Development of SNP set derived from single copy genes for MAB in *Brassica oleracea*

SEUNG WOO JIN

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

ABSTRACT

Brassica oleracea is one of the 10 most cultivated crops in the world. According to reports, the competitiveness of Korean traditional breeding is world-class for *B. oleracea*. With more advanced molecular breeding, Korea is expected to have global competitiveness in *B. oleracea*. For the development of SNP array sets, resequencing data of 44 accessions that have various agricultural traits was mapped to the reference genome. From more than 3 million SNPs, SNPs that were genotyped from more than 40 accessions are filtered. After the filtering steps that enhance the accuracy of markers, 3,446 SNPs from genic regions were filtered. Since paralogous sequences can interfere with the fluorescent signal, paralogous sequences of flanking sequence containing the SNPs were filtered out. From 849 SNPs in single copy genes, 240 markers that are distributed throughout the genome were designed. After SNP array experiments,

192 markers that have high accordance rate with the resequencing data were selected, and these markers were applied to 92 samples from LG Farmhannong. 150 markers that can be used in marker assisted breeding (MAB) were ultimately selected with the criteria of no call rate and clustering form of data points. The phylogenetic trees that were drawn with 150 markers cluster cabbage in group 1 and other *B. oleracea* subspecies in group 2, and the result matched well with the phenotype data. These results showed that these markers are working well and can instantly be utilized to facilitate the shortening of the breeding period.

Keywords: *Brassica oleracea*, SNP array, MAB

Student number: 2016-28934

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SNP	Single nucleotide polymorphism
MAB	Marker assisted breeding
QTL	Quantitative trait locus
CTAB	Cetyltrimethylammonium bromide
STA	Specific target amplification
PCR	Polymerase chain reaction
IFC	Integrated fluidic circuit
ASP	Allele specific primer
LSP	Locus specific primer
NTC	No template control

INTRODUCTION

Backcross breeding, first introduced in 1922, is one of the effective ways for introducing one or a few genes to an elite cultivar (Stoskopf et al. 1993). However, traditional backcrossing is a time-consuming way because it requires multiple crossing events. Marker assisted breeding (MAB) is an indirect selection process in which a trait of interest is selected based on a marker. Compared to traditional cross-breeding, MAB can shorten the breeding period by selecting the individual that has a similar genome with the recurrent parent (Hospital et al. 1992). It can also choose traits related to Quantitative Trait Locus (QTL) that are hard to detect with the eyes with the aid of molecular markers (Collard et al. 2008). Through these steps, MAB can greatly reduce the number of crossing times and increase breeding efficiency.

B. oleracea is one of the 10 major horticultural crops. It is comprised of many subspecies such as cabbage, cauliflower, kohlrabi, broccoli, etc., and most of these subspecies have high demands in the world. More than 76 million tons of *Brassica* vegetables are produced annually (http://faostat.fao.org/). The species in the *Brassica* genus contributes to human nutrition as well as plant evolutionary studies (Liu et al, 2014).

B. oleracea is a diploid plant but has hexaploid traits due to a few whole genome duplications and whole genome triplications throughout evolution (Liu et al, 2014). Such polyploidy events are known to affect genome structure and gene expression (Jiao et al. 2011, Doyle et al. 2010, Liu et al, 2014). Moreover paralogous sequences make developing molecular markers difficult. Due to the way primers work, primers can bind to undesirable positions.

To resolve this issue, we designed Fluidigm SNP arrays derived from single copy genes that have no paralogous sequences by removing them via Blast

search. With this method, we developed 150 markers with high recall rate. These markers were validated by applying them to 92 unknown samples provided by Farmhannong (Farmhannong Co., Ltd., Anseong-si, Gyeonggi-do, 17503, Korea)

These markers can instantly be used for facilitating MAB. Moreover, since these genes are derived from single copy genes, they can be directly linked to their phenotypes. These markers are expected to be utilized in further genetic research on *B. oleracea*.

Materials & Methods

1. Plant materials and SNP filtering process

In the previous study, 202 accessions were collected from various companies. Among them, 44 accessions with various agricultural traits were selected (Lee et al. 2015, Lee et al. 2016). Genomic DNA of these accessions were extracted from 2g samples of young leaves, using the cetyltri-methylammonium bromide (CTAB) protocol (Allen et al. 2006). The quantity and quality of the DNA were examined with NanoDrop ND-1000 (NanoDrop Technologies, inc., USA). Resequencing data of these accessions was produced with Illumina Hiseq 2000 and NextSeq with an average coverage of 6~7x. The reads were mapped to the reference genome completed in 2014 (Liu et al. 2014). By comparing the sequences of accessions and reference genome, more than 13 million SNPs were extracted.

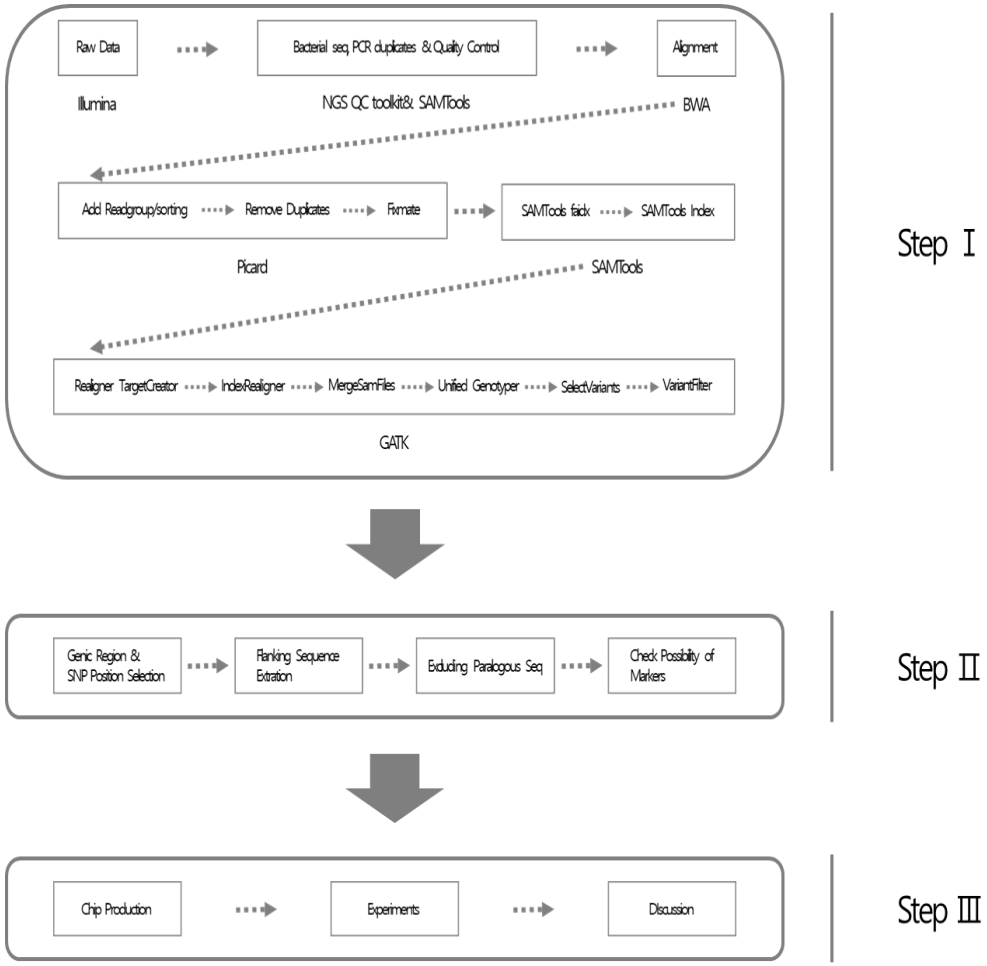


Figure 1. Overall workflow of SNP set development.

