Simulation of QTL by sequencing for agronomic quantitative trait loci detection in small to medium population size in soybean

D Satyawan*, H Rijzaani and I M Tasma

Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development (ICABIOGRAD), Indonesian Agency for Agricultural Research and Development, Ministry of Agriculture, Jalan Tentara Pelajar No. 3A, Bogor 16111, West Java, Indonesia

*E-mail: d.satyawan@gmail.com

Abstract. Quantitative Trait Loci by sequencing (QTL-seq) is a QTL detection method that utilizes the principles of Bulked Segregants Analysis. It detects alleles with extreme frequencies in whole-genome sequence data from two bulked populations with contrasting phenotypes. This approach is less laborious than QTL detection using linkage mapping, and the result had been shown to be comparable in the same mapping population. However, since the genomes of the two bulked populations are completely sequenced, it can facilitate further characterization of the QTL segment and the genes underlying the QTLs. In this study, QTL-seq was simulated using high-density SNP genotyping data from a recombinant inbred population consisting of 188 individuals. The genomes of both parents had been sequenced, and the SNP genotyping identified 2,207 SNP markers that were polymorphic and segregating in the population. Since the markers are dense enough and well distributed across the genome, they can be used to represent the alleles that can be obtained from whole genome resequencing of bulked individuals. The availability of genotype data for each individual in the mapping population also enabled the detection of QTL via linkage mapping. Using data generated from both approaches, various simulations were conducted to compare the results that could be obtained under ideal conditions, as well as less ideal ones such as when the QTL effects are small, the presence of skewed phenotype distribution, and a small number of bulked samples.

Keywords: SNP genotyping, quantitative trait locus detection, QTL-seq, whole-genome resequencing.

1. Introduction

Detection of chromosomal segments containing genes that regulate traits of interest in crops can be a lengthy and costly endeavour. Typical forward genetics approach relies on genetic mapping of the traits in a segregating population, where the location of the gene is inferred from recombination frequency between the trait and DNA markers distributed throughout the genome [1]. Depending on the choice of methods and technology, a significant amount of time, labour, and funding are required to prepare a mapping population, scoring the trait and genotype the DNA markers on each individual, as well as the processing and analyzing the data to generate the map.

Takagi et al. [2] proposed Quantitative Trait Loci by sequencing (QTL-seq) as an alternative mapping method to shorten the time, reduce workload, and reduce cost. It is a variant of Bulked Segregant Analysis (BSA) method [3], where individuals with the highest and lowest phenotypic scores are combined into two bulks, and genotyped using generation sequencing. The allele frequencies of Single Nucleotide Polymorphisms (SNPs) identified from the sequencing data are then compared between the two bulks. SNPs located near the causal gene should show contrasting frequencies in the two bulks, and the authors termed this contrast as SNP index.

The use of whole-genome sequencing simplifies the genotyping process, since screening for polymorphic markers is no longer required. All sequence variations within the population can also be identified from the sequence data [4,5], which will assist in candidate gene identification further down the line. The number of genotyped individuals is also dramatically reduced to just three: the two bulks and one of the parents. This also reduces the wet lab work required to extract the DNA, since each bulk can be treated as a single sample during DNA extraction.

However, some disadvantages can also arise. Since the bulk can only be identified after phenotyping, the total time for phenotyping and genotyping may be longer compared to traditional QTL analysis, where individuals can be genotyped soon after the plants germinate. Sequencing and bioinformatics analysis can also take around one month, so in some cases, this approach could be more time-consuming than regular QTL mapping. Genotyping must also be repeated for multiple traits, as different phenotypes will likely produce the different high and low bulk composition. This approach also needs high depth sequencing [5], which can be expensive for species with large genome size [6]. The resulting analysis also cannot estimate R2 value, as well as the effect of heterozygous alleles.

Despite those disadvantages, this method can still represent a significant saving in some situations and species. Consequently, we attempted to test the applicability of this method in our soybean breeding program under ideal and non optimal conditions. Several non ideal scenarios were tested, such as small sampling size, weak QTL effects and skewed distribution. The results were then compared with regular QTL analysis on a recombinant inbred lines population.

