Single Nucleotide Polymorphism Marker Discovery from Transcriptome Sequencing for Marker-assisted Backcrossing in *Capsicum*

Jin-Ho Kang$^{1,2,3}$$^\dagger$, Hee-Bum Yang$^{1,2,3}$$^\dagger$, Hyeon-Seok Jeong$^{1,2,3}$, Phillip Choe$^{1,2,3}$, Jin-Kyung Kwon$^{1,2,3}$, and Byoung-Cheorl Kang$^{1,2,3}*$$^\dagger$

$^1$Department of Plant Science, Seoul National University, Seoul 151-921, Korea
$^2$Vegetable Breeding Research Center, Seoul National University, Seoul 151-921, Korea
$^3$Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Korea

Abstract: Backcross breeding is the method most commonly used to introgress new traits into elite lines. Conventional backcross breeding requires at least 4-5 generations to recover the genomic background of the recurrent parent. Marker-assisted backcrossing (MABC) represents a new breeding approach that can substantially reduce breeding time and cost. For successful MABC, highly polymorphic markers with known positions in each chromosome are essential. Single nucleotide polymorphism (SNP) markers have many advantages over other marker systems for MABC due to their high abundance and amenability to genotyping automation. To facilitate MABC in hot pepper (*Capsicum annuum*), we utilized expressed sequence tags (ESTs) to develop SNP markers in this study. For SNP identification, we used Bukang F$_1$-hybrid pepper ESTs to prepare a reference sequence through de novo assembly. We performed large-scale transcriptome sequencing of eight accessions using the Illumina Genome Analyzer (IGA) IIx platform by Solexa, which generated small sequence fragments of about 90-100 bp. By aligning each contig to the reference sequence, 58,151 SNPs were identified. After filtering for polymorphism, segregation ratio, and lack of proximity to other SNPs or exon/intron boundaries, a total of 1,910 putative SNPs were chosen and positioned to a pepper linkage map. We further selected 412 SNPs evenly distributed on each chromosome and primers were designed for high throughput SNP assays and tested using a genetic diversity panel of 27 *Capsicum* accessions. The SNP markers clearly distinguished each accession. These results suggest that the SNP marker set developed in this study will be valuable for MABC, genetic mapping, and comparative genome analysis.

Additional key words: expressed sequence tag, linkage map, marker-assisted breeding

Introduction

Hot pepper (*Capsicum annuum*) is a crop of major economic importance that is commercially cultivated in China, Korea, the East Indies, and the United States of America, among many other countries (Shao et al., 2008). Worldwide production of hot peppers has been estimated to be 14-15 million tons a year (Weiss, 2002). In fact, hot pepper is the vegetable accounting for the largest planting area in Korea.

In commercial breeding, new traits are commonly introduced into elite breeding lines using conventional backcross methods, which involve time consuming efforts to transfer target genes into the genetic background of a recipient parent. Recently marker-assisted backcrossing (MABC), which is more efficient and faster than conventional backcrossing, has been generally accepted as an advanced plant breeding technique (Blum et al., 2002). For successful application...
of MABC, factors to consider include the number of target genes to be transferred, the genome size of the crop species, and the availability of a highly saturated map. Among those factors, the availability of highly polymorphic markers with known positions in each chromosome is critical (Varshney et al., 2009). Although molecular markers based on restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), and random amplified polymorphic DNA (RAPD) have been developed and used in practical plant breeding (Imelfort et al., 2009; Jung et al., 2010; Kang et al., 2001; Paran et al., 2004; Yoo et al., 2003), such markers still have limitations for use in MABC due to the difficulty of finding polymorphisms among breeding lines and of high throughput genotyping. Recently, single nucleotide polymorphism (SNP) markers have captured attention because of their potential for high-throughput detection and computerization with automated platforms (Jung et al., 2010; Vignal et al., 2002; Yi et al., 2006).

SNPs are single-base differences in DNA between accessions. In plants, SNPs generally occur in populations once every few hundred base pairs (Metzker, 2005). SNP markers can be developed through several methods. For example, SNPs can be identified by simply comparing a candidate sequence to a reference sequence (Nicolai et al., 2012), by whole genome sequencing (WGS; Goff et al., 2002; The Arabidopsis Genome Initiative, 2000), or by sequence alignment of expressed sequence tags (ESTs) to a reference sequence (Jones, 2009; Kota et al., 2001; Labate and Baldo, 2005). For the crops like hot pepper, in which the genome is huge, ESTs have been adopted as an alternative to WGS and as a substrate for cDNA array-based expression analyses (Kim et al., 2008; Rudd, 2003). ESTs are a few hundred base pairs of sequence derived from randomly selected cDNA clones prepared from specific tissues, and EST sequencing is inexpensive compared to WGS. These characteristics led us to test whether SNP markers could be developed from several pepper accessions using ESTs generated with next generation sequencing (NGS) technology.

