Published online September 5, 2019

A Systematic Gene-Centric Approach to Define Haplotypes and Identify Alleles on the Basis of Dense Single Nucleotide Polymorphism Datasets

Aurélie Tardivel, Davoud Torkamaneh, Marc-André Lemay, François Belzile, and Louise S. O'Donoughue*

A. Tardivel, D. Torkamaneh, M.-A. Lemay, F. Belzile, Dép. de Phytologie and Institut de Biologie Intégrative et des Systèmes, Univ. Laval, Quebec City, QC, Canada, G1V 0A6; L.S. O'Donoughue, CÉROM, Centre de recherche sur les grains Inc., 740 chemin Trudeau, Saint-Mathieu-de-Beloeil, Canada, QC J3G 0E2.

ABSTRACT Assessing the allelic diversity within a germplasm collection and identifying individuals carrying favorable alleles is challenging. Advances in high-throughput technologies allow the genotyping of many individuals for thousands of markers but bridging the gap between single nucleotide polymorphisms (SNPs) and relevant alleles remains difficult. We developed a systematic approach that defines haplotypes from large SNP catalogs that aims to identify haplotypes that can be equated to alleles at given genes. Unlike haplotype visualization tools, our approach selects SNP markers that flank a gene and define haplotypes that correspond to this gene's alleles. We tested this approach on four known soybean [*Glycine max* (L.) Merr.] maturity genes (*E1*, *GmGia*, *GmPhyA3*, and *GmPhyA2*) in a collection of 67 lines and two genotypic datasets [a SNP array and genotyping-by-sequencing (GBS)]. For *E1*, *GmGia*, and *GmPhyA3*, we identified SNP haplotypes such that the allele found at these genes could be accurately predicted from the haplotype in 97.3% of the cases. For these genes, of the 12 known alleles in the collection, 10 and 8 could be correctly predicted from the haplotypes found with the SNP array and GBS datasets, with success rates of 98 and 97% for all allele–line combinations, respectively. The approach proved equally successful for data derived from a SNP array and GBS. However, in the case of *GmPhyA2*, a lack of markers in the genomic region prevented the identification of alleles, regardless of the dataset. We demonstrate the feasibility and reproducibility of our approach and identify limits to its applicability.

Abbreviations: GBS, genotyping-by-sequencing; LD, linkage disequilibrium; r^2_{vs}, correlation coefficient considering relatedness and population structure; SNP, single nucleotide polymorphism; WGS, whole-genome sequencing.

CORE IDEAS

- A gene-centric approach for haplotype definition was developed and implemented in R.
- The tool allows for allelic characterization at given loci in germplasm collections.
- Allelic status at four maturity genes is predicted on the basis of marker genotyping data.

OVER THE LAST DECADE, the affordability and availability of genome-wide genotyping data of breeding lines through approaches such as GBS and microarrays has increased significantly, giving rise to new breeding approaches such as genomic selection. In parallel, genes underlying economic traits are constantly being reported and natural variation in these genes contributes to variation in phenotypes. For such known genes, it is useful to assess the allelic diversity captured in genotyped germplasm collections and to identify individuals carrying favorable alleles. Some mutations in or near a gene can cause a complete loss-of-function, whereas others may lead to a partial or leaky phenotype. Moreover, as independent mutations occur in distinct genetic backgrounds, neighboring quantitative trait loci, genes and alleles could be different. Thus two null alleles for the same gene,

Citation: Tardivel, A., D. Torkamaneh, M.-A. Lemay, F. Belzile, and L.S. O'Donoughue. 2019. A systematic gene-centric approach to define haplotypes and identify alleles on the basis of dense single nucleotide polymorphism datasets. Plant Genome 12:180061. doi: 10.3835/plantgenome2018.08.0061

Received 28 Aug. 2018. Accepted 1 Apr. 2019.
*Corresponding author (louise.odonoughue@cerom.qc.ca)

presumed to be equivalent in terms of their phenotypic consequences (e.g., causing a complete loss of function), could be strongly linked with different alleles at neighboring loci. In a breeding context, the introgression of these linked alleles in elite material may not be equally useful, as undesirable alleles may be transmitted through linkage drag (Narvel et al., 2001). In this context, and in addition to knowing which individuals carry a favorable allele at a given locus, distinguishing between alleles, even favorable ones, can be helpful for choosing the best source for a trait within the germplasm. Therefore, developing tools that are able to extract such information would prove beneficial.

Linkage disequilibrium (LD) between two markers describes the probability of these markers being inherited together more often than would be expected by chance. The study of LD is of great interest, as it is a key concept for genome-wide association studies, gene mapping, or understanding evolutionary processes. The level of LD is influenced, among other factors, by selection, recombination, mutation, population size, and genetic drift (Slatkin, 2008). The extent of LD is expected to decay proportionally with distance along the chromosome and the number of generations over which a combination of alleles has been transmitted. However, the distance over which LD decays can vary between species and collections of accessions (wild, landrace, and elite) (Hyten et al., 2007; Zhou et al., 2015). Similarly, within the genome, LD extends over much longer distances in pericentromeric regions than in the more gene-rich euchromatic regions (Bastien et al., 2014; Zhou et al., 2015). Linkage disequilibrium also decays much more rapidly in recombination hotspots, which are regions that exhibit unusually high recombination rates caused by unstable molecular structures or the presence of recombination initiation sites (Myers et al., 2005, 2008). Although markers separated by large physical distances (Mb) are expected to behave independently, the occurrence of LD between highly distant markers can be observed. This phenomenon can underlie bias attributable to various processes such as population structure or admixture, epistatic selection, genetic drift, structural variation, or genotyping artifacts (Koch et al., 2013).

Various measures of disequilibrium have been proposed, but *D*' and r^2 are the most commonly used. These are based on a common basic pairwise disequilibrium coefficient (*D*) but differ from each other in their consideration of rare alleles (reviewed by Zondervan and Cardon, 2004). Whereas a value of $D' = 1$ (or –1) between two markers means that the rarer allele at one marker occurs exclusively with one allele at the other marker (and thus there is no evidence of recombination), $r^2 = 1$ means that the two markers have identical allele frequencies and one allele predicts the other (a perfect correlation). More recently, an approach has been proposed by Mangin et al. (2012) to correct the measure of r^2 by considering relatedness (r^2_s), population structure (r^2_v), or both (r^2_{vs}), thus allowing the capture of “true” LD as opposed to estimates of LD that are biased by population structure and admixture.

For a given gene, a range of alleles can typically be found in a large collection. Analyzing LD in a region containing a gene of interest could reveal very close markers that are not in disequilibrium and distant markers that are in disequilibrium. Indeed, a range of alleles can exist at a given gene and markers in LD with each given mutation can be found in the vicinity of the gene. These markers may not be in disequilibrium with each other, so the extent of disequilibrium with each causal mutation may be different. The maintenance of LD between causal mutations and neighboring loci can vary depending on the allele frequencies and selection pressure. As explained by Sabeti et al. (2002), common alleles will typically present a short extent of LD, while rare alleles may have short or extended LD. However, alleles present at high frequency can maintain an atypically long range of LD if the LD arose in the gene pool through positive selection. In the context of breeding, factors like inbreeding and genetic bottlenecks can contribute to long-range LD for such alleles.