Table 1. Information of accessions

No	AC.No	Sub-species	Sequencer	Read Count	Table Bases	Coverage	NABIC accessions number
1	IM_Bol_01	Cabbage	HiSeq 2000	114,454,524	11,559,906,924	6.2x	NN-2496
2	IM_Bol_02	Cabbage	HiSeq 2000	113,830,992	11,496,930,192	6.2x	NN-2498
3	IM_Bol_03	Cabbage	HiSeq 2000	126,837,064	12,810,543,464	6.2x	NN-2495
4	IM_Bol_04	Cabbage	HiSeq 2000	100,552,700	10,155,822,700	6.2x	NN-2497
5	IM_Bol_05	Cabbage	HiSeq 2000	38,709,840	3,909,693,840	6.1x	NN-2499
6	IM_Bol_06	Cabbage	HiSeq 2000	36,389,648	3,675,354,448	5.7x	NN-2528
7	IM_Bol_07	Cabbage	HiSeq 2000	42,569,026	4,299,471,626	6.7x	NN-2531
8	IM_Bol_08	Cabbage	HiSeq 2000	44,600,458	4,504,646,258	7.0x	NN-2532
9	IM_Bol_09	Cabbage	NextSeq 500	28,561,204	4,296,548,943	6.7x	NN-2537
10	IM_Bol_10	Cabbage	NextSeq 500	32,094,538	4,827,824,278	7.5x	NN-2538
11	IM_Bol_11	Cabbage	NextSeq 500	34,441,430	5,180,936,024	8.0x	NN-2500
12	IM_Bol_12	Cabbage	NextSeq 500	25,347,188	3,812,412,023	5.9x	NN-2539
13	IM_Bol_13	Cabbage	NextSeq 500	20,362,678	3,062,717,870	4.7x	NN-2540
14	IM_Bol_14	Cabbage	NextSeq 500	29,666,958	4,462,572,566	6.9x	NN-2541
15	IM_Bol_15	Cabbage	NextSeq 500	27,212,296	4,093,204,740	6.3x	NN-2542

16	IM_Bol_16	Cabbage	NextSeq 500	25,875,536	3,891,772,517	6.0x	NN-2543
17	IM_Bol_17	Cabbage	NextSeq 500	24,091,034	3,613,319,229	5.6x	NN-2544
18	IM_Bol_18	Cabbage	NextSeq 500	23,080,080	3,470,177,537	5.4x	NN-2545
19	IM_Bol_19	Cabbage (Red)	NextSeq 500	22,034,600	3,314,510,653	5.1x	NN-2546
20	IM_Bol_20	Cabbage (Red)	NextSeq 500	20,132,954	3,028,742,165	4.7x	NN-2547
21	IM_Bol_21	Broccoli	HiSeq 2000	41,360,398	4,177,400,198	6.5x	NN-2529
22	IM_Bol_22	Broccoli	NextSeq 500	29,216,120	4,394,748,426	6.8x	NN-2548
23	IM_Bol_23	Broccoli	NextSeq 500	30,320,644	4,558,502,571	7.1x	NN-2549
24	IM_Bol_24	Broccoli	NextSeq 500	31,965,382	4,803,230,335	7.4x	NN-2550
25	IM_Bol_25	Broccoli	NextSeq 500	27,781,312	4,173,439,756	6.5x	NN-2551
26	IM_Bol_26	Cauliflower	HiSeq 2000	42,785,216	4,321,306,816	6.7x	NN-2553
27	IM_Bol_27	Cauliflower	NextSeq 500	22,939,574	3,447,289,984	5.3x	NN-2552
28	IM_Bol_28	Cauliflower	NextSeq 500	20,744,430	3,118,796,486	4.8x	NN-2553
29	IM_Bol_29	Cauliflower	NextSeq 500	21,752,896	3,270,437,504	5.1x	NN-2554
30	IM_Bol_30	Cauliflower	NextSeq 500	17,590,712	2,643,949,139	4.1x	NN-2555
31	IM_Bol_31	Kale	HiSeq 2000	37,437,444	3,781,181,844	5.9x	NN-2530
32	IM_Bol_32	Kale	NextSeq 500	29,375,128	4,417,524,971	6.8x	NN-2556
33	IM_Bol_33	Kale	NextSeq 500	28,857,440	4,333,159,487	6.7x	NN-2557
34	IM_Bol_34	Kale	NextSeq 500	26,945,414	4,049,362,678	6.3x	NN-2558

35	IM_Bol_35	Kale	NextSeq 500	32,888,608	4,946,458,550	7.7x	NN-2559
36	IM_Bol_36	Kohlrabi	HiSeq 2000	47,284,696	4,775,754,296	7.4x	NN-2535
37	IM_Bol_37	Kohlrabi	NextSeq 500	25,928,648	3,900,488,174	6.0x	NN-2560
38	IM_Bol_38	Kohlrabi	NextSeq 500	19,946,018	3,000,396,162	4.7x	NN-2561
39	IM_Bol_39	Kohlrabi (Purple)	NextSeq 500	32,485,848	4,886,191,711	7.6x	NN-2563
40	IM_Bol_40	Kohlrabi (Purple)	NextSeq 500	20,439,910	3,074,727,253	4.8x	NN-2564
41	IM_Bol_41	Kailan	HiSeq 2000	41,848,780	4,226,726,780	6.6x	NN-2536
42	IM_Bol_42	Kailan	NextSeq 500	25,646,552	3,853,116,381	6.0x	NN-2565
43	IM_Bol_43	Kailan	NextSeq 500	26,210,296	3,941,431,623	6.1x	NN-2566
44	IM_Bol_44	Brussels sprouts	HiSeq 2000	43,434,280	4,386,862,280	6.8x	NN-2534
				Average	3,998,159,703	6.2x	

2. SNP filtering

Among the 13 million SNPs, 3 million SNPs that were genotyped from more than 40 accessions were extracted. SNPs with more than 5 heterozygous genotypes and SNPs with less than 5 minor allele frequency were removed. SNPs that have other SNPs within 150bp on both sides of the flanking sequences were also removed to enhance the accuracy of markers (Lee et al. 2015, Lee et al. 2016). Among the resulting 125,677 SNPs, 3,446 SNPs that are located in genic regions were selected. 150bp of the flanking regions from both sides of the SNP, were extracted from an online source (*http://im-crop.snu.ac.kr/new/tools/candi_ole_subseq.pl*). Blast analysis was utilized to remove paralogous sequences. 849 SNPs from single copy genes were finally selected.

