misMM: An Integrated Pipeline for Misassembly Detection Using Genotyping-by-Sequencing and Its Validation with BAC End Library Sequences and Gene Synteny

Young-Joon Ko1, Jung Sun Kim2, Sangsoo Kim1*

1Department of Bioinformatics and Life Science, Soongsil University, Seoul 06978, Korea, 2Genomics Division, Department of Agricultural Biotechnology, National Institute of Agricultural Sciences, Rural Development Administration, Jeonju 54874, Korea

As next-generation sequencing technologies have advanced, enormous amounts of whole-genome sequence information in various species have been released. However, it is still difficult to assemble the whole genome precisely, due to inherent limitations of short-read sequencing technologies. In particular, the complexities of plants are incomparable to those of microorganisms or animals because of whole-genome duplications, repeat insertions, and Numt insertions, etc. In this study, we describe a new method for detecting misassembly sequence regions of *Brassica rapa* with genotyping-by-sequencing, followed by MadMapper clustering. The misassembly candidate regions were cross-checked with BAC clone paired-ends library sequences that have been mapped to the reference genome. The results were further verified with gene synteny relations between *Brassica rapa* and *Arabidopsis thaliana*. We conclude that this method will help detect misassembly regions and be applicable to incompletely assembled reference genomes from a variety of species.

Keywords: BAC end library, gene synteny, genotyping-by-sequencing, misassembly, next-generation sequencing, reference genome

Introduction

The genomics era has opened in earnest with the completion of the Human Genome Project. With the development of next-generation sequencing (NGS) technologies, the amount of genomics data has exploded, and sequencing targets have become very diverse. As of 2017, there are 7,930 species of eukaryotes, 192,677 species of bacteria, and 1,412 species of archaea that have been officially registered in NCBI. As the Nagoya Protocol is initiated, it is expected that these numbers will continue to increase in the future due to the policies of each country to secure information on biological genetic resources [1, 2]. Despite the fact that the cost of genomic analysis is declining, there are still a number of technical problems that make it difficult to sequence the genome completely [3]. For example, misassembly due to the inherent limitations of NGS technology is well known [4-6]. Especially in plants, there are many barriers that make plant genomes hard to sequencing, such as Numts, repeats, and genome duplication events [7-9].

Genotyping-by-sequencing (GBS) is a technology that allows high-throughput genotyping by applying NGS technology. It is used to analyze single nucleotide polymorphisms (SNPs) in populations to find molecular markers that are related to phenotype and genotype or to draw genetic linkage maps for plant breeding. By analyzing the pattern of GBS data along each chromosome, one can find out where the gene crossover occurs. On the other hand, a small block that interrupts an otherwise continuous GBS pattern is genetically non-ideal and implies a misassembled region. Therefore, we explored the application of GBS in the detection of misassemblies [10-12].

Brassicaceae is a mustard family containing 372 genera...
and 4,060 accepted species, and its varieties are cultivated as economically valuable crops not only in East Asia but also globally [13]. The triangle of U theory states that the differentiation of an allotetraploid of Brassica species—Brassica juncea (AABB), Brassica napus (AACC), and Brassica carinata (BBCC)—occurs due to the polyploidization of diploid Brassica species: Brassica rapa (AA), Brassica nigra (BB), and Brassica oleracea (CC). This theory has been proven by genomic analysis by NGS of Brassica species [14-25].

Research on the correlation between the genetic information and the nutrient content of crops has been actively conducted in Brassica genomes [26]. The recently published B. rapa V2.1 genome sequence shows much improved quality, as well as a number of misassembly corrections over the previous version, V1.5 [17]. This offers an interesting opportunity to test the potential of misassembly detection, based on GBS data.

In this study, we propose a user-friendly pipeline, called misMM, which automatically identifies misassembled candidate blocks (MCBs) and adjacent to destination blocks (ADBs) and plots the genetic map of MCBs by using raw GBS data. These results are verified by using the BAC end-sequence library published in NCBI and the gene synteny relation between Arabidopsis thaliana and B. rapa [28-31].