Some approaches and software tools have been designed to analyze and visualize patterns of LD in genetic data through the characterization of haplotype blocks. Three approaches have been used to define haplotypes: (i) the confidence interval approach (Gabriel et al., 2002), (ii) the four-gamete rule (Wang et al., 2002), and (iii) the Solid Spine of LD approach (Barrett et al., 2005). The elucidation of haplotype block structure has proven useful, for example, in LD analysis, genome-wide association studies or redefining intervals around a quantitative trait locus (Hwang et al., 2014; Bandillo et al., 2015; Contreras-Soto et al., 2017). On the other hand, haplotypes for a targeted gene of interest can be defined by selecting a block (interval) containing the gene or of several blocks surrounding the gene. Because markers in LD with the allelic variation at a gene are expected to be near this gene, the selection of the closest block(s) will define informative haplotypes. Depending on the approach and the selected parameters, variable results in terms of block definition can be obtained. In our experience, choosing the most appropriate method and set of parameters to define and select blocks, with the aim of systematically obtaining a good fit between haplotypes and alleles for a given gene, is not trivial. Indeed, the three standard approaches named above, though relatively efficient at identifying haplotypes surrounding the genes of interest, required several manual adjustments to the parameters and were not systematic and reproducible enough for our needs.

In this study, we have developed a versatile and systematic approach to facilitate the process of defining informative haplotypes from a set of high-density SNP markers. As opposed to the other approaches aimed at defining haplotype blocks (by grouping a set of markers inherited together on a chromosome scale), the approach we describe here is a gene-centric haplotyping process that aims to select only the markers near a gene that are in LD with this gene. This haplotyping approach has a different focus (centered around a gene of interest) and thus allows the identification of a set of SNPs that can be used to describe a target gene's haplotypes in a more focused and intuitive way. The objectives of this work were to develop a gene-centric haplotyping tool, which we called HaplotypeMiner, and to test how well this approach can capture the allelic diversity at genes of interest located in various regions of the soybean genome. We also aimed to assess its potential and limitations when different genotypic datasets are used.

MATERIALS AND METHODS

Plant Materials

A set of 67 Canadian soybean lines, chosen to be representative of short-season soybeans and allelic variation at four well-known maturity genes (*E1*, *GmGia*, *GmPhyA3*, and *GmPhyA2*), was used to examine and calibrate our approach. These lines are a subset of 102 short-season Canadian lines that have recently been analyzed via whole-genome sequencing (WGS) (Torkamaneh et al., 2018). From the full collection of 102 lines, we selected almost all registered varieties as well as a few advanced breeding lines of particular interest, given their maturity phenotype. Among these 67 lines, 50 were also in common with a previously published collection [Set C in Tardivel et al. (2014)]. Maturity groups ranged from 000 to II.

Genotyping

Three distinct sets of SNP markers, with physical positions based on Assembly 2 of the soybean reference genome (Wm82.a2; Schmutz et al., 2010; Song et al., 2016), were used in this work: (i) a WGS dataset, (ii) a simulated SNP array (SoySNP50K) dataset, and (iii) a GBS dataset. The WGS data for the 67 lines were simply extracted from the larger catalog of 102 lines for which such data were available. For the purpose of simulating SNP array data, the WGS data were filtered to remove indels, and residual heterozygotes were recoded as missing values. Indels and heterozygous variants were nevertheless used in the WGS dataset for assessing the true alleles, as it is known that indels serve as diagnostic variants for some alleles. Finally, SNP markers with $\geq$60% missing data were removed. To simulate a SNP array dataset, all polymorphic markers in the WGS catalog corresponding to nucleotide positions interrogated by the SNP array (Song et al., 2013) were extracted for the 67 Canadian lines used here. Finally, the same collection was genotyped via GBS. DNA extraction, preparation, and sequencing of the GBS libraries were performed as described by Tardivel et al. (2014). The Fast-GBS pipeline (Torkamaneh et al., 2017) was used for variant discovery and the resulting genotypes were filtered as described above for the WGS dataset, except that a more permissive threshold for missing data was used ($\leq$80%), given that a more complete set of variants facilitates the imputation of GBS data (Torkamaneh and Belzile, 2015). Imputation was performed for the GBS catalog only with Beagle version 4.0 and the default parameters (Browning and Browning, 2007). To prevent the erroneous imputation of missing SNPs that were caused by large deletions, we inferred the presence of previously published alleles known to result from such deletions by analyzing the presence or absence of reads at positions known to be captured through GBS. Specifically, to test for a 13-kb deletion in the *E3* gene (the *e3-tr* allele) (Watanabe et al., 2009), we assessed depth of coverage at three positions (*Gm*19:47641476, *Gm*19:47645909, and *Gm*19:47649947). Similarly, for a 130-kb deletion in the *E1* gene (the *e1-nl* allele)(Xia et al., 2012), we measured the depth of coverage at two positions (*Gm*06:20193636 and *Gm*06:20207387). Individuals with no reads mapping to these positions were deemed to carry the deletion allele. The GBS variant catalog was then edited for each individual carrying one of these two deletions. This information was coded as two supplementary "stand-in" SNP markers on chromosome 06 at position 20,080,746 and on chromosome 19 at position 47,638,565. One allele coded for the deletion; the alternate allele coded for the nondeleted version of these genomic regions.

Allelic Characterization Based on WGS

The WGS dataset was used for allelic characterization. This characterization was performed by inspecting the nonimputed WGS dataset at the known positions of the causal mutations reported in the literature (Supplemental Table S1). Large regions of missing data in the deleted genomic regions characteristic of *e1-nl* and *e3-tr* were also used to define individuals carrying such deletions. As the reference genome did not carry the *Ty1/copia*-like retrotransposon characteristic of the *e4(SORE-1)* allele in the *GmPhyA2* gene, the WGS data were not helpful for detecting this insertion. Therefore, all individuals were tested for the presence or absence of the ~6.2 kb retrotransposon insertion with the primers and polymerase chain reaction conditions described by Liu et al. (2008). Finally, data from a previous characterization of alleles at the *GmPhyA3* gene were also used for validation (Tardivel et al., 2014).