2. Materials and methods

The mapping population consisted of 188 recombinant inbred soybean lines derived from a cross between Tambora and B3293. The phenotype data for this simulation was taken from a single location at Cibalagung, Indonesia. Genotyping was performed using Illumina Soy SNP 6K on the Illumina iScan platform. QTL analysis was first performed using Windows QTL Cartographer [8] to identify strong and weak QTL for comparison with QTL-seq.

QTL-seq simulation was done by first identifying individuals with the highest and lowest phenotypic scores for plant height, seed weight and flowering time traits, to be bulked for SNP index calculation. The bulk size was set at 20 and 10 individuals to see the effect of bulk size on QTL detection. Marker data from each individual were then combined to form a pooled genotype data in each bulk. The delta SNP index (dSNP index) was calculated as the value of the frequency of a major allele in the 'high' bulk minus the frequency of that allele in the 'low' bulk.

dSNP index = (major allele frequency in high bulk)–(allele frequency in low bulk). As an example, a SNP has A and C alleles in the mapping population. The number of A allele in the high bulk of 20 individuals was 32, so its allele frequency is 32/(20×2) or 0.8. In the low bulk, there were only 7 A allele, so the frequency is 7/40 or 0.175. The dSNP index for that marker thus equals to 0.8–0.175 or 0.625. The resulting values for each marker were then plotted as a scatter plot using qqman package in R [8]. For easy comparison, the LOD scores from the same traits from composite interval mapping (CIM) analysis in Windows QTL Cartographer were also plotted in similarly.

3. Results and discussion

The SoySNP6K contains 5,236 SNPs markers selected from Song et al. (2013), which could be assayed concurrently for each individual's DNA. Among those SNP markers, 2,207 were polymorphic in the mapping population, and it should be dense enough to represent most segments that underwent recombination in the mapping population. Such a high marker density should be able to simulate the pattern generated by even denser marker generating technology like next generation sequencing (NGS).

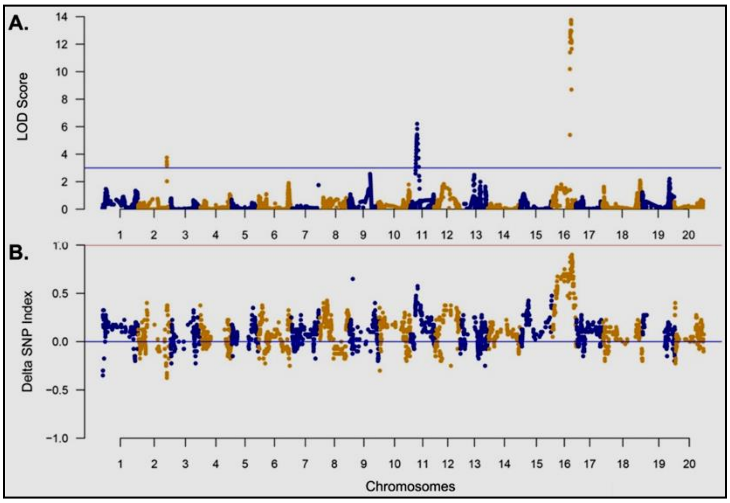

Figure 1. Manhattan plot of QTL detection using CIM in Windows QTL Cartographer (A) and QTL-seq (B) for plant height.

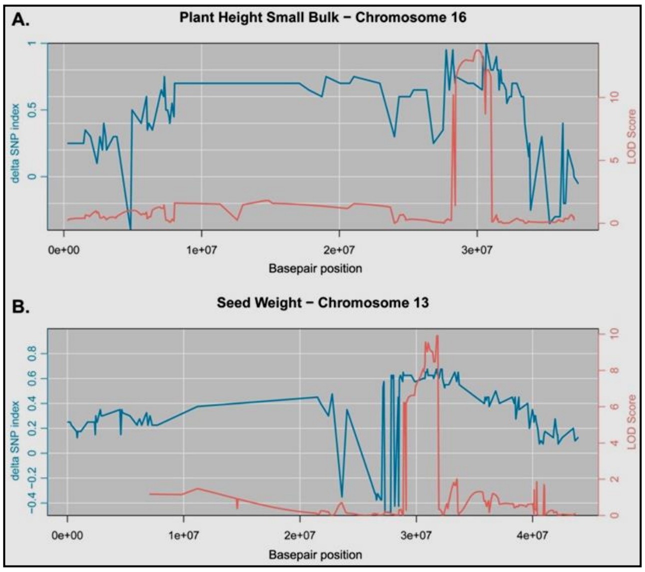

Figure 2. Line plot of LOD scores from Windows QTL Cartographer (blue) and dSNP index from QTL-seq (red) in chromosomes where a strong QTL was detected for plant height (A) and seed weight (B).