NGS is a fast and low cost method for the large-scale generation of reliable and robust transcript sequences and identifying and characterizing genetic polymorphisms in plants (Imelfort et al., 2009; Metzker, 2010). The Illumina Genome Analyzer (IGA) used to be the most widely used platform based on amplified sequencing features generated by bridge PCR (Shendure and Ji, 2008). Since the IGA reads only a short sequences (75-100 base pairs; Flicek and Birney, 2009), SNPs can be identified by either de novo assembly of short sequence reads or alignment to the reference. Several factors can interfere with correct sequence alignment; these include missing calls of overlapping genotypes (Anney et al., 2008), false discovery of polymorphic SNPs (Pettersson et al., 2008), homozygote to heterozygote miscalls (Teo et al., 2007), and allelic dropout (Pompanon et al., 2005).

Here, we describe a process for SNP development from transcriptome sequencing of peppers. The resulting SNP primers clearly distinguished between 27 tested Capsicum cultivars, demonstrating that the SNP markers developed in this study will be useful resources to facilitate MABC in pepper.

Materials and Methods

Plant Material Preparation for Eight Accessions

Jeju (Capsicum annuum), LAM32 (C. annuum), Tean (C. annuum), CM334 (C. annuum), SNU-001 (C. chinense), Yuwolcho (C. annuum), PI201234 (C. annuum) and YCM334 (C. annuum) were grown in a growth chamber with 12 h light at 25°C and 12 h dark at 18°C. Leaf tissues at the same stage from the eight accessions were collected. Total RNA was isolated from leaves of each accession with Trizol extraction buffer (Ambion, Carlsbad, CA, USA) as described in the manufacturer's protocol and used for sequencing of transcriptomes.

Sequencing of Transcriptomes for Eight Accessions

Jeju, LAM32 and Tean were sequenced via GAIIx sequencing with 116-bp single-end reads at the National Instrumentation Center for Environmental Management (NICEM). CM334 sequencing data of 101-bp paired-end reads were provided by Dr. Choi from the Plant Genomics Laboratory at Seoul National University. SNU-001, Yuwolcho, PI201234 and YCM334 were sequenced with the GAIIx sequencing platform using the 90-bp paired-end read sequencing method at Beijing Genome Institute (BGI).

Quality Trimming of Reads

The sequence data from each accession were trimmed to reduce the quality deterioration in the 3’-end and 5’-end regions, which negatively affects mapping and assembly. Sickle version 1.0 (https://github.com/najoshi/sickle) was used for quality trimming.

Reference Sequence Preparation

The reference sequence, assembled by commercially available CLC Genomics Workbench software, comprised 31,196 contigs from EST sequences that were mainly derived
from Korean F1 hybrid line Bukang, for which data are available at the Korea Research Institute of Bioscience & Biotechnology (KREBB) (Ashrafi et al., 2012; Kim et al., 2008).

**Table 1. Sequencing information from eight pepper accessions.**

<table>
<thead>
<tr>
<th>Cultivar</th>
<th>Origin</th>
<th>Raw reads Bases</th>
<th>Raw reads Reads</th>
<th>Raw length (bp)</th>
<th>Trimmed reads Bases</th>
<th>Trimmed reads Reads</th>
<th>Trimmed length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeju</td>
<td>Korea</td>
<td>4,341,320,416</td>
<td>37,425,176</td>
<td>116</td>
<td>3,690,172,704</td>
<td>36,990,461</td>
<td>99.8</td>
</tr>
<tr>
<td>LAM32</td>
<td>India</td>
<td>4,120,993,728</td>
<td>35,525,808</td>
<td>102</td>
<td>3,596,017,756</td>
<td>35,239,208</td>
<td>102</td>
</tr>
<tr>
<td>Tean</td>
<td>Korea</td>
<td>4,356,564,208</td>
<td>37,556,588</td>
<td>116</td>
<td>3,708,990,466</td>
<td>37,165,176</td>
<td>99.8</td>
</tr>
<tr>
<td>CM334</td>
<td>Mexico</td>
<td>7,111,852,582</td>
<td>70,414,382</td>
<td>101</td>
<td>5,297,351,327</td>
<td>66,591,676</td>
<td>99.8</td>
</tr>
<tr>
<td>SNU-001</td>
<td>Venezuela</td>
<td>3,575,010,600</td>
<td>39,722,340</td>
<td>90</td>
<td>3,493,467,320</td>
<td>39,637,637</td>
<td>88.1</td>
</tr>
<tr>
<td>Yuwolcho</td>
<td>Korea</td>
<td>3,579,091,560</td>
<td>37,767,684</td>
<td>90</td>
<td>3,438,315,513</td>
<td>39,381,505</td>
<td>87.5</td>
</tr>
<tr>
<td>PI201234</td>
<td>Germplasm in USA</td>
<td>3,575,048,220</td>
<td>39,722,758</td>
<td>90</td>
<td>3,493,467,360</td>
<td>39,637,637</td>
<td>88.1</td>
</tr>
<tr>
<td>YCM334</td>
<td>Taiwan</td>
<td>3,525,119,100</td>
<td>39,167,990</td>
<td>90</td>
<td>3,444,541,024</td>
<td>39,081,800</td>
<td>88.1</td>
</tr>
</tbody>
</table>