Haplotyping Based on the SNP Array and GBS

Haplotyping was performed on both genotypic datasets (GBS and the simulated SNP array) by selecting pairs of markers flanking the central nucleotide position in the four genes of interest and estimated to be in LD with each other. Two measures were first used to estimate disequilibrium between markers: the r^2 and the corrected r^2_{vs} measure, which takes information about genetic relatedness and population structure into account. To calculate the r^2_{vs} values for all pairwise SNP combinations across the chromosome, we used the LDcorSV package (Mangin et al., 2012; Desrousseaux et al., 2017). Kinship values were estimated via the centered identity-by-state method (Endelman and Jannink, 2012) but setting a higher threshold for the minor allele frequency ($\geq$5%). Population structure was estimated with fastSTRUCTURE (Raj et al., 2014). After filtering for minor allele count $\geq$ 4, SNPs in high LD ($r^2_{\mathrm{vs}} \geq 0.8$) were then grouped into tag SNPs to reduce the amount of redundancy. This tagging was performed independently on each side of the gene,

Table 1. Number of single nucleotide polymorphisms (SNPs) identified via whole-genome sequencing (WGS) and two genotyping tools [a simulated SNP array and genotyping-by-sequencing (GBS)] on the four chromosomes carrying maturity genes in a collection of 67 Canadian soybean lines. For each method, the number of SNPs remaining at two different minor allele frequency (MAF) thresholds is indicated.

Gm a2.v1	WGS		SNP array†		GBS	
Chromosome	MAF $\geq$ 0.01	MAF $\geq$ 0.05	MAF $\geq$ 0.01	MAF $\geq$ 0.05	MAF $\geq$ 0.01	MAF $\geq$ 0.05
*Gm*06	181,060	148,124	1,818	1,690	1,201	1,013
*Gm*10	152,913	131,835	1,947	1,749	1,195	975
*Gm*19	190,295	163,850	2,229	2,006	1,233	1,021
*Gm*20	148,002	90,071	1,562	1,245	1,197	746
All	3,406,724	2,693,744	38,264	33,152	24,048	18,797

† Simulated array data extracted from the WGS data.

and the marker closest to the center of the gene was kept as the representative tag SNP. For haplotyping, all pairs of tag SNPs flanking the central gene position with values of $r^2_{vs} \geq 0.5$ were used. To test the effect of the distance parameter, maximum distances of either 250 kb or 1 Mb between selected tag SNPs were tested. For each gene, the number of tag SNPs retained and the number of haplotypes generated for each gene were recorded. The size of the haplotypes was recorded as the distance between the two most distant tag SNPs retained for haplotyping.

To evaluate the accuracy of the diagnostic haplotypes for maturity alleles identified in this study, we used a set of 32 Plant Introduction lines for which GBS, SNP array (Copley et al., 2018), and WGS data (Torkamaneh et al., 2019) were available. The same set of markers that have been predicted to be associated with different haplotypes was extracted from GBS and SNP array data of 32 Plant Introduction lines and then compared with the WGS dataset to assess the accuracy of the predicted alleles.

The haplotyping approach described here was implemented in the R programming language (R Core Team , 2016) and is publicly available from GitHub as a package called HaplotypeMiner (github.com/malemay/HaplotypeMiner). The package includes functions to perform the steps of variant filtering, LD computation, SNP tagging, and haplotype enumeration seamlessly in a single function call. Computation time is usually <1 min for a single gene and a typical GBS or SNP array dataset. Parameters are customizable for the needs of a particular study and a set of graphical functions allows the rapid generation of plots and statistics from the output of the analysis. To compare HaplotypeMiner with other haplotyping methods, we also tested how the confidence interval (Gabriel et al., 2002), four-gamete rule (Wang et al., 2002), and Solid Spine of LD (Barrett et al., 2005) approaches performed with our datasets; the associated methods and results can be found in Supplemental File S2.

RESULTS

Genotypic Data

A WGS dataset comprising 3406,724 polymorphic SNPs was obtained for this collection of 67 soybean lines (Table 1). Of the ~60,000 SNPs that could be tested with the SNP array, 38,264 were polymorphic within these WGS data (with 6.5% missing data). This subset of polymorphic WGS-derived SNPs was extracted from the complete WGS dataset and used to constitute a simulated array-derived catalog of SNPs. From the same collection of lines, 24,048 polymorphic SNPs were called via a GBS approach. Analysis of read coverage in the GBS dataset detected 39 and 30 individuals devoid of reads mapping to the genomic intervals corresponding to deletions that characterized the *e1-nl* and *e3-tr* alleles, respectively. For the simulated SNP array, missing data at two SNPs within the E1 locus and one SNP in the *GmPhyA3* locus were used to diagnose the presence of the deletions. The overlap between the simulated array data and the GBS data was fairly limited, as only 927 SNP markers were in common between them.

Identification of Alleles at Maturity Genes in Each Line

The comparison of variants (SNPs and indels) in the WGS dataset with the known causal variants for the various alleles at the *E1*, *GmGia*, *GmPhyA3*, and *GmPhyA2* genes allowed us to identify most of the alleles at these four genes (Table 2). Individuals carrying the *e1-nl* and *e3-tr* alleles were characterized by missing data in the genomic intervals (~130 kb for the *E1* gene and ~13 kb for the *GmPhyA3* gene) corresponding to the causal deletions. For *GmPhyA2*, a previously unreported

Table 2. Alleles at four maturity genes and their frequency in a collection of 67 Canadian soybean lines as defined by whole-genome sequencing data or molecular validation for the large indels of *e1-nl*, *e3tr*, and *e4* (*SORE-1*) alleles.

Gene	Gene model	Allele	Count	%
E1	*Glyma.06g207800*	*E1*	1	1.5
		e1-nl	39	58.2
		e1-as	25	37.3
		e1-fs	2	3.0
GmGia	*Glyma.10g221500*	*E2-in*	13	19.4
		e2-ns	54	80.6
GmPhyA3	*Glyma.19g224200*	*E3Ha*	30	44.8
		E3Mi	1	1.5
		E/e3p.Thr832Ala	2	3.0
		e3tr	30	44.8
		e3-fs	1	1.5
		e3Mo	3	4.5
GmPhyA2	*Glyma.20g090000*	*E4*	36	53.7
		e4 (*SORE-1*)	17	25.4
		e4p.T832QfsX21	13	19.4
		Heterozygote	1	1.5