**Methods**

**Data source**

The end sequences of B. rapa accession Chiifu-401-42, a Chinese cabbage BAC library (KBrH, KBrB, and KBrS), were downloaded from NCBI and used to verify the putative misassembly genome regions. In order to investigate the gene synteny relation between A. thaliana TAIR10 and B. rapa genome V1.5, the corresponding general feature format (GFF) annotation files and protein sequences of each species were downloaded from http://ensemblgenomes.org and http://brassicadb.org, respectively. The GBS data were produced by a previous study that investigated the correlation between flavonoid content and the genotype of B. rapa in 69 individuals of a doubled haploid F2 generation obtained by microbial culture of an F1 generation cross of two subspecies—yellow sarson of LP08 (B. rapa ssp. tricolaris) and pak choi of LP21 (B. rapa ssp. chinensis)—with distinct morphologies [26]. From the study, genotype data were obtained at a total of 8,176 positions.

**Configuration of the misMM pipeline for misassembled block detection**

misMM, a pipeline for genome misassembled block detection, was written in a Linux shell and with Python ver. 2.7 in-house codes. The first step is preprocessing: after loading all GBS raw data files, markers with a missing value of over 8% were filtered out. If the neighboring positions had the similar GBS pattern with consistency, they were grouped into one block. Our script then automatically prepared the three kinds of input files (.loc, IDs, and maps) for MadMapper (UC Davis) [27], a package that specializes in recombinant inbred lines analysis using large genetic markers and easy visualizes the 2D pairwise matrix. The next step is the linkage grouping and block shuffling step, performed with MadMapper. By using the default parameters of MadMapper_RECBIT (rec_cut, 0.2; bit_cut, 100; data_cut, 25; allele_dist, 0.33; missing_data, 50; trio_analysis, TRIO; double_cross, 3), linkage grouping and marker extraction were performed by generating a pairwise matrix between GBS patterns of each block. Subsequently, block shuffling was performed by MadMapper_XDELA (marker fixation, FIXED; shuffle option, SHUFFLE; shuffle block, 6; shuffle step, 3) with each clustered block. At the end of this process, a pairwise matrix was made, and the output generated was used for generating the genetic map of MCBs.

![Fig. 1](image-url) Total workflow of misMM. GBS, genotyping-by-sequencing.
process, it plotted a genetic map diagram with putative misassembled blocks. In addition, it also generated 2D heatmap graphs for comparing before and after the block shuffling. All of the work flow of this pipeline is described in Fig. 1. The misMM pipeline scripts can be downloaded from http://sskimbnas.ipdisk.co.kr:80/publist/HDD1/misMM/misMM.tar.gz.

**Validation using BAC end sequences**

In order to confirm the misassembled blocks with experimental data, we extracted 41,969 pairs of end sequences from the BAC libraries (KBrS, KBrH, and KBrB) of *B. rapa* and carried out sequence alignment against the *B. rapa* reference genome sequence using Nucmer (MUMmer3.23) with the proper options (--maxmatch, use all anchor matches; -g, global alignment; -I, >95%; -r, sort output lines by reference). The Nucmer results were then filtered for discordant BAC end pairs with one end aligned to the MCB and the other end to the ADB.

**Validation using gene synteny relation between *A. thaliana* and *B. rapa***

For validation with gene synteny, the protein sequence of *B. rapa* were matched to those of *A. thaliana* using BLASTP (Blast 2.2.26), and the top four hits for each query were retained. The tabulated results were then sorted, based on the genomic coordinates of each protein, and the gene synteny relation was examined manually.

**Results and Discussion**

misMM was developed to provide a streamlined and yet simple-to-use pipeline for the detection of misassembled regions, so-called MCBs, based on GBS data (Fig. 1). This pipeline was tested with the GBS data of *B. rapa* against the *B. rapa* V1.5 reference genome, which is known to have some misassembled regions compared to the recently published V2.1 genome [17]. The original linkage score heatmap that was produced by MadMapper showed many off-diagonal cells with a low score that were often clustered in stretch (Fig. 2 left panel). The off-diagonal blocks scoring less than 0.33 were defined as MCBs (Table 1, Fig. 3). For each MCB, the corresponding ADB was identified by MadMapper, based on the linkage score (Table 1). The subsequent shuffled heatmap showed clean clustering, with no low-scoring off-diagonal blocks, implying the unambiguousness of the GBS pattern in detecting misassemblies (Fig. 2 right panel). The MCBs and ADBs were distributed throughout the entire pseudomolecule. A total of 16 MCBs had an average block size of 65,477 bp, and the largest one was 410,190 bp. The average size of the ADBs was 746,707 bp, with a maximum of 4,936,893 bp. The fact that only a few small MCBs were detected and that the corresponding ADBs were large in size implies that the *B. rapa* V1.5 genome is well assembled overall but has a few problematic regions, as shown by the recent update of the genome [17].

