTCATGTTCTCTGAAGATCAAATGG	GTCATGTTCTCTGAAGATCAAATGC	TGGAAGCGGCATTGCGAGA
PfSNP21	GACAATCATGCTCTCTCAATCTATGAATT	GACAATCATGCTCTCTCAATCTATGAATC	AGACTGGAGACTTGTTTCCGCA
PfSNP22	CTTTATTCGGGTCTTCAGCTCTCT	CTTTATTCGGGTCTTCAGCTCTCA	CCTACTAGGCAACTACAGACTGACC
PfSNP23	CCTAATCATTGGTATCCTCCTCCG	CCTAATCATTGGTATCCTCCTCCT	AGGGTACACAGTTTCATGATAATGA
PfSNP24	GCAAATGGCGGCTGATTTCA	GCAAATGGCGGCTGATTTCT	GCGAATGCGTATAGCGCCG
PfSNP25	CACAAGGCATTTCCAGTCCC	CCACAAGGCATTTCCAGTCTCT	ACTTGGCGACGAACCGAGG
PfSNP26	GCGGTGCAGAAATGGACC	AGCGGTGCAGAAATGGACA	ATTCCCATTCTCTGCAGAAGGGT
PfSNP27	GAGCCTTGATTTGAGTCGACG	GGAGCCTTGATTTGAGTCGACA	GGTGACATCCGCTCTCTGT
PfSNP28	CATCACCTGTCATGATGATGAAAGAC	CATCACCTGTCATGATGATGAAAGAA	CAGAATACCGCTTGACGCCC
PfSNP29	ACCTTTGTGCGCTATAACTGTGTT	ACCTTTGTGCGCTATAACTGTGTC	GAAGGTGGACTTAGACTGTTCCCA
PfSNP30	CAATATCACAACAGTTTCGAAGTAGACAC	CAATATCACAACAGTTTCGAAGTAGACAG	TCAAATTAACGAAATGTGTGTTTT
PfSNP31	ACTGGGGATGAAATTGACTATATGTATCG	GACTGGGGATGAAATTGACTATATGTATCA	ACAGTTTACGGTGGATGAGAGCA
PfSNP32	CACCATTTCATCAACGTCTCTCG	GCACCATTTCATCAACGTCTCTCT	GAGCTGAGAGAGTCGCTCCA
PfSNP33	GGCATATGCTCTTTAGTAACTCCCT	GGCATATGCTCTTTAGTAACTCCCA	CTAACATATCTTGAGTTTCAAACAT
PfSNP34	GCACCCCTTGTTACAATGTTATGTAC	GCACCCCTTGTTACAATGTTATGTAT	TGGTCATGATATTCATAGTTGATAG
PfSNP35	GTTGCAAGCTTCCGTGGAG	GTTGCAAGCTTCCGTGGAA	GGCAAGGAAGCTATCATTGCCA
PfSNP36	CAGATTCATGTGCCGCATCA	CAGATTCATGTGCCGCATCG	GCTCATAGGTGAAGTCAAAATACCT
PfSNP37	CAATCGATTCTTTTGTAGGCCA	CAATCGATTCTTTTGTAGGCCG	GCTGCTGTCAAAAAGGAGCCT
PfSNP38	CTTTTCTTTGAGAGTGTGTCGATGC	TCTTTTCTTTGAGAGTGTGTCGATGT	CTTCCAACGGACCGCTCTCT
PfSNP39	GTACACCAAATCGAGTGATTCCAGTA	CACCAAATCGAGTGATTCCAGTG	TGTTGTTCTCTCTTCACGTAGTTG
PfSNP40	TTCGAAGCTCTTCTTACAGGTTACC	TTCGAAGCTCTTCTTACAGGTTACG	GACCAATTGACCTTTGATTTAAAT
PfSNP41	AAGTCGGACATCTAGTTCAATCACAT	AGTCGGACATCTAGTTCAATCACAC	CCAACCGTCCATTTCCATGGT