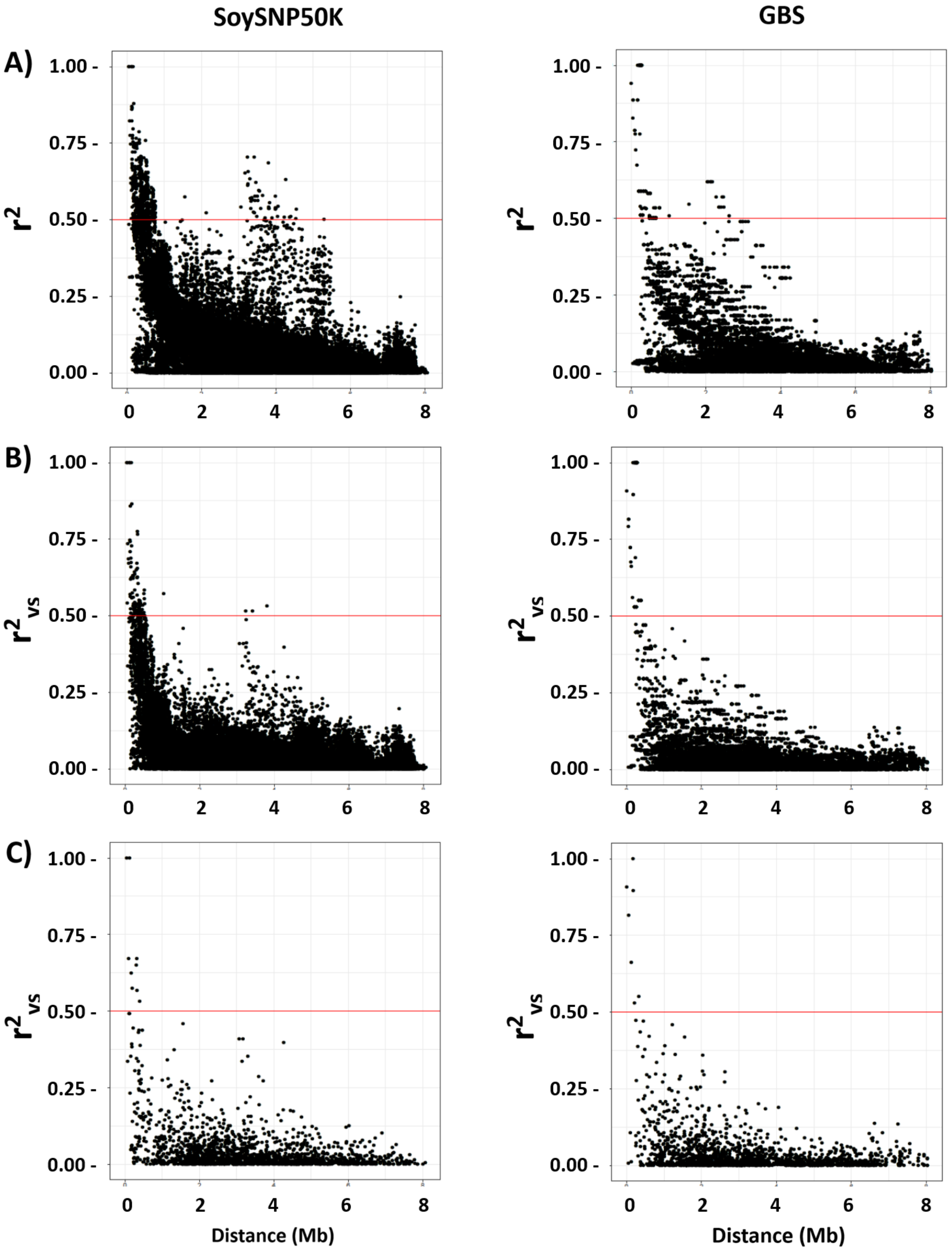

Fig. 1. Estimation of disequilibrium between pairs of markers flanking the *GmPhyA3* gene as a function of distance between those marker pairs in either genotyping-by-sequencing (GBS) or the simulated single nucleotide polymorphism (SNP) array dataset. (A) r^2 between pairs of markers (all SNPs) flanking the gene. (B) r^2 considering relatedness and population structure (r^2_{vs}) between pairs of markers (all SNPs) flanking the gene. (C) r^2_{vs} between pairs of tag SNPs flanking the gene. The horizontal red line corresponds to the r^2 or r^2_{vs} threshold used for the selection of flanking pairs.

single-nucleotide variant at Position 33,239,715 was found in the WGS catalog. This variant was found to correspond to a 1-bp deletion (A/–) in Exon 2 of the *GmPhyA2* gene, resulting in a frameshift mutation and a premature stop codon (Supplemental Fig. S1). As this nonfunctional variant has not been reported previously, it is proposed that this allele be named *e4p.T832QfsX21*. Finally, according to polymerase chain reaction, 17 individuals tested positive for the presence of the *Ty1/copia*-like retrotransposon characteristic of the *e4(SORE-1)* allele in the *GmPhyA2* gene. The known and new alleles described above and their prevalence in this collection of lines are shown in Table 2.

Gene-Centric Systematic Haplotyping

Tagging SNPs

Two measures (r^2 and r^2_{vs}) were used to estimate disequilibrium on the filtered set of markers (minor allele count $\geq$ 4). For r^2_{vs} estimation, we first assessed the population structure and relatedness within our collection of lines. Each line was assigned by fastSTRUCTURE to one of five (GBS dataset) or six (SNP array dataset) subpopulations. The relatedness between lines was assessed through an identity-by-state matrix calculated for each dataset. The difference between these two measures (r^2 and r^2_{vs}), for the *GmPhyA3* gene on two different genotyping datasets are illustrated in Fig. 1A, B. This figure shows that r^2_{vs}

Table 3. Number of tag single nucleotide polymorphisms (SNPs) retained for four chromosomes after tagging of SNPs in high linkage disequilibrium [r^2 for relatedness and population structure r^2_{vs}) ≥ 0.8] on each side of the gene under study using either the simulated SNP array or genotyping-by-sequencing (GBS).

Gene (chromosome)	Dataset	SNPs†	Tag SNPs
E1	SNP array	1760	320
(*Gm06*)	GBS	1088	294
GmGia	SNP array	1883	321
(*Gm10*)	GBS	1082	271
GmPhyA3	SNP array	2170	286
(*Gm19*)	GBS	1119	244
GmPhyA2	SNP array	1426	308
(*Gm20*)	GBS	945	257

† Number of SNPs used after applying a minor allele count ≥ 4 filter.

resulted in a more accurate view of LD, with fewer pairs of markers exhibiting r^2_{vs} values ≥ 0.5 than r^2. This is especially true for distant pairs of SNPs that exhibited suspiciously high LD values despite being very far apart.

Tagging was performed to identify and remove SNPs that were in high LD (r^2_{vs} ≥ 0.8) with each other and thus redundant for the purpose of this analysis. Single nucleotide polymorphism tagging was performed independently upstream and downstream of each targeted gene. As shown in Table 3, after tagging, between 286 and 321 tag SNPs were found for the simulated SNP array dataset and from 244 to 294 tag SNPs were found in the GBS dataset. Although the simulated SNP array dataset harbored a significantly higher number of SNPs than the GBS dataset (>1.5 times) before tagging (Table 1 and Table 3), it showed only a slightly higher number of tag SNPs. Furthermore, interestingly, although chromosome 19 showed the highest number of SNPs, this chromosome also showed the lowest number of tag SNPs. In Fig. 1B, C, comparing the LD landscapes for pairs of markers before and after tagging shows that the number of marker pairs flanking the *GmPhyA3* gene that exceeded the chosen selection threshold LD (r^2_{vs} ≥ 0.5) for pairs of tag SNPs was significantly reduced after tagging, thus providing a reduction in the number of pairs that would be retained for haplotyping.