*RIL line derived from a cross between Yolo wonder and CM334 in INRA (France).*
reads from Jeju, LAM32 and Tean; a total of 7.1 Gb with 70 million reads from CM334; and a total of 3.5-3.6 Gb with 39-40 million reads from SNU-001, Yuwolcho, PI201234, and YCM334 (Table 1). After quality trimming with Sickle version 1.0, the collected data were reduced to 3.6-3.7 Gb with 35-37 million reads of 100-102 bp read length for Jeju, Tean, and LAM32; 5.3 Gb with 67 million reads averaging 79.5 bp for CM334; and 3.4-3.5 Gb with 39-40 million reads averaging 87-88 bp for SNU-001, Yuwolcho, PI201234, and YCM334 (Table 1).

Alignment of Each Accession to the Reference Sequence

The total acquired reference sequence of 21,665 kb was de novo assembled from 31,196 contigs derived from Bukang ESTs, with an average contig length of 696 bp. There was a total of 1,071 kb of interspersed repeat elements (5,216 elements). Unclassified repeats accounted for 626 kb of this. Of the remaining repeat sequence, the largest portion (278 kb) was annotated as long terminal repeats (LTR), which are composed of several hundred base pairs. Long interspersed nuclear elements (LINEs), which encode a reverse transcriptase (RT) and other proteins, accounted for 118 kb. In addition, there was 49 kb of DNA transposable elements, which tend to have short inverted repeats at each end. In addition to these interspersed repeat elements, 102 kb and 103 kb were marked as simple repeat and low complexity regions, respectively. Therefore, a total of 1,276 kb (5.9%) was masked as repeat sequences.

The trimmed reads of eight accessions (Table 1) were aligned to the reference sequence and the range of alignment ratio between each accession and the reference was 52.6-70.5% (Table 2).

SNP Discovery and SNP Filtering

Sequences with read depth over 25x were used for SNP discovery, and those that were obviously distinguishable and different among accessions were defined as SNPs. Based on these criteria, a total of 58,151 SNPs were identified. From these, SNPs with only two types of genotypes and two alleles were filtered out. Among the remaining 57,502 SNPs, 33,315 were recognized as distinguishable in at least six different accessions (Table 3). In the next step, SNPs that showed a uniform segregation ratio were chosen. Among the 33,315 SNPs tested, 4,508 segregated 4:4 or 3:5 in the eight accessions. Finally, from those 4,508 SNPs, 1,910 were selected based on not having any other SNPs within 60 bp in either direction (Table 3).

Development and Validation of SNP Markers

Among 1,910 SNPs, a total of 1,282 were positioned on the hot pepper map produced by UC Davis. From these, we further selected 412 SNPs that showed clear and repeatable polymorphism, and that were evenly distributed over a ~3-cM interval in each chromosome (Supplementary Table 1). The linkage map showing the location of the 412 SNPs is presented in Fig. 1 and the SNP number per chromosome is given in Table 4. The entire length of the genetic map was 1,460.6 cM. There was an average of 34 SNPs in each linkage group, with the P1_Wild (177.1 cM) linkage group including the most SNPs. The P3 linkage group had the highest density (0.40 SNP/cM). The fewest

---

Table 2. Results of alignment of each accession to the reference sequence.