SNP_NAME	ASP1_SEQ	ASP2_SEQ	LSP_SEQ
PfSNP42	TCGACCATCTCTCAGGACCTAAAT	CGACCATCTCTCAGGACCTAAAC	TGACAAACATATTGTACAAGTGTTT
PfSNP43	TGAACGCGGCCATTCCA	GAACGCGGCCATTCCG	AGGTAATTGTTTAAGAATAAATTAT
PfSNP44	TGACGAGGGAAACCTGCG	ATGACGAGGGAAACCTGCA	CGAGACACACCCTCGACTCTC
PfSNP45	GTGAAATGTGGCAGCATTTGGA	TGAAATGTGGCAGCATTTGGG	TCACAGAAATGGCGTCCCGA
PfSNP46	CAATATTAATAATCCTGCAGCTAAGATCACT	CAATATTAATAATCCTGCAGCTAAGATCACC	AGTCGTGTTCTTGTAAGGAAGCA
PfSNP47	ACTCCAGGACATTCAAATCCTTAGAATC	ACTCCAGGACATTCAAATCCTTAGAATT	CTGCATTGAGCACCTACGCT
PfSNP48	GCCAGAACCTCTTCCAGAAAGAT	GCCAGAACCTCTTCCAGAAAGAC	TGGCAGGCGGGATCCAT
PfSNP49	CCATGCCTTCTTCATATCTCTATTCCA	CCATGCCTTCTTCATATCTCTATTCCG	CTCTAAGTGCCAAGACCCCGA
PfSNP50	ACGGAATGGAAACAGATGGGA	CGGAATGGAAACAGATGGGC	GCCGCCGCTCCACAT
PfSNP51	ATCGGGCTGCTAAAGCAGTTA	CGGGCTGCTAAAGCAGTTG	TCTTCCTTCGAGATTGTTTTGGTA
PfSNP52	CCCGCAATTCTCCGGAAG	CCCGCAATTCTCCGGAAG	TCTCCGTGATGAGGGCGTG
PfSNP53	GGTCGTTTGTGATGCAAAAGATCA	GGTCGTTTGTGATGCAAAAGATCG	CAGCCGCGAGAAGTCAGC
PfSNP54	AGAGCTCCGAGTGGATCAAGATA	GAGCTCCGAGTGGATCAAGATC	TGCTAGCCTTCAAATCCGCCA
PfSNP55	GAATTGATATGGATTTCTGCAGCTC	GAATTGATATGGATTTCTGCAGCTT	AGTGGAGTGCGTACTGCTTGT
PfSNP56	GACTTTAAATGCGAAAAAGGAGCAATTATC	GACTTTAAATGCGAAAAAGGAGCAATTATG	GGGTACTGCCCCACCCA
PfSNP57	GCCCGGTGTGTTCTTCACTA	CCCGGTGTGTTCTTCACTG	CACACCAACCCAGCAAATCCT
PfSNP58	CGGCAACTCTTCTGTCTCA	CGGCAACTCTTCTGTCTCG	TCTAGCTAGAATGTTTGAAGGTGA
PfSNP59	GATTCATCCAAGAAGAAAACATATCCACTTC	ATTCATCCAAGAAGAAAACATATCCACTTG	GGACAACAATTGTGCCGAGA
PfSNP60	GCCACCCGACGGGC	GCCACCCGACGGGT	TGGCCCAAGTGAAGACGA
PfSNP61	GCCGATTCTAGTTTCCCAAAGG	GCCGATTCTAGTTTCCCAAAGC	AGAAAGCGTGACCCCACTTGT
PfSNP62	GAGTTCGCGTTGATTGCAGT	AGTTCGCGTTGATTGCAGC	TCCGGTTCATGGAAGCACAGT
PfSNP63	CCAGACATCATCTGCATTTTGTAAACA	CCAGACATCATCTGCATTTTGTAACT	AGGTAGCGTACGTACTGATTTGCT
PfSNP64	TGGAAGTGTCAAGTTTCTAGCTTC	TGGAAGTGTCAAGTTTCTAGCTTT	ACCTCTATCAGTTCGAGGCCCTA
PfSNP65	GCTGCGCCCTCATCAAC	TGCTGCGCCCTCATCAAT	AAGCACAAACACTCTTTACAACATA
PfSNP66	GAGCGTACACCTACAGCCT	GAGCGTACACCTACAGCCC	CCGGCTGGAACATAACCAGG
PfSNP67	CAGTTTATCTGACCTGCACATTTGAC	CAGTTTATCTGACCTGCACATTTGAA	ACGTCAAATTAAGATCAGCTCTGG
PfSNP68	TTTTCTTGAAGGAAGGATCGTACA	TTCTTGAAGGAAGGATCGTACG	CTCAAGAGCAATGCAATGCCCA
PfSNP69	TGATGTTACTTGAGCTTGATTGCGT	GATGTTACTTGAGCTTGATTGCGC	AGAATGTCGTTGTAGCCGCGA
PfSNP70	CACTCAAGGTACTTCTTTAATTTTCTTTCTCTT	ACTCAAGGTACTTCTTTAATTTTCTTTCTCTG	GCTGCGATTGAAGGAAAGGCA
PfSNP71	CATGCAACCGTAACGGTCATATTT	CATGCAACCGTAACGGTCATATTC	GGCTGCATGGACAAACTGCG
PfSNP72	CTGAGGTCAGCTTTACGACCA	TGAGGTCAGCTTTACGACCG	GTGCTTCAACTTCGGCTGCA
PfSNP73	ACTATAAGCTCGACTTGTTTTCCG	ATACTATAAGCTCGACTTGTTTTCCA	GTGGACTTACACGCGCATGA
PfSNP74	CATTAAACCACAGAATCTACTGGTAGGATTA	AACCACAGAATCTACTGGTAGGATTG	CCACCACTGAGACAAAAGTTTGCT
PfSNP75	AGGAATGCAATCAACGTGAAAAGAC	AAGGAATGCAATCAACGTGAAAAGAT	ACCTCGAGAAAAGGCACTGGT
PfSNP76	CCAATTATTATGCCCTCATGCAATAGC	CCAATTATTATGCCCTCATGCAATAGT	AGTCCACGATTGGGATTGGT
PfSNP77	GATCAACATTAATGTTGGGGGTTATCAA	GATCAACATTAATGTTGGGGGTTATCAG	CCCTCTTCATTCTACACCGA
PfSNP78	GATGAAACACGGGGCCTA	TGGAACACGGGGCCTG	GCACACGAACCAAGCCT
PfSNP79	GTCCGTTGAAGTATGTGCTCCT	TCCGTTGAAGTATGTGCTCCA	TGCAGGTGTACATTTTAAGTGAGT
PfSNP80	AGACATTGTTAGCTCATCCAATTCATTAGA	ACATTGTTAGCTCATCCAATTCATTAGC	CCGGCATCTCCGACCTGAAA
PfSNP81	CTCCAGGTTGATATAAAGGGAGACC	CTCCAGGTTGATATAAAGGGAGACT	ACTCCCTGGGTCAACTTTCCC
PfSNP82	GCACAAGAGCAGGTAATTAACATACATC	GCACAAGAGCAGGTAATTAACATACATT	CCACACCTCCCCCTAACTGT
PfSNP83	TGACAAGAATGATTGAGATATCGACGT	GACAAGAATGATTGAGATATCGACGC	TTCTCATGAGTTCCGGAGAAGCA
PfSNP84	AAGACGGATAAGGACGTGAACAA	AAGACGGATAAGGACGTGAACAG	ACTCACTGAGCCTGGCAGC