Defining Haplotypes on the Basis of the Tag SNPs of Flanking SNPs

Pairs of tag SNPs flanking each gene were then selected with a maximal distance between tag SNPs of either 250 kb or 1 Mb, as we suspected that this parameter may have had a critical influence on the results. For each of the four maturity genes, the resulting collection of SNPs was used to identify the haplotypes surrounding each gene. The number of haplotypes obtained with both distances for each of the four genes can be seen in Table 4. The sizes of the haplotypes differed among genes but were mostly similar for a given gene across the different genotypic datasets. No pair of tag SNPs flanking the *GmPhyA2* gene was found to be in disequilibrium (r^2_{vs} > 0.5) in either dataset and thus no haplotype could be described for this gene.

Table 4. Detailed results of the haplotyping approach applied to four maturity genes that used a maximal distance between tag single nucleotide polymorphisms (SNPs) of either 250 kb or 1 Mb. The table shows the number of retained tag SNPs (in linkage disequilibrium with at least one other tag SNP on the opposite side of the gene), the number of SNPs supporting these tag SNPs, the size of the resulting haplotype block (distance between the two farthest tag SNPs), the total number of distinct haplotypes for both genotyping datasets, and the number of rare haplotypes (frequency < 0.05).

Dataset	Gene	Distance	Tag SNPs	Supporting SNPs	Size of haplotype block	Haplotypes	Rare haplotypes < 5%
					bp		
SNP	*E1*	250 kb	3	12	209,669	4	2
array		1 Mb	3	12	209,669	4	2
	GmGia	250 kb	4	55	388,835	4	1
		1 Mb	7	70	1,223,653	9	6
	GmPhyA3	250 kb	6	17	243,307	7	4
		1 Mb	10	53	406,898	9	6
	GmPhyA2	250 kb	0	0	–	–	–
		1 Mb	0	0	–	–	–
GBS	*E1*	250 kb	2	37	157,787	3	1
		1 Mb	2	37	157,787	3	1
	GmGia	250 kb	2	6	24,084	2	0
		1 Mb	4	9	1,089,481	5	3
	GmPhyA3	250 kb	7	23	328,748	7	4
		1 Mb	7	23	328,748	7	4
	GmPhyA2	250 kb	0	0	–	–	–
		1 Mb	0	0	–	–	–

In some cases, increasing the maximum distance between tag SNP pairs increased the number of haplotypes obtained; in other cases, this had no impact. Indeed, the use of a larger maximum distance (1 Mb) between tag SNP pairs significantly increased both the number of distinct haplotypes and haplotype length for the *GmGia* gene in both datasets. A difference in haplotype length and the number of distinct haplotypes was also observed for the *GmPhyA3* gene but only in the simulated SNP array dataset. Usually, the use of a larger distance increased the number of distinct haplotypes by documenting new low-frequency haplotypes (minor allele frequency < 5%). Differences in the number of SNPs supporting the selected tag SNPs (i.e., all SNPs were in high LD with the tag SNPs) were observed between the GBS and SNP array datasets. This difference in SNP coverage was not found to correlate with the quality of the defined haplotypes (i.e., the correspondence between haplotypes and alleles).

Correspondence between Haplotypes and Alleles

For each gene, the haplotypes obtained using both the SNP array and GBS datasets from our approach were compared with the alleles known to be carried by each line according to the WGS data (Table 5, Supplemental Fig. S2). With the simulated SNP array dataset and by using both maximal distances (250 kb or 1 Mb) (Table 4),

Table 5. Correspondence between defined haplotypes and alleles known to be carried by 67 soybean lines at three maturity genes using a maximal distance of 250 kb in the single nucleotide polymorphism (SNP) array and genotyping-by-sequencing (GBS) datasets.

		SNP array		GBS	
Gene	Alleles	*N*	Haplotype	*N*	Haplotype
E1	*E1*	1	A	3	A†
	e1-fs	2	B		
	e1-as	25	C	25	B
	e1-nl	39	D	39	C
GmGia	*E2-in*	11	A	13	A
		2	B		
	e2-ns	49	C	54	B
		5	D		
GmPhyA3	*E3Ha*	30	A	30	A
	e3Mo	4	B†	4	B†
	e3-fs				
	E/e3p.Thr832Ala	2	C	2	C
	E3Mi	1	D	1	D
	e3-tr	26	E	26	E
		2	F	2	F
		2	G	2	G

† These haplotypes are those that contained more than one allele.

four haplotypes were identified for the *E1* gene, which corresponded perfectly with the four alleles known to be present in these lines. With the GBS data, only three haplotypes were found by using either maximal distances. Two of these haplotypes (B and C) coincided with the *e1-nl* and *e1-as* alleles, respectively. However, Haplotype A grouped three lines carrying either the *E1* or *e1-fs* alleles.

For the *GmGia* gene (two alleles), the use of a maximum distance of 250 kb resulted in the definition of four different haplotypes with the SNP array dataset. The lines known to carry *E2-in* and e*2-ns* were each found to be split into two haplotypes (one major and one minor). However, each of the two haplotypes corresponding to these alleles was unique to that allele. With the GBS dataset, two different haplotypes were obtained that corresponded perfectly with the two alleles (*E2-in* and *e2-ns*) present in this collection of lines. However, with both datasets, the use of the larger maximal distance (1 Mb) yielded a larger number of haplotypes, essentially by splitting haplotypes into several subhaplotypes. For example, with the SNP array dataset, a total of nine haplotypes was observed, six of which were present at a very low frequency (<5%). Thus a 1-Mb maximal distance did not allow for a better equivalence between haplotypes and alleles for this gene.

For the *GmPhyA3* gene (six alleles), a short distance of 250 kb allowed us to identify seven distinct haplotypes with both the SNP array and the GBS datasets. The results were highly similar for both datasets. Three haplotypes were specific to three of the alleles (*E3Ha*, *E/e3p.Thr832Ala*, and *E3Mi*), whereas three other haplotypes were found to describe individuals sharing the *e3-tr* allele. The seventh haplotype was shared among three lines carrying the rare alleles *e3Mo* and *e3-fs*. When the maximal distance was increased to 1 Mb, the same results were obtained with the GBS dataset, whereas the SNP array dataset resulted in the definition of additional haplotypes that essentially broke down allele-specific haplotypes into two subhaplotypes without improved resolution of the *e3-Mo* and *e3-fs* alleles. Again, expanding the maximal distance between the flanking markers to 1 Mb did not improve the correspondence between haplotypes and alleles but resulted in the creation of new very low frequency haplotypes.