3. Gene function analysis

Prior to marker designing, we used Blast2GO to assign Gene Ontology (GO) terms to investigate the functions of the single copy genes that contain the filtered SNPs (Figure 10). Through GO analysis, these genes were assigned to three different categories: molecular function, cellular component, and biological processes.

4. Removing paralogous sequences

B. oleracea has undergone whole genome duplications and whole genome triplications. Although *B. oleracea* is a diploid plant, it acquired hexaploid traits

during evolution (Lie et al. 2014) (Figure 2). This results in paralogous sequences of target sequences in the same chromosome and other chromosomes as well. Due to the limit of the hybridization method, primers can bind to paralogous sequences that have very few differences (Figure 3). Thus allele sites with the 150 bp flanking sequences from both sides that are on 3,446 SNPs in genes were extracted form database produced from previous research (Lee et al. 2014, Lee et al. 2015). The resulting 301 bp sequences were aligned to the reference genome using Blast analysis to eliminate SNPs that are present in paralogous sequences. Conclusively, we extracted 849 SNPs derived from single copy genes.

5. Marker design

When marker designing, one should make sure the markers are distributed evenly throughout the chromosomes to represent the whole genome. To obtain this goal, each chromosome was divided into smaller regions. Bigger chromosomes such as chromosomes 1, 3, and 9 were divided into 6 parts and the rest into 5 parts (Table 3). In each region, candidate SNPs were randomly selected using python code. If the same SNP got selected, it was replaced manually. SNPs with flanking sequences were subjected to test the possibility of marker production through Fluidigm website (https://www.fluidigm.com/) to check the validity of markers. Markers that are designable are classified into High and Medium based on the GC content. Markers that are classified into the high category were selected. Markers that are classified into Medium were replaced to High category.

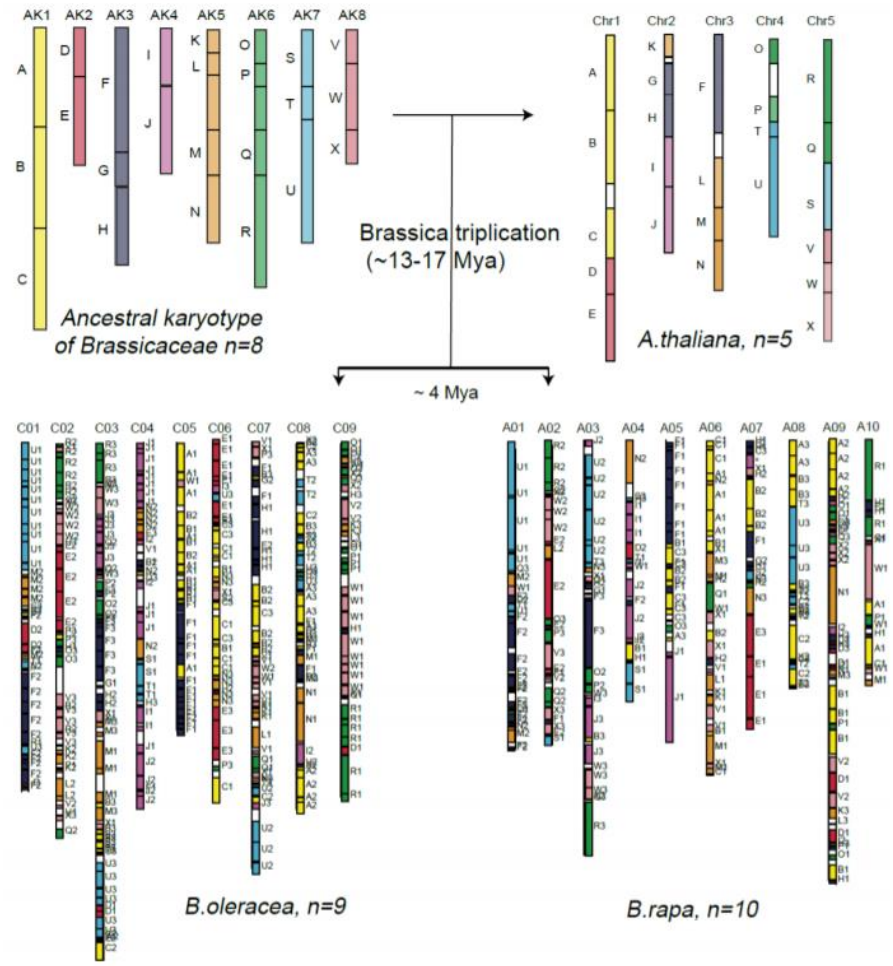


Figure 2. Evolutionary scenario of *Arabidopsis thaliana, Brassica rapa,* and *B. oleracea*

The 24 ancestral blocks (labelled A-X) were defined and colored in ancestor of *Brassicaceae* and in *A. thaliana* as described previously. The distribution of ancestral block in *B. oleracea* and *B. rapa* are obtained by alignment analysis between *A. thaliana* and *B. rapa, A. thaliana* and *B. oleracea* (Liu et al, 2014).

.

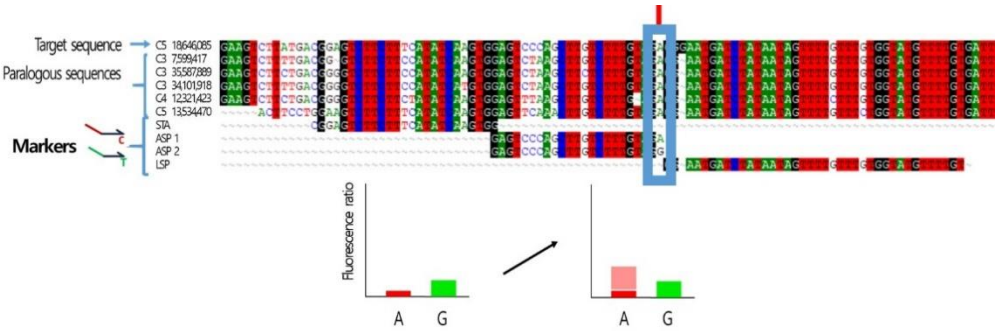


Figure 3. Change of genotyping due to the influence of paralogous sequences.
Paralogous sequences that make proper genotyping difficult, exists not only in the same chromosome but also in other chromosomes. The genotype is determined with the fluorescence ratio between ASP 1 and ASP 2. In this figure, ASP 1 detects A and ASP 2 detects G. However, due to paralogous sequences, samples that should be confirmed as G can be proven to be a heterozygous or homozygous A type since there are lots of paralogous sequences that ASP 2 can bind to.