<table>
<thead>
<tr>
<th>Cultivar</th>
<th>Trimmed reads</th>
<th>Aligned reads</th>
<th>Alignment ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeju</td>
<td>36,990,461</td>
<td>19,465,349</td>
<td>52.62</td>
</tr>
<tr>
<td>LAM32</td>
<td>35,239,208</td>
<td>20,122,901</td>
<td>57.1</td>
</tr>
<tr>
<td>Tean</td>
<td>37,165,176</td>
<td>20,983,023</td>
<td>56.46</td>
</tr>
<tr>
<td>CM334</td>
<td>66,591,676</td>
<td>36,054,433</td>
<td>54.14</td>
</tr>
<tr>
<td>SNU-001</td>
<td>39,338,990</td>
<td>24,801,440</td>
<td>63.05</td>
</tr>
<tr>
<td>Yuwolcho</td>
<td>39,381,505</td>
<td>26,073,255</td>
<td>66.21</td>
</tr>
<tr>
<td>PI201234</td>
<td>39,637,637</td>
<td>27,929,380</td>
<td>70.46</td>
</tr>
<tr>
<td>YCM334</td>
<td>39,081,800</td>
<td>25,936,924</td>
<td>66.37</td>
</tr>
</tbody>
</table>

Table 3. SNP filtering process.

<table>
<thead>
<tr>
<th>Filtering criteria</th>
<th>No. of SNPs remaining</th>
<th>No. of filtered SNPs</th>
<th>Filtering percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth filtering (&gt; 25 X)</td>
<td>58,151</td>
<td>649</td>
<td>1.11</td>
</tr>
<tr>
<td>Genotype/allele</td>
<td>57,502</td>
<td>24,187</td>
<td>42.06</td>
</tr>
<tr>
<td>No. of distinguishable accessions</td>
<td>33,315</td>
<td>28,807</td>
<td>86.46</td>
</tr>
<tr>
<td>Segregation ratio</td>
<td>4,508</td>
<td>2,598</td>
<td>57.63</td>
</tr>
<tr>
<td>Adjacent SNP</td>
<td>1,910</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Filtering percentage = (No. of SNPs removed by filter / No. of SNPs tested) × 100.
Fig. 1. Location of 412 SNP markers on pepper linkage map.
Table 4. The number of SNPs and density in each pepper linkage group.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>SNPs</th>
<th>Size (cM)</th>
<th>SNP density (SNP/cM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1_Wild</td>
<td>69</td>
<td>177.1</td>
<td>0.39</td>
</tr>
<tr>
<td>P2</td>
<td>43</td>
<td>113.8</td>
<td>0.38</td>
</tr>
<tr>
<td>P3</td>
<td>57</td>
<td>141.6</td>
<td>0.40</td>
</tr>
<tr>
<td>P4</td>
<td>32</td>
<td>144.1</td>
<td>0.22</td>
</tr>
<tr>
<td>P5</td>
<td>25</td>
<td>101.6</td>
<td>0.25</td>
</tr>
<tr>
<td>P6</td>
<td>36</td>
<td>136.6</td>
<td>0.26</td>
</tr>
<tr>
<td>P7</td>
<td>21</td>
<td>104.2</td>
<td>0.20</td>
</tr>
<tr>
<td>P8_Wild</td>
<td>8</td>
<td>50.3</td>
<td>0.16</td>
</tr>
<tr>
<td>P9</td>
<td>22</td>
<td>110.2</td>
<td>0.20</td>
</tr>
<tr>
<td>P10</td>
<td>35</td>
<td>98.0</td>
<td>0.36</td>
</tr>
<tr>
<td>P11</td>
<td>33</td>
<td>90.0</td>
<td>0.37</td>
</tr>
<tr>
<td>P12</td>
<td>31</td>
<td>106.3</td>
<td>0.29</td>
</tr>
<tr>
<td>Total</td>
<td>412</td>
<td>1460.6</td>
<td>0.28</td>
</tr>
</tbody>
</table>

SNPs mapped to linkage group P8_Wild. The number of SNPs per cM ranged from 0.16 to 0.40, with an average of 0.28 SNPs per cM. Accordingly, there an average of 1 SNP per 3.55-cM interval in the genetic map (Table 4 and Fig. 1).

To validate the 412 SNP markers, 24 *C. annuum* and 3 *C. chinense* accessions (Supplementary Table 2) were tested for diversity of pepper genotypes (Fig. 2). Between *C. annuum* accessions, the average number of SNPs was 143, ranging from 23 (ECW vs. ECW30R) to 205 (Takanotsume vs. YCM334). By contrast, between bell-type accessions of *C. annuum* (9003, Dempsey, ECW30R, ECW), the average number of SNPs was 51, ranging from 23 (ECW vs. ECW30R) to 69 (DRB vs. ECW). *C. chinense* accessions showed an average of 23 SNPs, ranging from 17 (Habanero vs. SNU11-001) to 27 (Habanero vs. PI159236). The SNP number between *C. annuum* and *C. chinense* accessions averaged 179, ranging from 149 (VK-515S vs Habanero) to 209 (Dempsey vs PI159236). Overall, the average SNP number between all *Capsicum* accessions was 150, ranging from 17 (Habanero vs SNU11-001) to 27 (Habanero vs PI159236).