In summary, this approach allowed us to define haplotypes at three of the four genes examined (*E1*, *GmGia*, and *GmPhyA3*) and to equate these haplotypes to alleles. Overall, 97.3% of the lines sharing the same haplotype were found to share the same allele at a locus. With the SNP array dataset, 10 of the 12 alleles known to be present in the dataset for those three genes were correctly identified, resulting in 98% (197 out of 201) of the alleles being correctly identified. For the GBS datasets, diagnostic haplotypes were found for 8 of the 12 alleles, with 97% (194 out of 201) of the alleles being correctly identified. Indeed, in the few cases where accessions with different alleles shared the same haplotype, these involved alleles that were present at low frequencies. Independent validation using a collection of unrelated lines (32 Plant Introduction lines) showed that with this validation dataset, the haplotypes and markers performed exactly the same way as with the collection of 67 Canadian lines, with the same difficulties in distinguishing the *e3Mo* and *e3-fs* alleles and, for the GBS dataset only, the *E1* and *e1-fs* alleles (Supplemental Table S2). Diagnostic markers identified in this study and their associated linked markers can be found in Supplemental Table S3. In the case of one gene (*GmPhyA2*), a lack of markers in high LD prevented the identification of alleles; this result was the same with both datasets and shows that for some genes, no matter which method is used, haplotyping cannot be used to derive diagnostic haplotypes.

DISCUSSION

A Systematic Approach for Defining Haplotypes

The selection of pairs of markers flanking the central position of a gene, in view of haplotyping, is structured into three steps and allows for the setting of four main parameters (Fig. 2). The three major steps are: (i) the estimation of disequilibrium, (ii) the identification of tag SNPs, and (iii) the selection of tag SNP pairs flanking the targeted gene. Our approach aims to allow an exploration of the haplotype diversity that is present in a collection for a given gene. Because it is centered on a gene, our strategy differs from previously described methods (the confidence intervals, the four-gamete rule, and the Solid Spine of LD), which tend to identify and describe blocks of markers on a large (chromosomal) scale (Gabriel et al., 2002; Wang et al., 2002; Barrett et al., 2005). Moreover, although the outputs from alternative approaches (which can be highly different from each other) are most often used for other purposes (large-scale LD analysis or genome-wide association studies), our gene-centric strategy specifically aims to

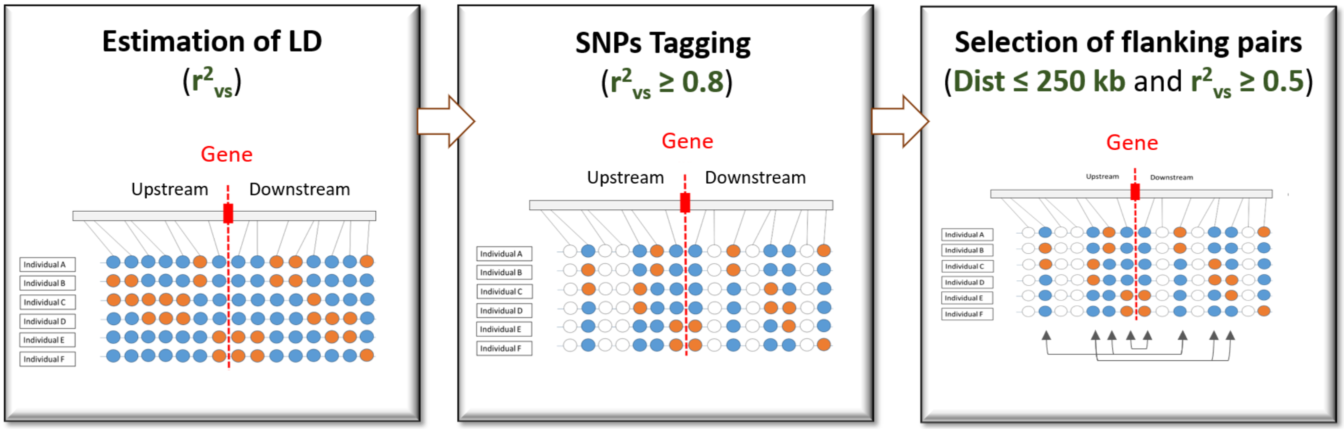

Fig. 2. Schematic illustration of the process used to identify a set of single nucleotide polymorphisms (SNPs) that can be used to define haplotypes near a gene of interest. Three main steps are used: (i) estimating linkage disequilibrium (LD) between SNPs by considering relatedness and population structure using the r^2 for relatedness and population structure (r^2_{vs}), (ii) clustering all SNPs located to one side of the gene that are in high LD with each other ($r^2_{vs} \geq 0.8$) and selection of tag SNPs, and (iii) identifying flanking tag SNPs above the chosen r^2_{vs} threshold and within a maximum distance.

identify and select a collection of SNPs that are useful for defining SNP haplotypes around a target gene.

In this work, we have demonstrated that the concept of selecting pairs of markers flanking a gene can lead to the accurate identification of haplotypes shared by individuals carrying the same allele. This strategy was largely successful in our collection of 67 inbred Canadian lines for three (*E1*, *GmGia*, and *GmPhyA3)* of the four genes studied via two different genotyping approaches, with an overall accuracy of 97.3%. Although there was only a small proportion of common markers between the GBS and simulated SNP array catalogs, our haplotype characterization results were equally successful between both datasets, indicating that our approach is genotyping-platform independent. Compared with the three standard haplotype block-based methods, HaplotypeMiner consistently performed similarly or better by decreasing the breakdown of individuals carrying the same allele into several haplotypes without functional relevance (see Supplemental File S2).

Parameter Settings

Several factors can affect the best settings for haplotype definition near a given gene; we have therefore allowed users to select the desired settings for some parameters: (i) the approach used to estimate LD (r^2, r^2_{vs}.), (ii) the threshold LD value for collapsing SNPs into a single tag SNP, (iii) the LD threshold, and (iv) the maximal distance between each member of a pair to be selected. Indeed, the species or the collection under study, the targeted gene, the local LD or density of SNPs, and the anticipated number of alleles at a gene are all factors that should be taken into consideration to produce the best haplotypes for the target region. During the analysis, results such as the size of the identified haplotypes, the number of haplotypes obtained compared with those expected (when known or suspected), and the number of very low frequency haplotypes can also guide the user in the adjustment of the parameters.

One of the customizable parameters in our approach is the selection of either r^2 or an alternative measure (such as r^2_{vs}, r^2_{v},or r^2_{s}) to select pairs of markers flanking a gene. In our case, allowing for relatedness and population structure in the estimation of LD (r^2_{vs}) was never found to reduce the effectiveness of haplotype definition; on the contrary, it allowed for a greater reduction in the number of tag pairs flanking a gene that were estimated to be in LD ($r^2_{vs} \geq 0.5$) (Fig. 1). We thus chose to use it by default. The impact of this correction on the selection of tag SNPs for haplotyping was found to vary depending on the targeted gene and the distance used [the shorter the distance, the less the impact of using r^2_{vs} was visible on the resulting haplotypes (data not shown)]. Prior to an analysis, the utility of the corrected measure of LD can be estimated by verifying the presence of unexpectedly high LD at distant loci when plotting the estimated LD between pairs of markers flanking the central position of the target gene as a function of the physical distance between the members of a pair of markers.