SNP_NAME	ASP1_SEQ	ASP2_SEQ	LSP_SEQ
PfSNP85	CGACTTCAGAGACTGGCGAC	CGACTTCAGAGACTGGCGAG	ACGACGCGAAAGAGCTCCA
PfSNP86	GCTTGGTTTTGATGGATGAATCACAA	GCTTGGTTTTGATGGATGAATCACAG	GGCCTTGTAACCTCAGTGCAGC
PfSNP87	CAAGATGAAGCGGACATCATCTCTA	AAGATGAAGCGGACATCATCTCTT	CCTCCATGAGCAGACTCAGGA
PfSNP88	CCCTCACTCAGCAAAATTCTAGCA	CCCTCACTCAGCAAAATTCTAGCT	CATCAGAGATTCTGCAGGAGGGA
PfSNP89	GACGTGATTTCGGTGACAAAATGAT	ACGTGATTTCGGTGACAAAATGAC	TGGGTTACGTATATCTAGTTTCCGA
PfSNP90	GTAACAAAAGAAAAGAAAAGAAATAAGGACAGG	TGTAACAAAAGAAAAGAAAAGAAATAAGGACAGT	TCTAGAATTATATAATCTCCAAGCA
PfSNP91	CAGCTTCAGTTCGACCGGT	AGCTTCAGTTCGACCGGC	ACGTGATAGGTGCCTTGCCA
PfSNP92	CGGCCTAATAACATCCCATGTAACCTC	CGGCCTAATAACATCCCATGTAACCTT	GCCTCCTTGTGGCCAACAT
PfSNP93	CTAAACGAGCCAGCATTTTAAATGGTA	ACGAGCCAGCATTTTAAATGGTC	GCCACAAGCTCGCACCT
PfSNP94	AGAAATGAATTAGACCATCAAATTAAGTCCG	CAGAAATGAATTAGACCATCAAATTAAGTCTT	GACACCTTTGAATTAATTTGTACA
PfSNP95	GGGAAATCCCACATCACCGA	GGAAATCCCACATCACCGC	CCCTCACGGTGCAGAACTTG
PfSNP96	GGAAACCTCTCACATGGGGTA	GGAAACCTCTCACATGGGGTC	CTCCATCCGAGCCACGAAC