Table 2. Number of SNPs and Indels depending on the number of genotyped accessions

No. of genotyped progenies	Type	C1	C2	C3	C4	C5	C6	C7	C8	C9	Total
≥ 1	SNPs	1,235,681	1,614,089	1,857,258	1,538,013	1,278,069	1,127,637	1,368,540	1,096,442	1,565,640	12,681,369
	InDels	144,889	181,495	238,362	180,450	149,789	133,469	163,057	138,108	184,345	1,513,964
≥ 10	SNPs	1,156,476	1,501,049	1,727,331	1,425,330	1,192,043	1,064,576	1,286,382	1,029,849	1,458,887	11,841,923
	InDels	130,528	162,108	214,382	161,948	134,984	121,120	146,692	124,898	166,145	1,362,805
≥ 20	SNPs	1,041,341	1,337,656	1,534,284	1,285,986	1,059,155	967,438	1,161,774	933,056	1,298,979	10,619,669
	InDels	116,675	144,079	190,251	145,437	119,917	109,845	132,145	112,434	147,865	1,218,648
≥ 30	SNPs	826,386	1,035,680	1,175,383	1,020,130	834,701	764,939	927,649	741,311	1,008,291	8,334,470
	InDels	89,957	108,017	142,433	112,107	91,197	84,571	102,556	85,748	111,897	928,483
≥ 40	SNPs	330,572	410,004	443,238	417,810	347,780	309,761	378,957	309,992	404,452	3,352,566
	InDels	33,469	39,248	49,062	42,305	34,261	31,638	39,089	32,279	41,161	342,512
44	SNPs	47,572	61,589	62,655	61,916	54,121	45,626	54,190	48,103	60,691	496,463
	InDels	3,569	4,337	5,177	4,694	3,832	3,411	4,286	3,598	4,589	37,493

Table 3. Candidate SNPs and number of markers per chromosomes

Chromosomes	Genetic distance	SNPs (40/44 common)	SNP in genes	SNP in single copy genes	Candidate probes	Validated Probes	Selected Probes	Finally selected Probes	Marker density(cM)
1 chr	**98.9**	**330,572**	**504**	**122**	**32**	**29**	**23**	**19**	**5.2**
2 chr	**91.9**	**410,004**	**330**	**69**	**33**	**23**	**17**	**15**	**6.1**
3 chr	**134.2**	**443,238**	**579**	**140**	**41**	**35**	**26**	**24**	**5.6**
4 chr	**103.9**	**417,810**	**403**	**119**	**37**	**29**	**24**	**22**	**4.7**
5 chr	**93.9**	**347,780**	**372**	**92**	**31**	**28**	**23**	**22**	**4.3**
6 chr	**72.5**	**309,761**	**244**	**59**	**40**	**30**	**21**	**19**	**3.8**
7 chr	**62.6**	**378,957**	**160**	**43**	**28**	**23**	**13**	**8**	**7.8**
8 chr	**64.2**	**309,992**	**321**	**87**	**25**	**27**	**18**	**5**	**12.8**
9 chr	**110.6**	**404,452**	**533**	**118**	**38**	**33**	**27**	**16**	**6.9**
Total	**833**	**3,352,566**	**3,446**	**849**	**304**	**257**	**192**	**150**	**6.4**

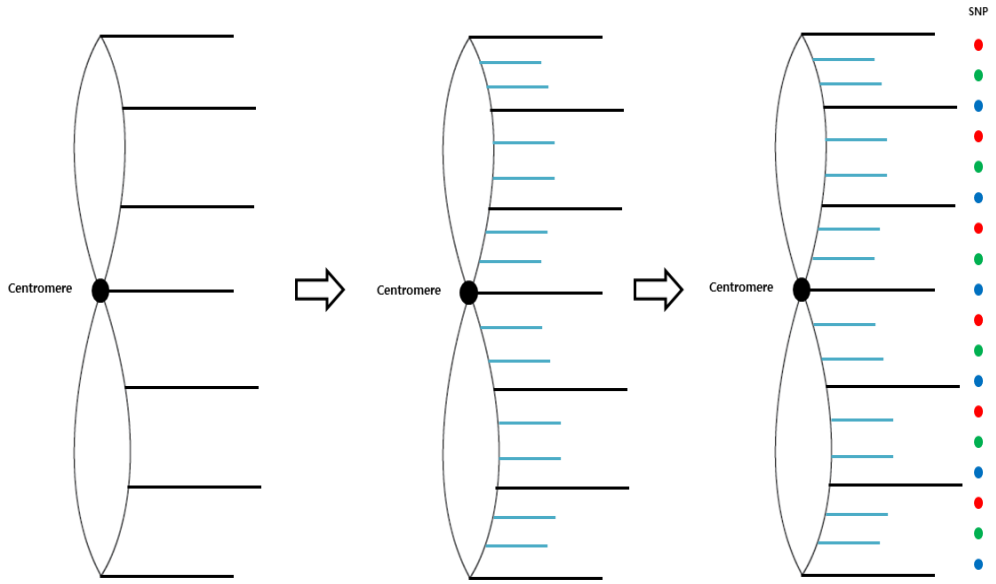


Figure 4. Principle of marker designing location.

First, each chromosome was divided into 5~6 parts(black line) based on the genetic map information and each smaller part was split again into 3 smaller regions (blue line) based on the physical map information.

6. Fluidigm experiments

Fluidigm experiment is composed of 5 steps: STA, priming, loading, PCR and scanning. In plants, STA (Specific Target Amplification) is a necessary step to remove polysaccharides, which animals do not have, and to reduce size because capillaries of IFC (Integrated Fluidic Circuits) are narrow. STA is done by going through a normal PCR process with 14, while Target Specification and Target-Specific Priming are done in the SNP array. STA products are diluted to 1/100 with water for higher DNA purity. Priming is the process of injecting oil to clean the capillaries and loading is the process of injecting samples and assays into SNP array. Another PCR is done inside the chip, and it is composed of Target Specification and Target-specific Priming (Figure 5). In Target Specification, ASP (Allele-Specific Primer) and LSP (Locus-Specific Primer) are used for a second PCR. Fluorescent tags attach to specific sites located on the tail side of ASP. Finally the ratio of fluorescence is measured with Fluidigm EP1 reader (Fluidigm USA).

.

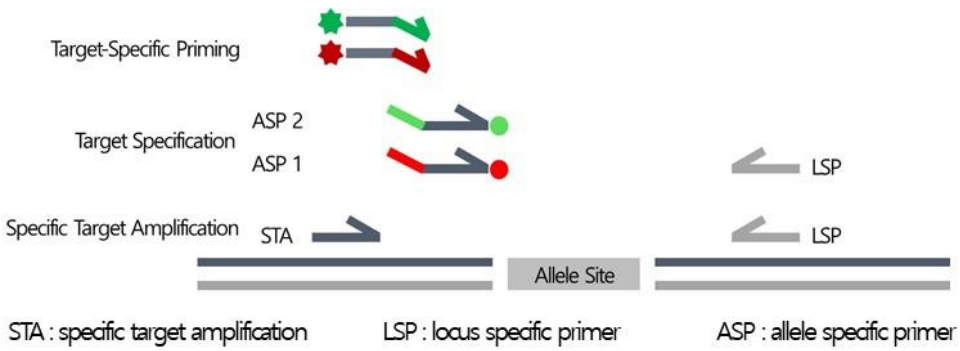


Figure 5. Principle of chip experiment.