We used two sets of data to validate that the ADBs were indeed in the neighboring area of the MCBs. The first one was used to find discordant BAC end pairs with one end
| Block No. | Chr No. | Start position | End position | Block size (bp) | Chr No. | Start position | End position | Block size (bp) | Synteny relation | Count of BAC end |
|----------|---------|----------------|--------------|----------------|---------|----------------|--------------|----------------|-----------------|-----------------|
| 1        | A01     | 10,335,503     | 10,336,457   | 955            | A07     | 2,718,763      | 2,760,427    | 41,665         | No gene         | 1               |
|          |         |                |              |                |         | 2,970,361      | 3,340,395    | 370,035         |                 |                 |
|          |         |                |              |                |         | 4,284,326      | 5,685,009    | 1,400,684       |                 |                 |
|          |         |                |              |                |         | 5,782,516      | 8,114,350    | 2,331,835       |                 |                 |
|          |         |                |              |                |         | 8,306,623      | 8,390,460    | 83,838          |                 |                 |
|          |         |                |              |                |         | 8,462,236      | 9,063,378    | 601,143         |                 |                 |
| 2        | A01     | 11,453,104     | 11,488,558   | 35,455         | A04     | 3,271,457      | 4,978,203    | 1,706,747       | Related         | 6               |
|          |         |                |              |                |         | 5,227,803      | 6,734,498    | 1,506,696       |                 |                 |
|          |         |                |              |                |         | 7,605,871      | 7,605,928    | 58              |                 |                 |
| 3        | A01     | 11,830,981     | -            | 1              | A05     | 10,274,396     | 14,490,617   | 4,216,222       | Related         | 1               |
|          | A07     | 13,576,261     | -            | 1              |         | 14,602,065     | 14,704,957   | 102,893         |                 |                 |
|          | A08     | 1,389,252      | 1,419,543    | 30,292         |         | 15,735,698     | 7,090,412    | 121,934         |                 |                 |
|          |         |                |              |                |         | 7,782,948      | 551,732      |                 |                 |                 |
|          |         |                |              |                |         | 8,040,473      | 21,480       |                 |                 |                 |
| 4        | A01     | 17,853,386     | 17,853,417   | 32             | A03     | 28,233,583     | 28,599,515   | 365,933         | Related         | 2               |
|          | A01     | 21,422,470     | 21,756,693   | 334,224        |         | 28,622,787     | 29,191,693   | 568,907         |                 |                 |
|          | A02     | 26,385,973     | 26,386,023   | 51             |         | 29,806,067     | 31,527,446   | 1,721,380       |                 |                 |
| 5        | A01     | 23,266,604     | 23,424,555   | 157,952        | A06     | 10,280,840     | 10,357,155   | 76,316          | Related         | 13              |
|          | A02     | 13,440,136     | 13,440,137   | 2              |         | 10,732,633     | 14,236,176   | 3,503,544       |                 |                 |
|          | A02     | 21,066,162     | 21,066,274   | 113            |         | 14,450,457     | 14,559,524   | 109,068         |                 |                 |
|          |         |                |              |                |         | 8,950,753      | 10,162,388   | 1,211,636       |                 |                 |
| 6        | A01     | 8,706,169      | 8,950,670    | 244,502        | A09     | 11,293,419     | 11,528,445   | 235,027         | Related         | 63              |
|          | A06     | 19,457,789     | 19,703,630   | 245,842        |         | 11,610,344     | 14,668,929   | 3,058,586       |                 |                 |
|          |         |                |              |                |         | 14,915,372     | 15,794,376   | 879,005         |                 |                 |
|          |         |                |              |                |         | 15,949,568     | 18,361,629   | 2,412,062       |                 |                 |
|          |         |                |              |                |         | 19,064,188     | 22,487,944   | 3,423,757       |                 |                 |
|          |         |                |              |                |         | 22,634,782     | 23,337,316   | 702,535         |                 |                 |
|          |         |                |              |                |         | 23,489,828     | 23,489,836   | 9               |                 |                 |
| 7        | A02     | 21,427,161     | -            | 1              | A10     | 11,579,416     | 11,600       | 1,462,830       | No gene         | 0               |
|          |         |                |              |                |         | 1,638,829      | 1,766,618    | 839             |                 |                 |
|          |         |                |              |                |         | 1,792,156      | 5,224,789    | 5,489           |                 |                 |
|          |         |                |              |                |         | 5,224,789      | 1,538,439    |                 |                 |                 |
|          |         |                |              |                |         | 5,459,255      | 123,820      |                 |                 |                 |
|          |         |                |              |                |         | 5,693,352      | 44,601       |                 |                 |                 |
| 8        | A03     | 15,343,238     | -            | 1              | A08     | 2,368,697      | 3,803,367    | 1,434,671       | Related         | 2               |
|          |         |                |              |                |         | 4,037,929      | 4,357,798    | 319,870         |                 |                 |
|          |         |                |              |                |         | 4,835,708      | 48,204       |                 |                 |                 |
|          |         |                |              |                |         | 5,046,948      | 858,396      |                 |                 |                 |
|          |         |                |              |                |         | 7,510,164      | 384,146      |                 |                 |                 |
|          |         |                |              |                |         | 8,722,062      | 1,162,227    |                 |                 |                 |
|          |         |                |              |                |         | 9,002,058      | 1,550        |                 |                 |                 |
| 9        | A05     | 8,144,773      | 8,162,600    | 17,828         | A08     | 19,141,593     | 19,320,715   | 179,123         | Related         | 16              |
|          |         | 8,217,883      | 8,234,265    | 16,383         |         | 19,394,784     | 19,497,679   | 102,896         |                 |                 |
|          |         | 8,250,451      | 8,352,384    | 101,934        |         | 19,568,112     | 19,621,358   | 53,247          |                 |                 |
|          |         |                |              |                |         | 19,674,213     | 19,711,491   | 37,279          |                 |                 |
Table 1. Continued