Table 4
SNP genotyping comparison of 96 markers using Fluidigm and GBS

Sum. of SNP assay	18432	-	-
No. of rxn for No Call in Fluidigm or GBS	2497	-	-
No. of rxn to comparable Fluidigm and GBS results	15935	100.0%	-
Identical genotyping in Fluidigm & GBS	14658	91.9%	-
Different genotyping in Fluidigm & GBS	1277	9.1%	100%
Case1: No. of rxn where Fluidigm (He) and GBS (Ho)	1071	-	83.9%
Case2: No. of rxn where Fluidigm (Ho) and GBS (He)	191	-	15.2%
Case3: No. of rxn where Fluidigm (Ho) and GBS (Ho)	12	-	0.9%

Discussion

MAS makes crop breeding programs faster and more accurate, especially if the life cycle of the crop is long, or there are not enough research personnel for the target crop. When multiple genes are introduced into different chromosomes, difficulties may arise due to unintentional dragging of nearby genes. Therefore, markers are needed that cover all the chromosomes evenly in order to investigate the recovery of the genetic background after new genes are introduced. This can be done either with a chromosome-level reference genome or with a complete linkage map of the genome. Although new reference genomes are being reported at a rapid rate, there are still many crops, especially minor crops, for which chromosome-level reference genomes are not available. It is always desirable to have a reference genome when designing MARS markers; however, it can be time consuming and costly to create a new high-quality reference genome. As an alternative approach, complete linkage maps can be used to develop MARS marker sets for crops lacking a suitable reference genome.

To create a set of MARS markers for perilla, which is a ploidy orphan crop with sparse background research and no available chromosome-level reference genome, we made a linkage map using F_2 populations. It is particularly challenging to select allelic-specific SNP markers that work well in polyploid crops, because it is not possible to determine how much similarity exists between subgenomes, especially in the absence of a high-quality reference genome. An important and difficult attempt to effectively distinguish each subgenome in polyploids such as bread wheat was made previously (Makhoul et al. 2020). Even with these difficulties, we were able to select appropriate candidate markers through creation of a linkage map. Although the 20 LGs, which equaled the number of perilla chromosomes, were not fully identified, a high success rate was confirmed in the experimental verification process while covering the entire chromosome.

The number of LGs differed depending on the population used to create the genetic map. Sixteen LGs were formed in #973109, whereas 21 LGs were formed in #973116 (Table S1 and Table S2). When the corresponding LGs were compared on the genetic maps of the two populations, LG01 in #973109 was divided into LG05 and LG20 in #973116. Conversely, LG06 in #973116 was divided into LG09, LG10, and LG11 in #973109 (Table 3). Combining two LGs into one LG or dividing one LG into two LGs is expected to fail to accurately separate the LG because of chromosomes with high similarity between subgenomes. This problem can be solved through genome assembly using long-read sequencing and comparison with genetic maps.

The application of PCR-based genotyping is essential for the rapid utilization of molecular markers in practical MAS breeding programs. We confirmed that the marker conversion from GBS markers to PCR-based KASP markers had a high success rate of over 91%. Thus, by applying the genetic map to perilla without a chromosome-level reference genome, we secured a molecular marker set that covered the entire genome. Our results show how a genetic map can be used to classify subgenomes in tetraploid crops. When we confirmed the converted SNP markers in the F_2 lines, 8.1% of the genotyping results for the markers were different between the GBS and the Fluidigm assay. If we look at the two sets of genotyping results in more detail, we find that certain markers

were more likely than others to differ between the two methods. Hence, this type of comparison may be a good way to check the accuracy of markers in the marker selection process.

Declarations

Ethics approval and consent to participate (Human Ethics, Animal Ethics or Plant Ethics): Not applicable.