ASP has a site for a fluorescent tag to bind to, and the quencher is detached while the polymerase elongates the product. ASP 1 and ASP 2 distinguishes the allele. If one ASP is attached, it will give a red or green fluorescence meaning homozygous, while both ASP types attaching indicates heterozygosity through a blue fluorescence

7. Markers screening

After applying 480 markers to 43 *B. oleracea* accessions, 2 *Brassica napus*, 1 *Raphanus rapanistrum* and 2 No Template Control (NTC), we screened markers based on the accordance rate between chip data and resequencing data and data clustering form. If both genotypes do not match, it was treated as 2 mismatches. In case of 1 genotype matches, It was treated as 1 mismatch. 203 markers with more than 80% accordance rate were selected. If the markers shared similar genetic positions, only one of them was selected. 192 markers that represent the whole genome were chosen.

8. Markers selection

192 markers were tested on 92 unknown samples. These markers were scored from 1 to 3 grade with no call rate. With high no call rate, recall rate of experiments decreases. 1st grade was assigned to markers with less than 5 no calls, 2nd grade for markers with 6 to 10 no calls, and 3rd grade for markers with more than 11 no calls. To increase marker accuracy and efficiency, markers were graded again with genotype frequency. In each grade, ‘skewed’ grade was assigned to markers that have major allele frequency over 95%, ‘no use’ to markers that showed no call rates over 10% and ‘evenly distributed’ to the rest. 134 evenly distributed 1st grade and 16 evenly distributed 2^{ns} markers, total 150 markers, were finally selected.

9. Phylogenetic tree

A phylogenetic tree was drawn using 150 markers using the genotyping result. Heterozygous results were marked as H/H and No calls were marked as -/-.

Homozygous results were treated as no change. Phylogenetic analysis was performed using neighbor-joining method of powermarker (https://brcwebportal.cos.ncsu.edu/powermarker/) with 1000 bootstrap replicates.

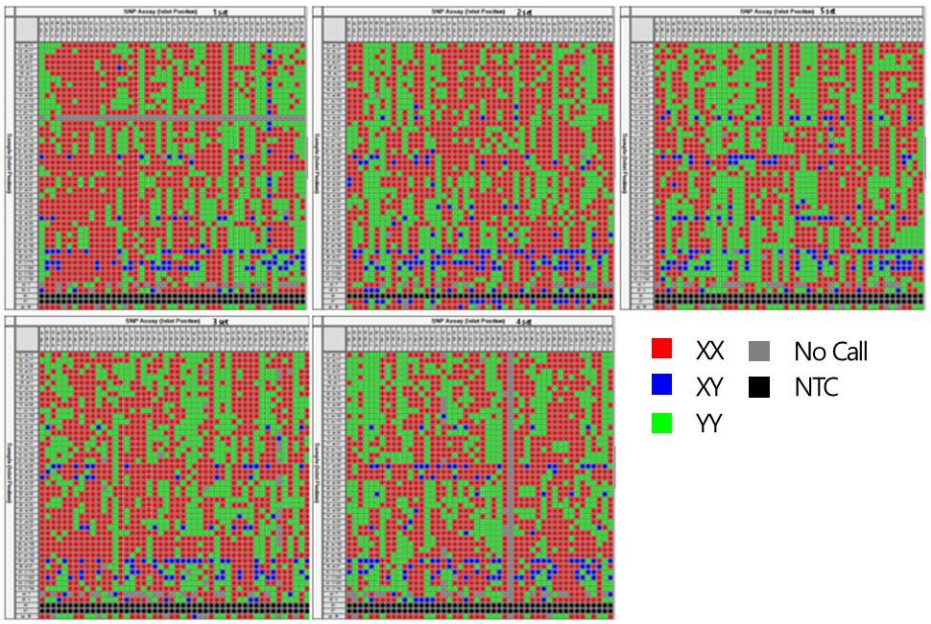


Figure 6. Result of 240 SNP markers tested on 43 *B. oleracea* accessions, 2 *Brassica napus*, 1 *Raphanus rapanistrum* and 2 No Template Control (NTC).

Red and green data points mean homozygous while blue data points mean heterozygous. Black data points are NTC which is used as standard.

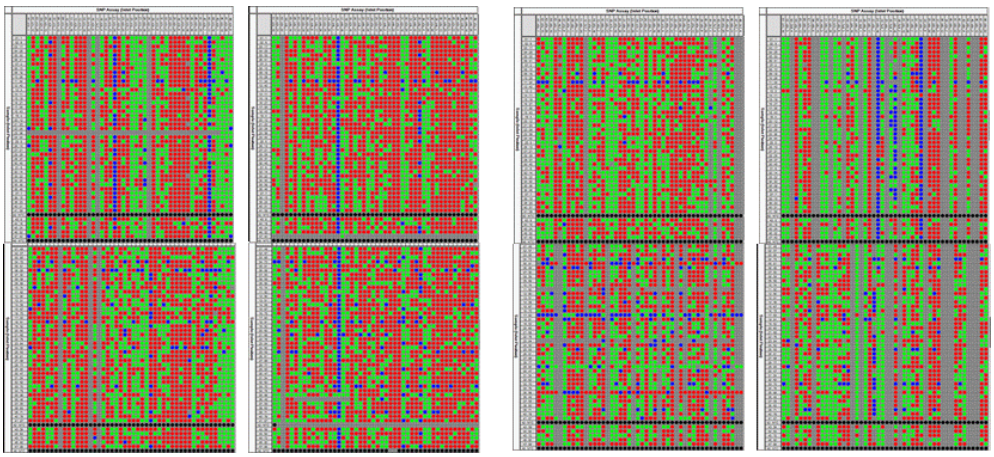


Figure 7. Chip results of 192 markers applied to 92 samples from Farmhannong.

Results

1. 849 SNPs derived from single copy genes

From the previous research (Lee et al. 2014, Lee et al. 2015), 3 million SNPs that more than 40 out of 44 accessions commonly contain, were extracted. According to the order of the workflow (Figure 1), SNP calling was performed. Within these 125,677 SNPs, 3,446 SNPs in genic regions were selected. With 150bp flanking sequences from both sides of the SNP, Blast search was carried out to remove the paralogous sequences that showed similarity with the sequences containing the candidate SNPs. 849 SNPs derived from single copy genes were considered when designing the markers.

2. 480 markers design

240 markers were designed from 3,446 SNPs and 849 SNPs each, 480 markers in total. The initially designed 240 markers from 3,446 SNPs showed a low success rate because of paralogous sequences. However, 240 markers designed from single copy genes produced success rates of over 80%. Experimental results testing these 240 markers revealed that 203 markers showed high accordance rate with the resequencing data. In total, 308 markers, 68 markers initially designed from 3,446 SNPs and 240 markers derived from 849 SNPs, were designed from single copy genes. After comparing the chip data and resequencing data, 257 markers were evaluated as reliable (Table 3). Considering accordance rate and genetic position, 203 markers were rated as high quality and 192 markers were finally selected to be applied to unknown samples based on the genetic position. Average marker density was 5.6 cM.

Moreover, Jeong (2015) and Herzog and Frisch (2011) reported that an average marker density of 6.52 cM and 10 cM is high enough for MAB, respectively. 5.6 cM marker density from this study is expected to be enough for the objective.