| Block No. | Misassembled candidate block | Adjacent to destination block | Synteny relation | Count of BAC end |
|-----------|------------------------------|-------------------------------|------------------|-----------------|
|           | Chr No. | Start position | End position | Block size (bp) | Chr No. | Start position | End position | Block size (bp) |               |
| 10        | A05     | 9,669,449      | 10,079,638   | 410,190         | A01     | 10,376,926      | 10,494,252   | 117,327         | Related 23    |
|           |         |                |              |                 |         | 10,687,155      | 11,394,090   | 706,936         |                 |
|           |         |                |              |                 |         | 11,519,211      | 11,744,579   | 225,369         |                 |
|           |         |                |              |                 |         | 11,900,084      | 16,836,976   | 4,936,893       |                 |
|           |         |                |              |                 |         | 16,848,291      | 17,125,316   | 277,026         |                 |
|           |         |                |              |                 |         | 17,226,336      | 17,789,202   | 562,867         |                 |
|           |         |                |              |                 |         | 17,860,877      | 18,575,071   | 714,195         |                 |
|           |         |                |              |                 |         | 9,305,241       | 9,707,259    | 402,019         |                 |
|           |         |                |              |                 |         | 9,711,107       | 10,267,937   | 556,831         |                 |
| 11        | A07     | 2,319,220      | 2,321,114    | 1,895           | A01     | 24,324,484      | 24,353,432   | 28,949          | Related 0      |
|           |         |                |              |                 |         | 24,402,955      | 24,488,344   | 85,390          |                 |
|           |         |                |              |                 |         | 24,619,832      | 24,806,034   | 186,203         |                 |
|           |         |                |              |                 |         | 24,920,288      | 24,920,419   | 132             |                 |
| 12        | A07     | 3,920,950      | 4,009,069    | 88,120          | A10     | 11,579,416      | 1,638,829    | 1,462,830       | Related 0      |
|           | A08     | 3,927,665      | -            | 1               |         | 176,000         | 1,766,618    | 839             |                 |
|           |         |                |              |                 |         | 1,786,668       | 1,792,156    | 5,489           |                 |
|           |         |                |              |                 |         | 3,686,351       | 5,224,789    | 1,538,439       |                 |
|           |         |                |              |                 |         | 5,335,436       | 5,459,255    | 123,820         |                 |
|           |         |                |              |                 |         | 5,648,752       | 5,693,352    | 44,601          |                 |
| 13        | A07     | 8,271,542      | 8,274,604    | 3,063           | A03     | 12,032,914      | 12,032,953   | 40              | Related 5      |
|           |         |                |              |                 |         | 12,049,203      | 12,406,487   | 357,285         |                 |
|           |         |                |              |                 |         | 12,473,776      | 13,917,498   | 1,443,723       |                 |
|           |         |                |              |                 |         | 14,019,224      | 14,200,642   | 181,419         |                 |
|           |         |                |              |                 |         | 14,222,379      | 14,355,939   | 133,561         |                 |
| 14        | A08     | 11,266,789     | -            | 1               | A03     | 25,996,840      | 26,033,638   | 36,799          | Related 0      |
|           |         |                |              |                 |         | 26,067,147      | 27,037,677   | 970,531         |                 |
|           |         |                |              |                 |         | 27,139,966      | 27,943,662   | 803,697         |                 |
| 15        | A05     | 8,552,907      | 8,593,005    | 40,099          | A02     | 11,596,619      | 13,185,910   | 1,589,292       | Related 0      |
|           | A08     | 1,584,456      | 1,594,851    | 10,396          |         | 13,498,575      | 14,449,253   | 950,679         |                 |
|           |         |                |              |                 |         | 14,804,284      | 18,303,089   | 3,498,806       |                 |
|           |         |                |              |                 |         | 18,558,399      | 19,378,431   | 820,033         |                 |
|           |         |                |              |                 |         | 19,548,342      | 19,666,145   | 117,804         |                 |
|           |         |                |              |                 |         | 19,787,176      | -             | 1               |                 |
| 16        | A08     | 4,941,300      | 4,969,852    | 28,553          | A10     | 11,146,660      | 11,437,447   | 290,788         | Related 0      |
|           |         |                |              |                 |         | 11,664,229      | 11,814,368   | 49,742          |                 |
|           |         |                |              |                 |         | 11,764,627      | 12,032,475   | 104,223         |                 |