Consent for publication: All authors consent to the publication of this paper.

Funding: This work was supported by a grant from the National Agricultural Genome Project (No. PJ013355032021), Rural Development Administration, Republic of Korea.

Competing Interests: On behalf of all authors, the corresponding author states that there are no relevant financial or non-financial conflicts of interest related to the content of this article.

Author contribution statement: YJ, BM, JK, SH, and JH conceived and designed the experiments; YJ and BM conducted the SNP analysis, linkage map construction, and data analysis; MN constructed the GBS library; MH developed the mapping population; TH provided the draft genome; KL participated in discussions about the experiments; SH and JH wrote and reviewed the manuscript.

Data Availability: The datasets generated during and/or analysed during the current study are available in the NABIC (National Agricultural Biotechnology Information Center, <https://nabic.rda.go.kr/>) repository, [NN-5662~5712].

Acknowledgements: Not applicable.

References

1. Ahmadikhah A, Mirarab M, Pahlevani MH, Nayyeripasand L. (2015) Marker-assisted backcrossing to develop an elite cytoplasmic male sterility line in rice. *The Plant Genome*. 8(2):plantgenome2014.2007.0031.
2. Ahmed HM. (2018) Ethnomedicinal, phytochemical and pharmacological investigations of *Perilla frutescens* (L.) britt. *Molecules*. 24(1).
3. Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF. (2006) A modified protocol for rapid dna isolation from plant tissues using cetyltrimethylammonium bromide. *Nat Protoc*. 1(5):2320-2325.
4. Asif M. (2011). Health effects of omega-3,6,9 fatty acids: *Perilla frutescens* is a good example of plant oils. *Orient Pharm Exp Med*. 11(1):51-59.
5. Bielenberg DG, Rauh B, Fan S, Gasic K, Abbott AG, Reighard GL, Okie WR, Wells CE. (2015) Genotyping by sequencing for snp-based linkage map construction and qtl analysis of chilling requirement and bloom date in peach [*Prunus persica* (L.) batsch]. *PLoS One*. 10(10):e0139406.
6. Cox MP, Peterson DA, Biggs PJ. (2010) Solexaqa: At-a-glance quality assessment of illumina second-generation sequencing data. *BMC Bioinformatics*. 11:485.
7. Dhyani A, Chopra R, Garg M. (2019) A review on nutritional value, functional properties and pharmacological application of perilla (*Perilla frutescens* L.). *Biomedical and Pharmacology Journal*. 12(2):649-660.
8. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. (2011) A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PLoS One*. 6(5):e19379.
9. Eun MH, Han JH, Yoon JB, Lee J. (2016) Qtl mapping of resistance to the cucumber mosaic virus p1 strain in pepper using a genotyping-by-sequencing analysis. *Horticulture, Environment, and Biotechnology*. 57(6):589-597.
10. Hussain W, Baenziger PS, Belamkar V, Guttieri MJ, Venegas JP, Easterly A, Sallam A, Poland J. (2017) Genotyping-by-sequencing derived high-density linkage map and its application to qtl mapping of flag leaf traits in bread wheat. *Sci Rep*. 7(1):16394.
11. Kang Y-J, Lee B-M, Nam M, Oh K-W, Lee M-H, Kim T-H, Jo S-H, Lee J-H. (2019) Identification of quantitative trait loci associated with flowering time in perilla using genotyping-by-sequencing. *Molecular biology reports*. 46(4):4397-4407.
12. Kim H, Yoon JB, Lee J. (2017) Development of fluidigm snp type genotyping assays for marker-assisted breeding of chili pepper (*Capsicum annuum* L.). *Horticultural Science & Technology*. 35(4):465-479.
13. Kim H-K, Cho Y-S, Yang J-W, Choi Y-W, Kang J-S, Lee Y-J, Son B-G. (2010) Analysis of genetic variation of *Perilla frutescens* var. *crispa* germplasm using RAPD. *Journal of Life Science*. 20(1):119-123.
14. Kim JE, Oh SK, Lee JH, Lee BM, Jo SH. (2014) Genome-wide snp calling using next generation sequencing data in tomato. *Mol Cells*. 37(1):36-42.
15. Kim WJ, Ryu J, Im J, Kim SH, Kang S-Y, Lee J-H, Jo S-H, Ha B-K. (2018) Molecular characterization of proton beam-induced mutations in soybean using genotyping-by-sequencing. *Molecular Genetics and Genomics*. 293(5):1169-1180.
16. Kosambi DD. (2016) The estimation of map distances from recombination values. *Dd kosambi*. Springer. p. 125-130.
17. Lee MH, Oh KW, Kim MS, Kim SU, Kim JI, Oh EY, Pae SB, Yeo US, Kim T-H, Lee JH. (2018). Detection of qtls in an interspecific cross between *Perilla citriodora* × *P. hirtella* mapping population. *Korean Society of Breeding Science*. 50(1):13-20.
18. Lee MH, Yang KW, Ha TJ, Jung CS, Pae SB, Hwang CD, Park CH, Baek IY, Kim HK, Park SK. (2011) Development of scar marker for identification of the perilla species. *Korean J Breed Sci*. 43(4):189-193.
19. Li H, Durbin R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*. 25(14):1754-1760.