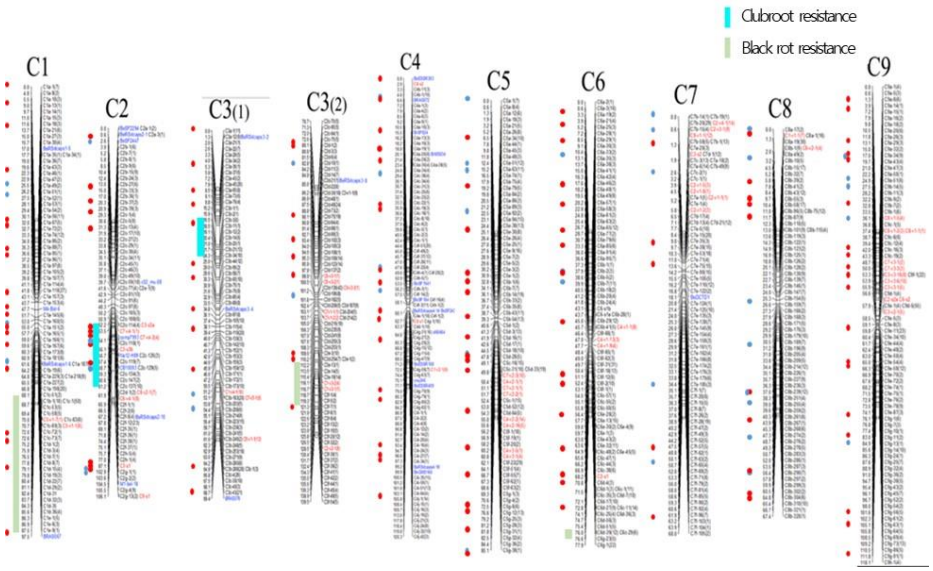


Figure 8. Position of markers.

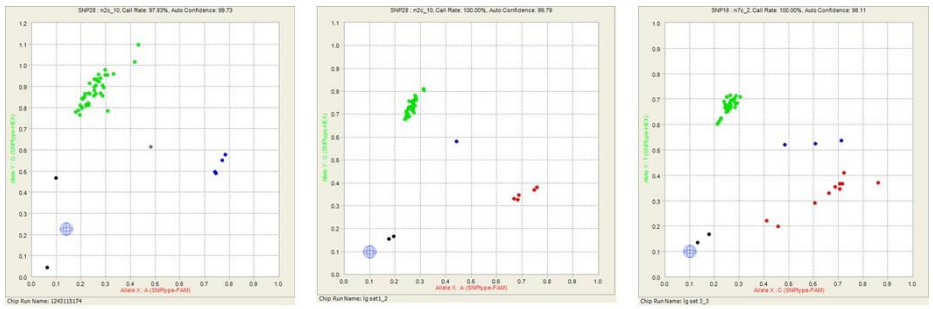
203 markers are marked on the genetic map of *B. oleracea* based on the genetic position. Blue sections represent clubroot resistance QTL and green section represents black rot resistance QTL that are studied in the previous research (Lee et al. 2014, Lee et al. 2015).

3. Validation of SNP markers

After applying 192 markers to 92 unknown samples, 150 markers that were classified into the Evenly Distributed category marked 1st and 2nd grade, were finally selected (Table 4). Sixteen 2nd grade markers that are classified into the Evenly Distributed category are expected to successfully genotype unknown samples.. In order to make sure that 1st grade markers can be applied right away, markers that showed even a little flaw were rated 2nd.13 out of the 16 markers rated 2nd could perform similarly to 1st grade markers after minor corrections (Figure 9). 134 1st grade markers and sixteen 2nd grade markers that were corrected were used to draw a phylogenetic tree (Figure 11). In this figure, both trees showed successful categorization of cabbage in group 1 and other *B. olecracea* subspecies in group 2.

Table 4. Screening result based on grade and genotype frequency.

Grade	**Frequency**	**Number**
1st grade	Skewed	7
	Evenly Distributed	134
	No Use	0
2nd grade	Skewed	0
	Evenly Distributed	16(13)
	No Use	8
3rd grade	Skewed	0
	Evenly Distributed	3
	No Use	24
	Total	192



(a) (b) (c)

Figure 9. Example of 2nd grade markers.

(a) genotyping result with auto call. Without correction, blue data points lie where red data points should be displayed. (b) genotyping result after manually changing the blue data points to red data points. (c) genotyping with dispersed data points. Although these markers are capable of accurately genotyping unknown samples, they are ranked 2nd due to the high quality standard set for 1st grade markers.

Discussion

1. 849 SNPs derived from single copy genes

From 120,000 SNPs, 3,446 SNPs in genic regions were extracted. In step 2 of the workflow (Figure 1), we extracted 150bp from both sides of the SNP, and paralogous sequences were filtered out by Blast analysis. In Fluidigm web site (https://www.fluidigm.com/), marker validity was confirmed. Candidate sequences classified into Not Designable and Medium were replaced to sequences classified into the High category. In step 2 and 3 (Figure 1), marker value and accuracy were improved. From 3,446 SNPs in genic regions, 849 SNPs derived from single copy gene regions were selected for marker design. The genes that contain these SNPs were analyzed with Gene Ontology to identify their functions. To get more detailed information on the functions of these genes, we searched the Brassica database (http://ocri-genomics.org/bolbase/). For unsearchable genes in the Brassica database, we searched for functions of orthologous genes in *A. thaliana*. Some of the functions of genes that 150 markers were designed from are related to flower development, disease resistance, and ethylene receptor that are important in agriculture. However, not all 150 genes were annotated. But they have possibility to change phenotype even through one nucleotide change. If phenotype investigation is done together, function of these genes have possibility to be tracked.

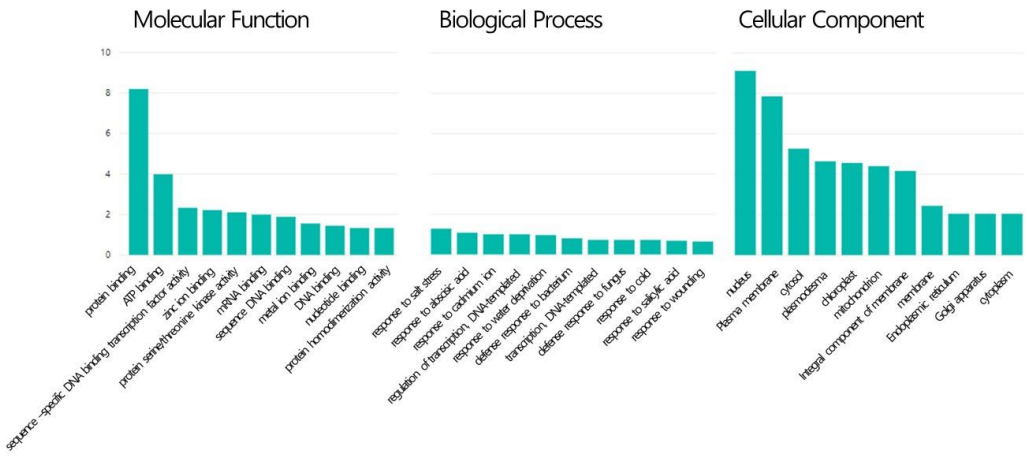


Figure 10. GO annotation result of 849 single copy genes.