Collapsing SNPs in high LD to retain a single tag SNP by using a higher threshold (e.g., $r^2_{vs} = 1$) leads to no loss of information from the dataset but has the disadvantage of being sensitive to genotyping errors. On the contrary, a lower threshold, as was used here ($r^2_{vs} = 0.8$), will mask small genotyping errors or rare recombination events but may also lead to the loss of haplotypes specific to rare alleles. For example, in the GBS dataset and for the *E1* gene, the use of a LD threshold of $r^2_{vs} = 1$ resulted in four distinct haplotypes instead of three and thereby allowing us to distinguish the two rare alleles *E1* ($N = 1$) and *e1-fs* ($N = 2$) (data not shown). In this case only, the use of a higher threshold for defining tag SNPs was found to be valuable.

For selecting marker pairs, the higher the critical threshold of LD (r^2_{vs} threshold) or the smaller the distance between markers, the lower the risk of documenting an excessive number of haplotypes, but the risk of losing information is greater (the number of haplotypes may be lower than expected). Our results showed that in this collection, if a haplotype specific to one allele is not detected within a maximum distance of 250 kb, the expansion of the interval does not make it possible to improve the match between haplotypes and alleles. Indeed, compared with 1 Mb, a shorter distance of 250 kb did not cause any loss of information and, in some cases, did not change the number and size of the haplotypes. Usually, the use of a greater distance between flanking tag SNPs led to an increased number of haplotypes but, in general, through the addition of low frequency haplotypes (<5%) that had no functional relevance to the alleles under investigation.

Limiting Aspects and Considerations

We successfully identified haplotypes offering a very good match with the allelic status at three of the four genes under study. For the most part, rare alleles (such as *E1* and *e3-fs)* proved difficult to capture accurately through the identification of a corresponding haplotype. Because we only retained SNPs with a minor allele count of ≥4, obtaining markers specific to rare alleles can be difficult, whatever the haplotyping approach. Furthermore, concerning rare alleles, particular caution may be warranted if one uses GBS genotyping data, as this may affect the accuracy of the imputation step. The detection of haplotypes also depends on the ability to genotype diagnostic markers for a given allele correctly. For example, in the case of the *GmPhyA3* gene, if a marker specific to the *e3Mo* allele had been present, then *e3-fs* would not have been confused with *e3Mo*. In our genotyping datasets, these two alleles (*e3-fs* and *e3Mo*) shared a highly similar haplotype in the vicinity of the gene. No marker in LD with *e3Mo* was found in the SNP array data and only one (but distant) was found in the GBS dataset. Because of their high similarity, these two alleles were also misidentified in Tardivel et al. (2014). The failure to distinguish two alleles can also be explained by the history of the appearance of an allele. Indeed, if a mutation is recent, the SNP landscape around this gene will not have had time to diverge. Markers specific to a recent mutation will thus be rare and it will prove difficult to capture a distinct haplotype. In the case of *e3Mo*, only resequencing data allowed the identification of markers in proximity to and in LD with this allele (data not shown). Thus the ability to obtain haplotypes that group individuals sharing the same allele is dependent on the technical ability to genotype markers in disequilibrium with each of the alleles. The ability to capture informative markers in the vicinity of the gene is dependent on the density of genotyping. For this approach to work best, the genotyping should be sufficiently dense to deliver at least one marker on each side of the gene that is in LD with each allele present at a given gene.

Although we successfully identified haplotypes that group individuals sharing the same allele (one haplotype = one allele), all individuals sharing a given allele were not systematically pooled into one single haplotype (one allele ≠ one haplotype). Such alleles with multiple haplotypes were particularly frequent for *e3-tr* and *E2-in*. Indeed, one interesting aspect of our approach would be to document the presence of recombination events around a given allele. This information is not exhaustive and will vary according to the frequency of the parental haplotype and the recombinant haplotype. This information could be valuable in a breeding context (e.g., selection of parents) to minimize linkage drag with unfavorable linked loci. This aspect can be considered when setting parameters for haplotyping.

The inability to detect haplotypes for the *GmPhyA2* gene was intriguing, especially since both the GBS and SNP array were equally unsuccessful. On chromosome 20, no markers in LD with any of the three alleles were detected on both sides of the gene. A comparison with the resequencing data demonstrated that the inability to document markers in LD with the *E4* and *e4* (*SORE-1*) alleles was not only caused by a lack of SNP coverage but resulted from an unusual LD landscape around this gene (data not shown). Further studies will be needed to determine if the LD landscape around the *GmphyA2* gene is specific to this early-maturing Eastern Canadian collection. In the case where there are truly no markers in LD with the different alleles (recent mutations, double recombinations, and recombination hotspots), it will prove difficult, if not impossible, to identify informative haplotypes, no matter which approach is used.

Breeding aims to produce new lines that carry an array of alleles that jointly produce a superior phenotype. A key part of this work is to assess the allelic diversity present in germplasm collections and to identify individuals carrying favorable alleles at these genes. In this work, we demonstrate how our approach can provide accurate and essential information for breeding by delivering a quick and clear picture of the allelic diversity for a gene within a given germplasm collection. This approach was found to be highly accurate, with an overall success rate of 97.3% in terms of grouping individuals sharing the same allele into a shared haplotype, thus predicting the allele from the haplotype. This approach was also found to be reproducible on two distinct genotypic datasets (SNP array and GBS) with similar rates of success. Failures were also highly concordant and shared across datasets, showing that the observed limitations were attributable to the paucity of informative variants in the vicinity of one gene. Ultimately, our approach for identification of haplotypes from large SNP datasets represents a promising approach to routinely assess allelic variation in large collections.

Supplemental Information

Supplementary Table S1. Description of the various maturity genes and alleles studied and their expected position on genome *Gm*.a2.v1

Supplementary Table S2. Maturity gene alleles present in a collection of 32 Plant Introduction lines based on WGS and prediction of allelic status from SoySNP50k array and GBS genotypic data using the markers and haplotypes defined in the current study with a collection of 67 Canadian lines.

Supplementary Table S3. Description of the markers defining haplotypes and markers linked to them ($r^2_{vs} > 0.8$) for use in marker-assisted selection.

Supplementary Figure S1. Gene structure of the *GmPhyA2* gene with (A) the location of the *e4p.T832QfsX21* loss-of-function allele and the five other alleles reported to date, (B) alignment of the second exon sequence around the 1-bp deletion characteristic of *e4p.T832QfsX21*, and (C) the frameshift mutation specific to *e4p.T832QfsX21*, resulting in a premature stop codon

Supplementary Figure S2. Detailed haplotypes obtained for the *E1* (a and b), *GmGia* (c and d), and *GmPhyA3* (e and f) genes obtained with both the simulated SoySNP50K (a, c, and d) and GBS (b, d, and f) datasets.

Supplementary File 2. Comparison of Haplotype-Miner to other haplotyping approaches

Conflict of Interest Disclosure

The authors declare that there is no conflict of interest.