aligned to the MCB and the other end aligned to the ADB. For example, the MCB of block number 2 in Table 1 was located in pseudomolecule A01, ranging from 11,453,104 to 11,488,588, while its corresponding ADBs were found in A04. Table 2 shows the mapping results of the six BAC end pairs of this block, the sizes of which ranged from 671 bp to 1,000 bp, with a mapping identity higher than 97.93%. While one end of the BAC pairs was mapped to the corresponding MCB in A01, all of the other ends were mapped within the ADB, ranging from 3,271,457 to 4,978,203 in A04. Likewise, 10 out of 16 blocks listed in Table 1 could be confirmed by the BAC end results. The true locations of these blocks could be estimated within the span of the corresponding BAC (average 110 kbp). The rest could not be confirmed, probably due to the distance between the MCB and ADB, making it incompatible with the BAC size.

The other validation method was the use of the gene synteny relation. Compared to the *A. thaliana* genome, there...
is evidence that the *B. rapa* genome has undergone triplication [32]. Accordingly, most of the *A. thaliana* genes are preserved in gene syntenic blocks at three different places. Within block number 2 in Table 1, two *B. rapa* genes are annotated: Bra033489 and Bra033490 (Table 3). For all 16 genes flanking these two genes, orthologs were identified by BLASTP (Table 4). Eight *A. thaliana* genes in the middle—including the orthologs of two genes, AT4G14330 and AT4G14350—were out of order and broke the continuity of the synteny in the region. This is consistent with our finding that this MCB is truly misplaced in *B. rapa* genome V1.5. The true locations of the two *B. rapa* genes in this MCB can be inferred by mapping the flanking genes of AT4G14330 and AT4G14350 to the *B. rapa* genome (Table 5). Indeed, a total of six *A. thaliana* flanking genes were mapped to the *B. rapa* orthologs that were found in the corresponding ADBs. As expected, the gene synteny of this region is also well preserved. In this way, we can estimate the approximate relative locations of these two genes. Based on this relationship, an analysis was carried out with regard to the relationship of the protein orthologs and gene coordination between the two species. First, two genes were annotated in an example block (Table 3). When these two genes were found in a table arranged by the coordinates of the *B. rapa* gene, there was no continuity between the ortholog genes and the surrounding genes (Table 4). But, when we sorted this based on the coordination of *A. thaliana*, the ortholog genes belonging to the ADB were located consecutively around the gene belonging to the MCB (Table 5). Furthermore, the gene order that was inferred here was confirmed in the updated *B. rapa* V2.1 genome that was recently published [17].