Top 11 GO terms in each classification. 829 SNPs were assigned to 4,673 GO terms. 902 terms were assigned to Molecular Function and 1,272 to Biological Process, and 2,498 to Cellular Component.

2. Phylogenetic analysis

Brassica genus consists of many important agricultural crops such as broccoli, cauliflower, kalian, kohlrabi, kale etc. In this study, we constructed phylogenetic trees based on the SNP chip result of 44 accessions that our lab possess, and 92 samples from Farmhannong with powermarker (https://brcwebportal.cos.ncsu.edu/powermarker/). Samples were divided into 2 groups, group 1 with cabbage, and group 2 with *B. oleracea* subspecies. In group 2, broccoli, cauliflower, kalian, kohlrabi, kale were grouped well (Figure 11, (a)). These trees clearly showed the evolutionary relationship of *B. oleracea* subspecies. If there are SNPs that could separate group 1 and group 2, and each of the *B. oleracea* subspecies in group 2, these SNPs are expected to further enhance the evolutionary study of *B. oleracea*.

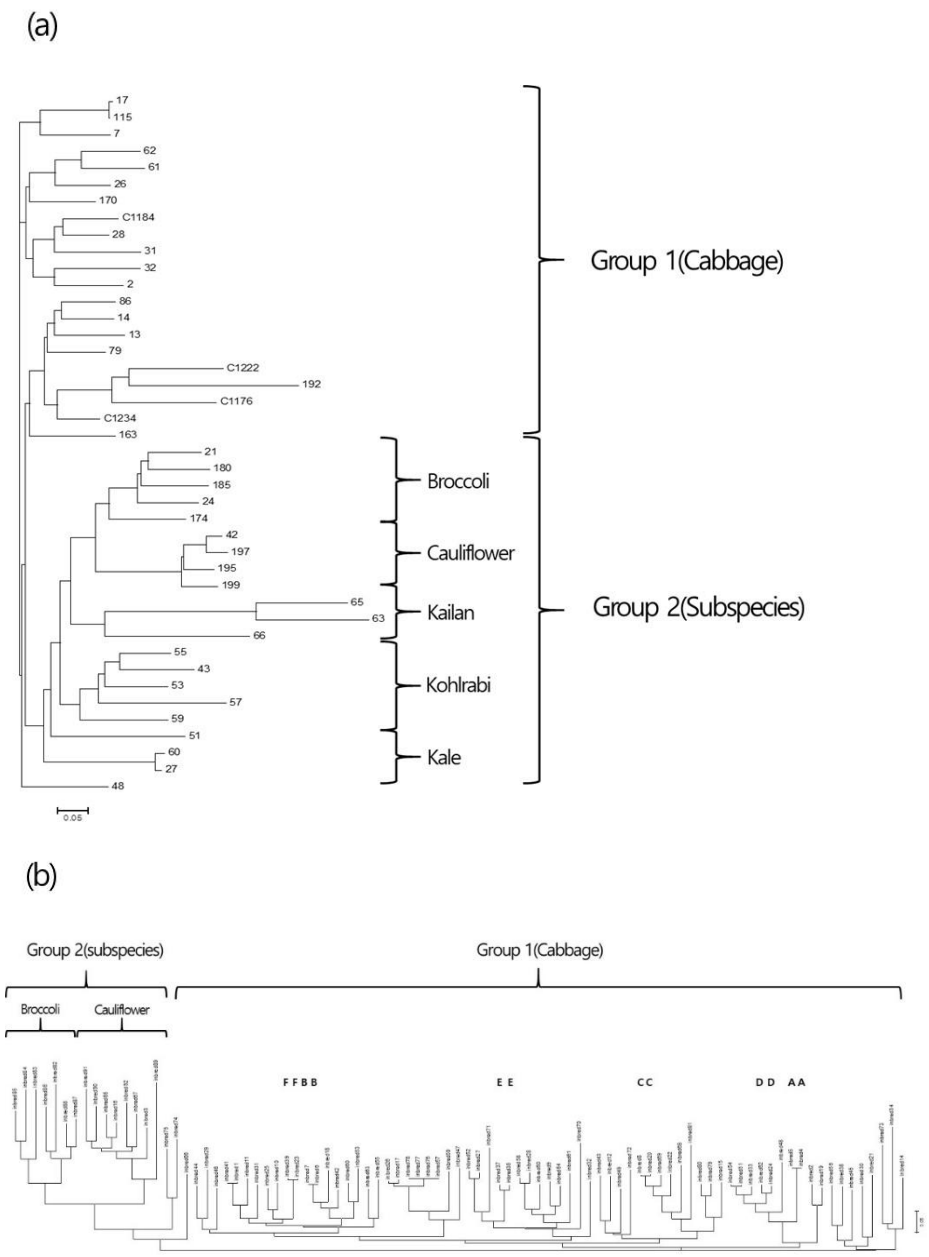


Figure 11. Phylogenetic tree using 150 markers.

(a) Phylogenetic tree drawn with 44 accessions. Group 1 which is composed of cabbage and group 2 which is composed of *B. oleracea* subspecies were clustered well. (b) Phylogenetic tree drawn with 92 accessions.Letters from A to F were assigned by Farmhannong on accessions based on the phenotype without considering genetic traits.

3. 150 markers derived from single copy genes

We validated 150 markers two times with phylogenetic tree analysis. Group 1 with cabbage and group 2 with *B. oleracea* subspecies were clustered. Moreover, letters A~F were assigned based on the phenotype without considering genetic traits. Samples with the same letters were grouped well (Figure 11 (b)). It appears that markers were applied well not only to samples that are not genotyped, but also in genotyped samples (Figure 11). Although marker density on chromosome 8 is a bit higher than 10 cM, the marker density of other chromosomes are low enough to complete a high quality marker set (Table 3). An average marker density of 6.4 cM is expected to be enough for MAB (Jeong et al 2015). If supplementary markers are designed on chromosome 8 this marker set can be improved. This marker set can be applied right away to shorten the breeding period and genotype unknown samples.

In conclusion, 150 markers derived from single copy genes were developed from 44 *B. oleracea* resequencing data. The 849 single copy genes that can possibly be linked to phenotypes can be useful resources for developing useful *B. oleracea* varieties. These markers are expected to instantly be utilized for MAB. Moreover, these markers can be used for phylogenetic relationship analysis in cabbage and related *B. oleracea* subspecies

References

Allen, G. C., Flores-Vergara, M. A., Krasnyanski, S., Kumar, S., & Thompson, W. F. (2006). A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nature Protocols, 1*(5), 2320-2325. doi:10.1038/nprot.2006.384

Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. (2000). *Nature, 408*(6814), 796-815. doi:10.1038/35048692

Collard, B. C. Y., & Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B-Biological Sciences, 363*(1491), 557-572. doi:10.1098/rstb.2007.2170

Doyle, J. J., & Egan, A. N. (2010). Dating the origins of polyploidy events. *New Phytol, 186*(1), 73-85. doi:10.1111/j.1469-8137.2009.03118.x

Duran, C., Appleby, N., Edwards, D., & Batley, J. (2009). Molecular Genetic Markers: Discovery, Applications, Data Storage and Visualisation. *Current Bioinformatics, 4*(1), 16-27. doi:10.2174/157489309787158198

Herzog, E., & Frisch, M. (2011). Selection strategies for marker-assisted backcrossing with high-throughput marker systems. *Theoretical and Applied Genetics, 123*(2), 251-260. doi:10.1007/s00122-011-1581-0

Hospital, F., Chevalet, C., & Mulsant, P. (1992). Using markers in gene introgression breeding programs. *Genetics, 132*(4), 1199-1210.