For strong QTLs, such as one found in chromosome 16 for plant height (LOD score >13), the locus also showed up as the highest peak in QTL-seq (Figure 1). Likewise, a major QTL for seed weight was detected in chromosome 13 (LOD score >10). Closer inspection of chromosome 16 revealed that the base interval identified by both QTL-seq and regular QTL analysis was mostly similar, with the highest peaks from QTL-seq flanked the same interval that had the highest LOD score in composite interval mapping. However, the shape of the peak from each method was very different (Figure 2).

Composite interval mapping was able to estimate the location of the QTL even when there is no marker data for that segment, an ability that is absent in QTL-seq analysis. For the strong QTL for seed weight in chromosome 13, the peak shapes from each method were also different, but as in plant height, the highest dSNP index values and the highest LOD scores were found to flank largely similar chromosome intervals (Figure 3).

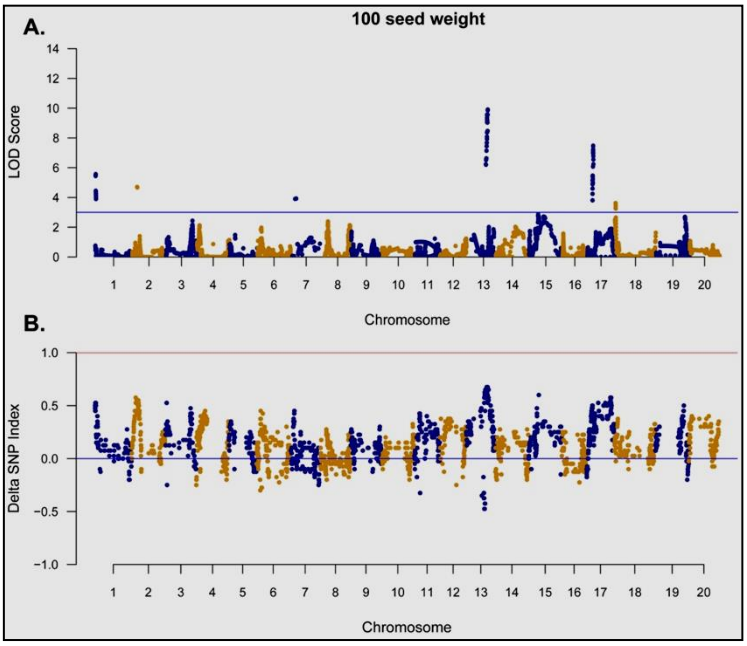

Figure 3. Manhattan plot of QTL detection using CIM in Windows QTL Cartographer (A) and QTL-seq (B) for seed weight.

For weaker QTLs found in plant height and seed weight mapping (LOD score <8), the loci were not always reliably detected using QTL-seq. The second highest QTL for plant height in chromosome 11 was also detected in QTL-seq, but the highest dSNP index in this chromosome was lower than one found in chromosome 9 (Figure 1B), where no significant QTL were detected using composite interval mapping.

The second most significant QTL for seed weight in chromosome 17 was not detected in QTL-seq, as another locus in the same chromosome and several loci in other chromosomes had a higher dSNP index than this QTL (Figure 3B). QTL-seq thus performed less reliably for less dominant QTLs in this mapping population.