In recent years, studies of expression quantitative trait loci that affect mRNA expression or protein expression using SNPs and studies to find markers that affect the environmental adaptation of plants have been becoming widely embraced [33]. For such works, accurate reference genome assembly is required. Toward that goal, our misMM pipeline is a useful tool for the identification of misassemblies in complex genomes using GBS data.

Table 3. Information on genes included in example misassembled candidate block

| Chr No. | Type | Start point | End point | *Brassica rapa* ID |
|---------|------|-------------|-----------|-------------------|
| A01     | Gene | 11,455,026  | 11,470,735| Bra033489         |
| A01     | Gene | 11,451,545  | 11,454,600| Bra033490         |

Table 2. Example of validation of BAC end library results

| BAC ends library ID | gi No. | Length (bp) | Identity (%) | Chr No. | Start position | End position |
|---------------------|--------|-------------|--------------|---------|----------------|--------------|
| KBrB037L22F         | 84732862 | 671         | 97.93        | A01     | 11,474,904     | 11,475,144   |
| KBrB037L22R         | 84732863 | 671         | 99.4         | A04     | 4,869,416      | 4,870,085    |
| KBrB039C19R         | 84732869 | 869         | 99.65        | A01     | 11,471,320     | 11,472,188   |
| KBrB039C19F         | 84732868 | 822         | 99.76        | A04     | 4,884,036      | 4,884,855    |
| KBrB043O24F         | 84737591 | 874         | 99.89        | A01     | 11,452,951     | 11,453,822   |
| KBrB043O24R         | 84737592 | 816         | 100          | A04     | 4,884,025      | 4,884,840    |
| KBrB077H15F         | 84762968 | 617         | 98.92        | A01     | 11,474,904     | 11,475,144   |
| KBrB077H15R         | 84762969 | 646         | 100          | A04     | 4,884,386      | 4,885,085    |
| KBrB097P17F         | 114827207| 1,000       | 98.2         | A01     | 11,471,303     | 11,472,294   |
| KBrB097P17R         | 114827208| 937         | 98.16        | A04     | 4,883,252      | 4,884,169    |
| KBrH087A11R         | 84341421 | 831         | 99.88        | A01     | 11,466,761     | 11,467,587   |
| KBrH087A11F         | 84341072 | 844         | 99.63        | A04     | 4,977,838      | 4,978,643    |
Table 4. Example of protein ortholog list, sorted by Brassica rapa gene coordination

| ID          | Chr No. | Start position | End position | ID          | Chr No. | Start position | End position | Comments       |
|-------------|---------|----------------|--------------|-------------|---------|----------------|--------------|----------------|
| Bra033497   | A01     | 11,382,249     | 11,386,827   | AT4G15570   | Chr4    | 8,892,607     | 8,898,999    |                |
| Bra033496   | A01     | 11,388,925     | 11,390,027   | AT4G15563   | Chr4    | 8,890,879     | 8,892,526    |                |
| Bra033495   | A01     | 11,393,659     | 11,396,663   | AT4G15560   | Chr4    | 8,883,907     | 8,887,565    |                |
| Bra033494   | A01     | 11,410,610     | 11,412,043   | AT4G15550   | Chr4    | 8,877,590     | 8,879,327    |                |
| Bra033493   | A01     | 11,412,702     | 11,414,443   | AT4G15545   | Chr4    | 8,875,918     | 8,877,799    |                |
| Bra033492   | A01     | 11,445,862     | 11,446,743   | AT5G49420   | Chr5    | 20,034,674    | 20,036,170   |                |
| Bra033491   | A01     | 11,450,091     | 11,451,172   | AT4G14320   | Chr4    | 8,241,732     | 8,243,910    |                |
| Bra033490   | A01     | 11,451,545     | 11,454,600   | AT4G14330   | Chr4    | 8,244,194     | 8,247,444    | Misassembled  |
| Bra033489   | A01     | 11,455,026     | 11,470,735   | AT4G14350   | Chr4    | 8,256,086     | 8,260,787    | Misassembled  |
| Bra039534   | A01     | 11,504,946     | 11,505,630   | AT2G35280   | Chr2    | 14,859,378    | 14,860,200   |                |
| Bra039535   | A01     | 11,504,946     | 11,505,422   | AT2G35280   | Chr2    | 14,859,378    | 14,860,200   |                |
| Bra039536   | A01     | 11,504,946     | 11,505,630   | AT2G35280   | Chr2    | 14,859,378    | 14,860,200   |                |
| Bra039538   | A01     | 11,510,855     | 11,512,648   | AT3G59380   | Chr3    | 21,944,178    | 21,945,943   |                |
| Bra039539   | A01     | 11,514,776     | 11,515,144   | AT4G15530   | Chr4    | 8,864,828     | 8,870,967    |                |
| Bra039540   | A01     | 11,516,583     | 11,521,200   | AT4G15530   | Chr4    | 8,864,828     | 8,870,967    |                |
| Bra039541   | A01     | 11,521,728     | 11,523,067   | AT4G15520   | Chr4    | 8,862,815     | 8,864,618    |                |