20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. (2009) The sequence alignment/map format and samtools. *Bioinformatics*. 25(16):2078-2079.
21. Li L, Zhao S, Su J, Fan S, Pang C, Wei H, Wang H, Gu L, Zhang C, Liu G et al. (2017) High-density genetic linkage map construction by f2 populations and qtl analysis of early-maturity traits in upland cotton (*Gossypium hirsutum* L.). *PLoS One*. 12(8):e0182918.
22. Makhoul M, Rambla C, Voss-Fels K, Hickey L, Snowdon R, Obermeier C. (2020) Overcoming polyploidy pitfalls: A user guide for effective snp conversion into kasp markers in wheat. *Theoretical and Applied Genetics*. 133:2413-2430.
23. Martin M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* 17(1):3.
24. Nitta M, Lee JK, Kang CW, Katsuta M, Yasumoto S, Liu D, Nagamine T, Ohnishi O. (2005) The distribution of perilla species. *Genetic Resources and Crop Evolution*. 52(7):797-804.
25. Oh E, Lee MH, Kim JI, Kim S, Pae S-B, Ha TJ. (2018) Estimation of oil yield of perilla by seed characteristics and crude fat content. *Korean Journal of Crop Science*. 63(2):158-163.
26. Park G, Jang HA, Jo S-H, Park Y, Oh S-K, Nam M. (2018) Development of snp marker set for marker-assisted backcrossing (MABC) in cultivating tomato varieties. *Korean Journal of Agricultural Science*. 45(3):385-400.
27. Poland JA, Brown PJ, Sorrells ME, Jannink JL. (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*. 7(2):e32253.
28. Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, He Z. (2017) Crop breeding chips and genotyping platforms: Progress, challenges, and perspectives. *Molecular plant*. 10(8):1047-1064.
29. Rossetto M, Henry RJ. (2014) Escape from the laboratory: New horizons for plant genetics. *Trends in plant science*. 19(9):554-555.
30. Sa KJ, Choi IY, Park KC, Lee JK. (2018) Genetic diversity and population structure among accessions of *Perilla frutescens* (L.) britton in east asia using new developed microsatellite markers. *Genes Genomics*. 40(12):1319-1329.
31. Van Ooijen J. (2006) Joinmap® 4, software for the calculation of genetic linkage maps in experimental populations. *Kyazma BV, Wageningen*. 33(10.1371).
32. Verma S, Gupta S, Bandhiwal N, Kumar T, Bharadwaj C, Bhatia S. (2015) High-density linkage map construction and mapping of seed trait qtls in chickpea (*Cicer arietinum* L.) using genotyping-by-sequencing (GBS). *Sci Rep*. 5:17512.
33. Voorrips R. (2002) Mapchart: Software for the graphical presentation of linkage maps and qtls. *Journal of heredity*. 93(1):77-78.
34. Ye G, Smith KF. (2008) Marker-assisted gene pyramiding for inbred line development: Basic principles and practical guidelines. *Int J Plant Breed*. 2(1):1-10.

Figures

Figure 1

A perilla genetic linkage map with 1123 markers. A high-density genetic map was constructed by merging the genetic maps of the two F2 groups of Chajogi #107-1 × Namcheon (#973109) and Pureunchajogi #115-2 × Namcheon (#973116). Red triangles indicate the 96 markers selected for the MARS marker set.

Figure 2

SNP genotyping of 96 MARS markers with Fluidigm. Bars represent the SNP genotype ratios in the #973109 (A) and #973116 (B) F2 population. The solid line shows the auto confidence of the 96 markers.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SuppTables.docx](#)