ACKNOWLEDGMENTS

This work was funded by The Canadian Field Crop Research Alliance and Agriculture and Agri-Food Canada under the Agri-Innovation Program, along with Genome Canada and Genome Québec funds awarded to the SoyaGen project. MAL is supported by a Natural Sciences and Engineering Research Council of Canada, Canada Vanier Graduate Scholarship and a scholarship from the AgroPhytoSciences NSERC CREATE Training Program. The authors thank Tanya Copley for critical review of the manuscript.

REFERENCES

Bandillo, N., D. Jarquin, Q. Song, R. Nelson, P. Cregan, J. Specht, et al. 2015. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome 8(3). doi:10.3835/plantgenome2015.04.0024

Barrett, J.C., B. Fry, J.D.N.J. Maller, and M.J. Daly. 2005. Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263–265. doi:10.1093/bioinformatics/bth457

Bastien, M., H. Sonah, and F. Belzile. 2014. Genome wide association mapping of *Sclerotinia sclerotiorum* resistance in soybean with a genotyping-by-sequencing approach. Plant Genome 7(1). doi:10.3835/plantgenome2013.10.0030

Browning, S.R., and B.L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81(5):1084–1097. doi:10.1086/521987

Contreras-Soto, R.I., F. Mora, M.A.R. de Oliveira, W. Higashi, C.A. Scapim, and I. Schuster. 2017. A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. PLoS One 12(2):e0171105. doi:10.1371/journal.pone.0171105

Copley, T.R., M.-O. Duceppe, and L.S. O'Donoughue. 2018. Identification of novel loci associated with maturity and yield traits in early maturity soybean plant introduction lines. BMC Genomics 19(1):167. doi:10.1186/s12864-018-4558-4.

Desrousseaux, D., F. Sandron, A. Siberchicot, C. Cierco-Ayrolles, and B. Mangin. 2017. LDcorSV: Linkage disequilibrium corrected by the structure and the relatedness. R package version 1.3.2. R Foundation for Statistical Computing. https://CRAN.R-project.org/package=LDcorSV (accessed 13 June 2019).

Endelman, J.B., and J.L. Jannink. 2012. Shrinkage estimation of the realized relationship matrix. G3 (Bethesda) 2(11):1405–1413. doi:10.1534/g3.112.004259

Gabriel, S.B., S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, et al. 2002. The structure of haplotype blocks in the human genome. Science 296(5576):2225–2229. doi:10.1126/science.1069424

Hwang, E. Y., Q. Song, G. Jia, J. E. Specht, D. L. Hyten, J. Costa, and P.B. Cregan. 2014. A genome-wide association study of seed protein and oil content in soybean. BMC Genomics 15(1):1. doi:10.1186/1471-2164-15-1

Hyten, D.L., I.Y. Choi, Q. Song, R.C. Shoemaker, R.L. Nelson, J.M. Costa, et al. 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175(4):1937–1944. doi:10.1534/genetics.106.069740

Koch, E., M. Ristroph, and M. Kirkpatrick. 2013. Long range linkage disequilibrium across the human genome. PLoS One 8(12):e80754. doi:10.1371/journal.pone.0080754

Liu, B., A. Kanazawa, H. Matsumura, R. Takahashi, K. Harada, and J. Abe. 2008. Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. Genetics 180(2):995–1007. doi:10.1534/genetics.108.092742

Mangin, B., A. Siberchicot, S. Nicolas, A. Doligez, P. This, and C. Cierco-Ayrolles. 2012. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity 108(3):285–291. doi:10.1038/hdy.2011.73

Myers, S., L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310(5746):321–324. doi:10.1126/science.1117196

Myers, S., C. Freeman, A. Auton, P. Donnelly, and G. McVean. 2008. A common sequence motif associated with recombination hot spots and genome instability in humans. Nat. Genet. 40(9):1124–1129. doi:10.1038/ng.213

Narvel, J. M., D.R. Walker, B.G. Hector, J.N. All, W.A. Parrott, et al. 2001. A retrospective DNA marker assessment of the development of insect resistant soybean. Crop Sci. 41(6):1931–1939. doi:10.2135/cropsci2001.1931

R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/ (accessed 13 June 2019).

Raj, A., M. Stephens, and J.K. Pritchard. 2014. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. Genetics 197(2):573–589. doi:10.1534/genetics.114.164350

Sabeti, P.C., D.E. Reich, J.M. Higgins, H.Z. Levine, D.J. Richter, S.F. Schaffner, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832–837. doi:10.1038/nature01140

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463(7278):178–183. doi:10.1038/nature08670

Slatkin, M. 2008. Linkage disequilibrium understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. 9(6):477–485. doi:10.1038/nrg2361

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, et al. 2013. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One 8(1):e54985. doi:10.1371/journal.pone.0054985

Song, Q., J. Jenkins, G. Jia, D.L. Hyten, V. Pantalone, S.A. Jackson, et al. 2016. Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. BMC Genomics 17:33. doi:10.1186/s12864-015-2344-0

Tardivel, A., H. Sonah, F. Belzile, and L.S. O'Donoughue. 2014. Rapid identification of alleles at the soybean maturity gene E3 using genotyping by sequencing and a haplotype-based approach. Plant Genome 7(2). doi:10.3835/plantgenome2013.10.0034

Torkamaneh, D., and F. Belzile. 2015. Scanning and filling: Ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing data. PLoS One 10(7):e0131533. doi:10.1371/journal.pone.0131533

Torkamaneh, D., J. Laroche, M. Bastien, A. Abed, and F. Belzile. 2017. Fast-GBS: A new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics 18(1):5. doi:10.1186/s12859-016-1431-9

Torkamaneh, D., J. Laroche, A. Tardivel, L. O'Donoughue, E. Cober, I. Rajcan, et al. 2018. Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean. Plant Biotechnol. J. 16(3):749–759. doi:10.1111/pbi.12825

Torkamaneh D., J. Laroche, B. Valliyoda, L. O'Donoughue, E. Cober, I. Rajcan, et al. 2019. Soybean haplotype map (GmHapMap): A universal resource for soybean translational and functional genomics. BioRxiv. doi:10.1101/534578

Wang, N., J.M. Akey, K. Zhang, R. Chakraborty, and L. Jin. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. Am. J. Hum. Genet. 71(5):1227–1234. doi:10.1086/344398

Watanabe, S., R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, Y. Nakamoto, et al. 2009. Map-based cloning of the gene associated with the soybean maturity locus *E3*. Genetics 182:1251–1262. doi:10.1534/genetics.108.098772

Xia, Z., S. Watanabe, T. Yamada, Y. Tsubokura, H. Nakashima, H. Zhai, et al. 2012. Positional cloning and characterization reveal the molecular basis for soybean maturity locus *E1* that regulates photoperiodic flowering. Proc. Natl. Acad. Sci. USA 109:E2155–E2164.

Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu, W. Li, et al. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. 33(4):408–414. doi:10.1038/nbt.3096

Zondervan, K.T., and L.R. Cardon. 2004. The complex interplay among factors that influence allelic association. Nat. Rev. Genet. 5(2):89–100. doi:10.1038/nrg1270