Alternative alignments due to genome triplication have been removed. ADB, adjacent to destination block; MCB, misassembled candidate blocks.

ORCID: Young-Joon Ko: http://orcid.org/0000-0002-2386-6355; Jung Sun Kim: http://orcid.org/0000-0001-7945-0829; Sangsoo Kim: http://orcid.org/0000-0001-9836-9823

Authors’ contribution

Conceptualization: YJK, JSK
Data curation: YJK
Formal analysis: YJK
Funding acquisition: SK
Methodology: YJK
Writing – original draft: YJK
Writing – review & editing: SK, JSK

Acknowledgments

This work was funded by a program (PJ01167402) of the RDA (Rural Development Administration) and a program (NRF-2012M3A9D1054705) of the NRF (National Research Foundation of Korea) and computationally supported by a program (NRF-2010-0018156, NTIS1711048528) of the KISTI GSDC.

References

1. Comizzoli P, Holt WV. Implications of the Nagoya Protocol for genome resource banks composed of biomaterials from rare and endangered species. *Reprod Fertil Dev* 2016 Feb 24 [Epub]. https://doi.org/10.1071/RD15429.
2. Schindel DE, du Plessis P. Biodiversity: reap the benefits of the
Genomics & Informatics Vol. 15, No. 4, 2017

Nagoya Protocol. Nature 2014;515:37.

3. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shen-dure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 2013;31:1119-1125.

4. Muggli MD, Puglisi SJ, Ronen R, Boucher C. Misassembly detection using paired-end sequence reads and optical mapping data. Bioinformatics 2015;31:i80-188.

5. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 2008;9:R55.

6. Zhu X, Leung HC, Wang R, Chin FY, Yiu SM, Quan G, et al. misFinder: identify mis-assemblies in an unbiased manner using reference and paired-end reads. BMC Bioinformatics 2015;16:386.

7. Ko YJ, Kim S. Analysis of nuclear mitochondrial DNA segments of nine plant species: size, distribution, and insertion Loci. Genomics Inform 2016;14:90-95.

8. Freeing M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol 2009;60:433-453.

9. Yi M, Cho YS, Guang X, Kang SG, Jeong JY, Cha SS, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet 2014;46:88-92.

10. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 2011;6:e19379.

11. Poland JA, Brown PJ, Sorrells ME, Jannink JL. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 2012;7:e32253.

12. The Plant List. Version 1.1. Published on the internet. The Plant List 2013. Accessed 2017 Oct 1. Available from: http://www.theplantlist.org.

13. Nagaharu U. Genome analysis in Brassica. Plant genetics: early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 2014;345:950-953.

14. Elshire RJ, Glaubitz JC, Ronen R, Boucher C. Misassembly detection using paired-end sequence reads and optical mapping data. Bioinformatics 2015;31:i80-188.

15. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 2008;9:R55.

16. Zhu X, Leung HC, Wang R, Chin FY, Yiu SM, Quan G, et al. misFinder: identify mis-assemblies in an unbiased manner using reference and paired-end reads. BMC Bioinformatics 2015;16:386.