Discussion

Various approaches have been used to find SNPs in hot pepper, such as targeting sequences using COSII markers (Jung et al., 2010) or PCR using primers based on BAC sequences (Yang et al., 2009). Recently, ESTs have been widely used to find SNPs in crop varieties including tomato and barley because of their abundance and easy accessibility to the data (Kota et al., 2001; Labate and Baldo, 2005). The present work demonstrates that EST-derived SNP discovery using NGS is advantageous in hot pepper as well.

We prepared the reference sequence from Bukang ESTs using an approach similar to that described by Nicoli et al. (2012). However, we employed different SNP filtering methods to find valuable SNPs more effectively. By preparing a lengthy reference sequence after de novo assembly, we generated an alternative to WGS for reference sequence alignment to identify SNPs. Repeat areas are distributed throughout the genome in hot pepper. Therefore when preparing a reference sequence, repeat masking should be performed to find more accurate SNPs. In NGS for Jeju, LAM32, and Tean, parts of some reads contained low quality data and quality trimming resulted in small reductions in bases but not read number. However, for CM334, there was large amount of reduction due to low quality in both bases and reads. The sequences of SNU-001, Yuwolcho, PI201234, and YCM34 were high quality and did not show no much difference before and after trimming. Most accessions aligned to the reference with 50-70% of reads alignment. The variations might be due to the different genetic relationships between the accessions and the reference.

Overall, we identified many SNPs by aligning NGS sequence to the reference. However, the SNPs were not all equally valuable. Several factors can lead to false SNP findings, including base calling errors from NGS (Nielsen et al., 2011), miss-calls including overlapping genotypes (Anney et al., 2008), and false discovery of polymorphic SNPs that are actually monomorphic (Pettersson et al., 2008). To address this, we performed quality filtering of the SNPs. Our in-house SNP filtering process significantly decreased the SNP numbers: starting from 58,151 SNPs, the SNPs were filtered down to 1,910. After positioning the SNPs on the genetic linkage map, developing and testing SNP markers using the Fluidigm® EP1™ genotyping system, 412 SNP markers were finally
selected and validated using 27 Capsicum accessions.

For MABC, background selection focuses on both reduction of donor segments around target genes and recurrent parent genome recovery. Successful MABC depends on genome size, population size, marker density in the map and the use of high throughput marker systems. Recently, Herzog and Frisch (2013) conducted computer simulations of MABC in genetic models of sugar beet, rye, sunflower and rapeseed,
using the 10% quantile (Q10) value, the arithmetic mean and the standard deviation of the probability distribution of the proportion of recipient genome in the entire genome of selected individuals (as a percentage), for determining every backcross generation to measure recurrent parent genome recovery. Based on their data, the optimum designs, which minimize the required number of marker data points for target Q10 values of 96-99% in generation BC2 or BC3 in the model plants, employed marker densities of 2-20 cM intervals between markers and a population size of 50-100. In this study, the entire length of the genetic map covered by the SNP markers was 1,460.6 cM, suggesting that 73-730 markers (for 2-20 cM intervals) are needed for MABC in pepper. Our marker set produced an average of 150 SNPs between 27 accessions (corresponding to ~10 cM intervals), implying that our newly developed SNP markers should be useful for MABC between 27 accessions. SNP numbers between C. annuum and C. chinense were much higher than those between accessions from the same species such as bell types of C. annuum and C. chinense. This result implies that MABC could be more difficult in closely-related accessions than in distant accessions. Despite this, our new SNP markers clearly distinguished 27 Capsicum accessions.

Among the accessions tested, AC2212 was positioned in the C. annuum cluster, rather than in the C. chinense cluster. AC2212 was originally classified as C. chinense based on agronomic information from the Centre for Genetic Resources, the Netherlands (http://applicaties.wageningenur.nl/applications/cgngenis/; Wahyuni et al., 2011). However, Wahyuni et al. (2013) reported out that AC2212 likely belongs to C. annuum instead of C. chinense, based on AFLP marker analysis. This result thus indicates that MABC could be more difficult in closely-related accessions than in distant accessions. Despite this, our new SNP markers clearly distinguished 27 Capsicum accessions.

In conclusion, our SNP markers derived from transcriptome sequences will be valuable tools for MABC. In addition, the markers can be used for genetic mapping, comparative mapping and genetic diversity analyses in pepper and related species.

Literature Cited