Jeong, H. S., Jang, S., Han, K., Kwon, J. K., & Kang, B. C. (2015). Marker-assisted backcross breeding for development of pepper varieties (Capsicum annuum) containing capsinoids. *Molecular Breeding, 35*(12). doi:10.1007/s11032-015-0417-z

Jiao, Y. N., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., . . . dePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature, 473*(7345), 97-U113. doi:10.1038/nature09916

Jin-Ho Kang, Hee-Bum Yang, Hyeon-Seok Jeong, Phillip Choe, Jin-Kyung Kwon, Byoung-Cheorl Kang. (2014). Single Nucleotide Polymorphism Marker Discovery from Transcriptome Sequencing for Marker-assisted Backcrossing in Capsicum. Korean Journal of Horticultural Science & Technology, 32(4), 535-543. Kifuji, Y., Hanzawa, H., Terasawa, Y., Ashutosh, & Nishio, T. (2013). QTL analysis of black rot resistance in cabbage using newly developed EST-SNP markers. *Euphytica, 190*(2), 289-295. doi:10.1007/s10681-012-0847-1

Lee, J., Izzah, N. K., Choi, B. S., Joh, H. J., Lee, S. C., Perumal, S., . . . Yang, T. J. (2016). Genotyping-by-sequencing map permits identification of clubroot resistance QTLs and revision of the reference genome assembly in cabbage (Brassica oleracea L.). *DNA Research, 23*(1), 29-41. doi:10.1093/dnares/dsv034

Lee, J., Izzah, N. K., Jayakodi, M., Perumal, S., Joh, H. J., Lee, H. J., . . . Yang, T. J. (2015). Genome-wide SNP identification and QTL mapping for black rot resistance in cabbage. *Bmc Plant Biology, 15*. doi:10.1186/s12870-015-0424-6

Liu, S. Y., Liu, Y. M., Yang, X. H., Tong, C. B., Edwards, D., Parkin, I. A. P., . . . Paterson, A. H. (2014). The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications, 5*. doi:10.1038/ncomms4930

Mammadov, J., Aggarwal, R., Buyyarapu, R., & Kumpatla, S. (2012). SNP markers and their impact on plant breeding. *Int J Plant Genomics, 2012*, 728398. doi:10.1155/2012/728398

Parkin, I. A. P., Koh, C., Tang, H. B., Robinson, S. J., Kagale, S., Clarke, W. E., . . . Sharpe, A. G. (2014). Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biology, 15*(6). doi:10.1186/gb-2014-15-6-r77

Stoskopf, N. C., & Tomes, D. T. (1993). *Plant breeding: theory and practice* (No. 633 S86).

Town, C. D., Cheung, F., Maiti, R., Crabtree, J., Haas, B. J., Wortman, J. R., . . . Bancroft, I. (2006). Comparative genomics of Brassica oleracea and Arabidopsis thaliana reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell, 18*(6), 1348-1359. doi:10.1105/tpc.106.041665

Wang, W. X., Huang, S. M., Liu, Y. M., Fang, Z. Y., Yang, L. M., Hua, W., . . . Zeng, A. S. (2012). Construction and analysis of a high-density genetic linkage map in cabbage (Brassica oleracea L. var. capitata). *Bmc Genomics, 13*, 12. doi:10.1186/1471-2164-13-523

ABSTRACT IN KOREAN

양배추는 세계 10 대 작물 중 하나로, 매년 2 백만 헥타르 이상의 지역에서 재배되고 있으며, 96 억 달러의 세계 채소 종자 시장에서 1/3 정도를 차지하고 있는 작물이다. 2012 년에 농림수산식품기획평가원에서 나온 보고서에 따르면, 한국은 양배추 전통 육종 분야에서는 세계적인 수준의 기술력을 가지고 있는 것으로 평가된다. 이에 육종 기술을 더 발전시킨다면 한국은 양배추 육종 분야에서 세계적인 경쟁력을 갖출 수 있을 것으로 보인다. 이 연구에서는 농업적으로 다양한 형질을 가졌다고 평가되는 양배추 44 계통의 resequencing data 를 평균 6.2x 의 coverage 로 생산하였고, reference genome 에 mapping 하여 SNP 를 calling 하였다. 44 개 계통 중 40 개 이상의 계통에서 genotyping 이 되는 300 만개 이상의 SNP 들 중, 마커의 정확도와 신뢰도를 높이기 위한 필터링 과정을 거쳐 12 만개 정도의 SNP 를 선발하였다. Gene 지역에 있는 3,446 개의 SNP 를 선발하였으며, 전체 유전체 중, 유일하게 존재하는 single copy gene 849 개를 최종적으로 선발하였다. 849 개 candidate SNP 들은 유사한 서열을 가진 sequence 들이 FLuidigm 의 형광 신호 해석을 방해하는 것을 방지하며, gene 들의 기능을 밝혀낼 시, 즉시 phenotype 으로 연결될 수 있다는 점에서 가치가 높을 것으로 보인다. 3,446 개의 gene 영역과 849 개의 single copy gene 영역에서 각각 240 개씩 유전체 전반에 걸쳐

분포하도록 마커를 디자인 하였다. 실험실에서 보요하고 있는 44 개 계통에 대하여 마커 실험을 진행한 뒤, resequencing data 와 높은 일치율을 보이는 192 개의 마커를 선발하였다. 마커를 한번 더 validation 하기 위해 팜한농에서 제공한 92 개 샘플에 대하여 다시 한번 적용하였고, major allele 가 95% 이하인 마커 150 개를 선발하였다. 이들 마커로 실험실에서 보유한 44 개 계통과 팜한농에서 제공한 92 개 계통에 대하여 phylogenetic tree 를 그려보았고, 양배추로 구성된 group 1 과 아종들로 구성된 group 2 로 잘 분리되는 것을 확인하였고, 팜한농에서 표현형으로 분류한 표식들과도 일치하는 것을 확인하였다. group 2 에서는 아종들끼리도 잘 분류되는 것으로 보아 그 활용도가 높을 것으로 보인다. 본 연구에서는 양배추 여교배 세대촉진을 위한 150 개의 Fluidigm SNP marker 를 gene 지역에서 개발하였고, 이들 마커가 잘 적용됨을 확인하였다. 추후 여교배 육종에 즉시 활용가능할 것으로 보이며, gene 연구에도 도움을 줄 수 있을것으로 보인다.