17. Ko YJ, Kim S. Analysis of nuclear mitochondrial DNA segments of nine plant species: size, distribution, and insertion Loci. Genomics Inform 2016;14:90-95.

18. Freeing M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol 2009;60:433-453.

19. Yi M, Cho YS, Guang X, Kang SG, Jeong JY, Cha SS, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet 2014;46:88-92.

20. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 2011;6:e19379.

21. Poland JA, Brown PJ, Sorrells ME, Jannink JL. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 2012;7:e32253.

22. The Plant List. Version 1.1. Published on the internet. The Plant List 2013. Accessed 2017 Oct 1. Available from: http://www.theplantlist.org.

23. Nagaharu U. Genome analysis in Brassica. Plant genetics: early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 2014;345:950-953.

24. Elshire RJ, Glaubitz JC, Ronen R, Boucher C. Misassembly detection using paired-end sequence reads and optical mapping data. Bioinformatics 2015;31:i80-188.

25. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 2008;9:R55.

26. Yi M, Cho YS, Guang X, Kang SG, Jeong JY, Cha SS, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet 2014;46:88-92.

27. The Plant List. Version 1.1. Published on the internet. The Plant List 2013. Accessed 2017 Oct 1. Available from: http://www.theplantlist.org.

28. Nagaharu U. Genome analysis in Brassica. Plant genetics: early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 2014;345:950-953.

29. Elshire RJ, Glaubitz JC, Ronen R, Boucher C. Misassembly detection using paired-end sequence reads and optical mapping data. Bioinformatics 2015;31:i80-188.

30. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 2008;9:R55.

31. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 2011;6:e19379.

32. The Plant List. Version 1.1. Published on the internet. The Plant List 2013. Accessed 2017 Oct 1. Available from: http://www.theplantlist.org.

33. Nagaharu U. Genome analysis in Brassica. Plant genetics: early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 2014;345:950-953.

34. Elshire RJ, Glaubitz JC, Ronen R, Boucher C. Misassembly detection using paired-end sequence reads and optical mapping data. Bioinformatics 2015;31:i80-188.

35. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 2008;9:R55.

36. Yi M, Cho YS, Guang X, Kang SG, Jeong JY, Cha SS, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet 2014;46:88-92.

37. The Plant List. Version 1.1. Published on the internet. The Plant List 2013. Accessed 2017 Oct 1. Available from: http://www.theplantlist.org.

38. Nagaharu U. Genome analysis in Brassica. Plant genetics: early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 2014;345:950-953.

39. Elshire RJ, Glaubitz JC, Ronen R, Boucher C. Misassembly detection using paired-end sequence reads and optical mapping data. Bioinformatics 2015;31:i80-188.

40. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 2008;9:R55.

41. Yi M, Cho YS, Guang X, Kang SG, Jeong JY, Cha SS, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet 2014;46:88-92.

42. The Plant List. Version 1.1. Published on the internet. The Plant List 2013. Accessed 2017 Oct 1. Available from: http://www.theplantlist.org.

43. Nagaharu U. Genome analysis in Brassica. Plant genetics: early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 2014;345:950-953.

44. Elshire RJ, Glaubitz JC, Ronen R, Boucher C. Misassembly detection using paired-end sequence reads and optical mapping data. Bioinformatics 2015;31:i80-188.

45. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 2008;9:R55.

46. Yi M, Cho YS, Guang X, Kang SG, Jeong JY, Cha SS, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet 2014;46:88-92.

47. The Plant List. Version 1.1. Published on the internet. The Plant List 2013. Accessed 2017 Oct 1. Available from: http://www.theplantlist.org.

48. Nagaharu U. Genome analysis in Brassica. Plant genetics: early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 2014;345:950-953.

49. Elshire RJ, Glaubitz JC, Ronen R, Boucher C. Misassembly detection using paired-end sequence reads and optical mapping data. Bioinformatics 2015;31:i80-188.

50. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 2008;9:R55.

51. Yi M, Cho YS, Guang X, Kang SG, Jeong JY, Cha SS, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet 2014;46:88-92.

52. The Plant List. Version 1.1. Published on the internet. The Plant List 2013. Accessed 2017 Oct 1. Available from: http://www.theplantlist.org.

53. Nagaharu U. Genome analysis in Brassica. Plant genetics: early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. Science 2014;345:950-953.