Another factor that can influence QTL detection in QTL-seq is the size of the bulk. We observed an increase of dSNP index values when the bulk size was reduced from 20 to 10 at the major QTL for plant height at chromosome 16, although the QTL interval also became enlarged compared to CIM result (Figure 4A and 4B). Logically, the use of smaller bulk size will eliminate individuals with average phenotype scores [11], resulting in the elimination of weak QTL alleles and larger differences of SNP index between the bulk at the expense of decreased sensitivity toward weak-effect QTL. The potential downside is that smaller bulk size will also make the analysis more sensitive to phenotyping error, as individuals that are phenotyped incorrectly will subsequently represent a larger percentage of the bulk. Bulk size reduction appeared not to diminish the ability of QTL-seq to detect weaker QTL, as illustrated by the weaker QTL for plant height at chromosome 11. As in the major QTL for this trait, the dSNP index was higher than when bulk size was set at 20, but the peak value has shifted away from the peak LOD score detected by CIM (Figure 4C and 4D). In this case, we conclude that although smaller bulk size can improve the contrast between bulks, it can reduce the accuracy of the detection of QTL region by several hundred thousand base pairs.

The last non ideal condition that we tested was skewed phenotypes. In CIM, it is possible to transform the phenotype data of a skewed population to make it resemble a normal distribution [11], and CIM will use the transformed data to find the QTL location. In QTL-seq, phenotype data are only used to select the individuals, so unless the transformation reshuffle the order of the highest to lowest individuals in the phenotype table, the same individuals will still be selected for the highest and lowest bulk respectively, which will negate the benefit of phenotype data transformation.

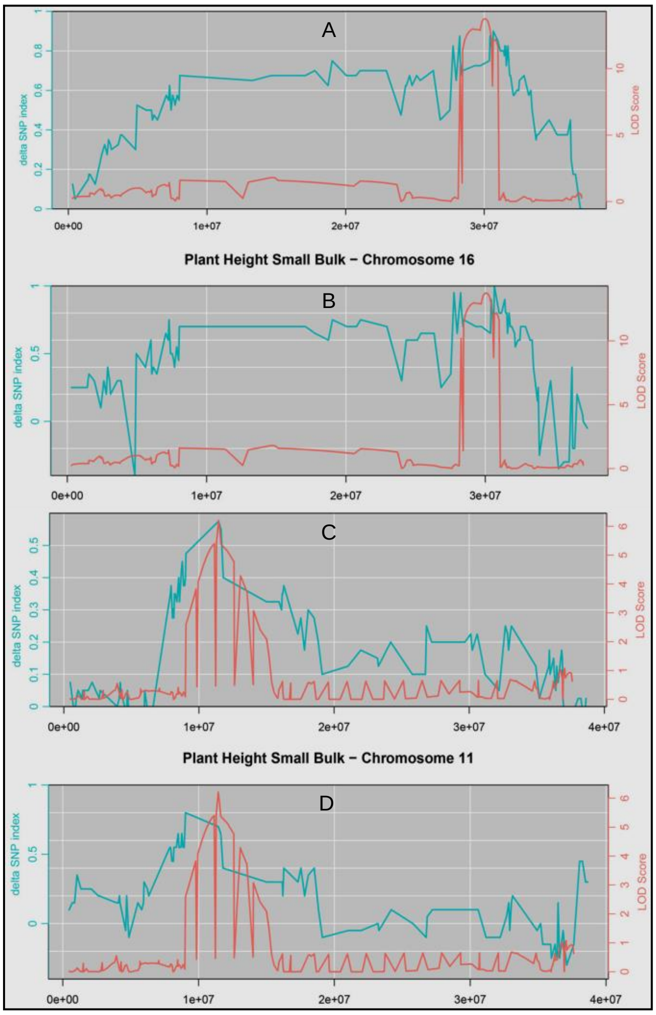

Figure 4. Line plots of LOD scores from Windows QTL Cartographer (blue) and dSNP index from QTL-seq (red). For a strong QTL, a bulk size of 20 individuals (A) and 10 individuals (B) detected similar interval as QTL. For a weaker QTL, the larger bulk size (C) is more consistent with CIM results than the smaller bulk size (D).

The trait that exhibited skewed distribution was flowering time, where the population skewed toward early flowering and a small number of individuals flowered late (Figure 5A). The bulk of early flowering individuals thus could potentially contain a very diverse genetic composition, as the earliest flowering time contained a large percentage of the population, resulting in low contrast between the late and early flowering bulks. As expected, under such condition QTL-seq did not produce outstanding dSNP index values in loci that were detected by CIM as the location of flowering time QTL in chromosome 11 and chromosome 16 (Figure 5B and 5C).

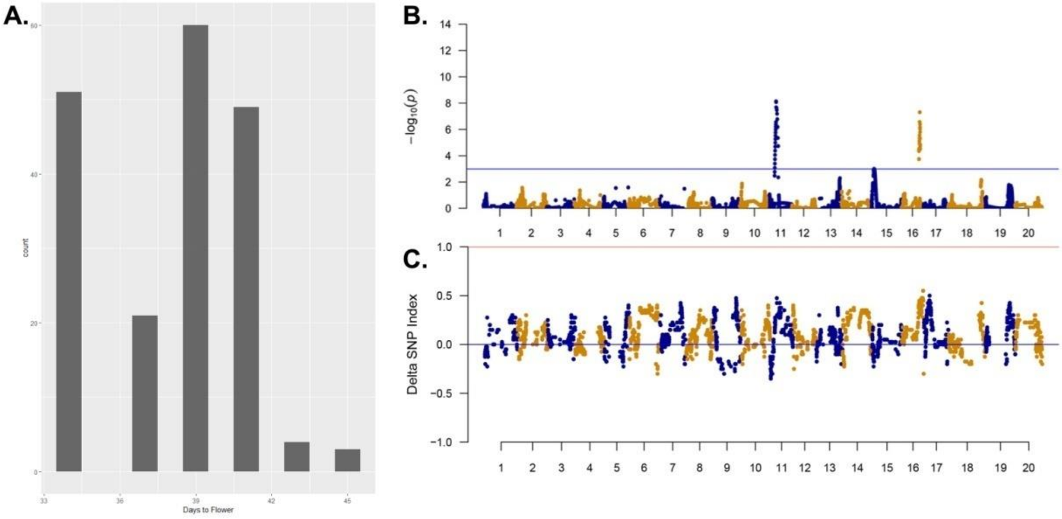

Figure 5. Histogram of flowering time trait in the population (A) and Manhattan plots of QTL detection using CIM in Windows QTL Cartographer (B), and QTL-Seq (C) for flowering time.

4. Conclusions

QTL-seq could reliably detect strong QTL, but less sensitive for detecting weaker effect QTLs. Charts from dSNP index values had different pattern from LOD score charts from linkage-based QTL analysis, but the highest peaks from both methods flanked the same loci, although the QTL interval may shift when the size of the bulk is small and the QTL is relatively weak. QTL-seq was also found to be sensitive to skewed trait distribution. Nevertheless, QTL-seq could be a more economical and less laborious alternative QTL detection method for less complex traits.

5. References

[1] Collard B C Y, Jahufer M Z Z, Brouwer J B and Pang E C K 2005 An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts *Euphytica* **142** 169–96

[2] Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S, Innan H, Cano L M, Kamoun S and Terauchi R 2013 QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations *Plant J.* **74** 174–83

[3] Michelmore R W, Paran I and Kesseli R V 1991 Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. U. S. A.* **88** 9828–32

[4] Varshney R K, Nayak S N, May G D and Jackson S A 2009 Next-generation sequencing technologies and their implications for crop genetics and breeding *Trends Biotechnol.* **27** 522–30

[5] Trick M, Adamski N, Mugford S G, Jiang C C, Febrer M and Uauy C 2012 Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat *BMC Plant Biol.* **12** 14

[6] Ray S and Satya P 2014 Next generation sequencing technologies for next generation plant

breeding *Front. Plant Sci.* **5** 367

[7] Wang S, Basten C J and Zeng Z B 2005 Windows QTL cartographer version 2.5 *Stat. Genet. North Carolina State Univ. Raleigh*

[8] Turner S D 2014 qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots *bioRxiv* 5165

[9] Song Q, Hyten D L, Jia G, Quigley C V, Fickus E W, Nelson R L and Cregan P B 2013 Development and evaluation of SoySNP50K, a high-density genotyping array for soybean ed T Zhang *PLoS One* **8** e54985

[10] Magwene P M, Willis J H and Kelly J K 2011 The statistics of bulk segregant analysis using next-generation sequencing ed A Siepel *PLoS Comput. Biol.* **7** e1002255

[11] Goh L and Yap V B 2009 Effects of normalization on quantitative traits in association test *BMC Bioinformatics* **10** 415