ABSTRACT

Title of Document: ASSOCIATION ANALYSIS IN SOYBEAN

Eun-Young Hwang, Doctor of Philosophy, 2008

Directed By: Associate Professor Jose M. Costa, Department of Plant Science and Landscape Architecture

Association analysis is a new approach to identify the location of gene(s)/allele(s) of interest. There are a number of factors determining the feasibility of whole-genome association analysis which include the level of linkage disequilibrium (LD) and the magnitude of population structure in a population. The goal of this study was to evaluate the success of whole-genome association analysis in soybean germplasm accessions using DNA markers across the whole genome. Firstly, the extent of LD and the presence of population structure were estimated. Secondly, whole-genome association analysis was performed to detect the location of the allele/gene controlling flower color, pubescence color, and seed protein quantitative trait loci (QTLs) in 319 soybean [*Glycine max* (L.) Merr.] germplasm accessions. The soybean germplasm accessions had a relatively low level of LD which declined very rapidly to 0.8 in less than 4 Kbp as indicated by r^2 as well as highly diverse population structure. Despite the low LD and the presence of high population structure, whole-genome case-control analysis successfully detected the

65 bp insertion in the *GmF3'5'H* (GenBank acc. AY117551) gene controlling purple *vs*. white flower color, as well as a single base deletion in the *F3'H* (GenBank acc. AB191404) gene controlling tawny *vs*. gray pubescence color. However, there were 28 gray pubescence lines that did not contain the deletion suggesting that there is a second mutation determining the pubescence color alteration. In the case of seed protein QTL, whole-genome regression analysis detected one of four previously reported seed protein QTLs which reside on linkage group (LG) E and new seed protein QTL on LG K. The detection of three other previously reported seed protein QTLs on LGs A1, I and M was not successful. It is unclear why association analysis was not successful in the detection of the three previously reported QTLs. However, a number of reasons including incomplete adjustment for population structure, lack of statistical power, an inadequate number of genetic markers in light of the low level of LD, and the power of association analysis to detect alleles with relatively modest genetic effects are suggested as possible reasons.

ASSOCIATION ANALYSIS IN SOYBEAN.

By

Eun-Young Hwang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:
Associate Professor Jose M. Costa, Chair
Adjunct Professor Perry B. Cregan
Professor Heven Sze
Professor William J. Kenworthy
Professor Marla S. McIntosh

Dedication

I would like to dedicate this dissertation to my parent, Doho Hwang and Leadong Lee, for giving me endless love, reliance, and encouragement through my whole life and to my husband, Daeho Jin, for supporting me with infinite love.

Acknowledgements

I would like to thank everybody who has been helping me to complete my Doctor of Philosophy degree in the department of Plant Science and Landscape Architecture. I thank and appreciate Dr. Perry B. Cregan for leading me on the journey to be a soybean geneticist and for being a great advisor through out the time required to complete this research and to write the dissertation. Moreover, he has demonstrated a great attitude as a scientist and unlimited generosity as a person. I would like to thank my advisor Dr. Jose M. Costa who has given me much thoughtful academic advice and who has always been on my side. I also thank Dr. William J. Kenworthy for being a part of my dissertation committee and allowing me to use the nitrogen analyzer for the analysis of germplasm accessions used for my dissertation. Without his help I should not have finished the analysis and would not have finished my research in a timely fashion. I thank Dr. Marla S. McIntosh for encouraging me and showing me how to lead a life as a women scientist. I also thank Dr. Heven Sze who has given her time as a member of my dissertation committee. I am grateful to all my colleagues, Dr. Qijiang Song, Dr. David L. Hyten, Mrs. Ronghui Yang, Mr. Edward Fickus, and Mr. Chuck Quigley. They assisted, advised, and supported my research and my life, and entertained me over the many years of friendship as well as taught me what is important in life and how to deal with reality.

Table of Contents

List of Tables

List of Tables

Chapter 1: Literature Review

Introduction

Much research has been conducted by plant molecular geneticists to identify the genome positions, and in some instances to isolate genes or quantitative trait loci (QTLs) that underlie traits of interest. The genome positions of genes responsible for qualitative traits such as flower color, pubescence color or some disease resistances are relatively easier to identify than the genes controlling quantitative traits. Many important agricultural traits such as seed chemical composition, seed weight, seed size, or yield etc. are expressed as quantitative traits. Quantitative traits show continuous variation, are controlled by multiple genes each segregating in a Mendelian fashion, are easily influenced by environmental factors, and may also be affected by the interaction between genes and the environment. The study of these complex traits generally requires statistical approaches. Conventional linkage analysis or QTL analysis is based upon the recombination frequency between a genetically controlled phenotype and DNA markers. The performance of such an analysis is dependent upon a pre-existing molecular genetic map of the organism under study in which DNA markers are positioned at defined intervals.

The most commonly used DNA markers including, RFLP (restriction fragment length polymorphism), RAPD (random amplified polymorphic DNA), microsatellites or simple sequence repeats (SSRs), single nucleotide polymorphisms (SNPs), and other available DNA markers are used to identify the location of genes/QTLs responsible for natural variation in structured populations. Structured populations such as recombinant inbred lines (RILs), near-isogenic lines (NILs), or backcross-derived populations are generally developed by crossing individuals with extreme values for a trait of interest i.e., high *versus* low seed oil or protein concentration or resistance *versus* susceptibility to pathogens or insects, etc. The individuals in the structured population have extensive linkage disequilibrium (LD) that is, the non-random association of alleles at different genetic loci in a population. Extensive LD results from the limited number of recombinations that occur between homologous chromosome pairs over the small number of generations required for population development. As a result, in a conventional linkage analysis, markers separated from the gene of interest by as much as 10-20 cM remain associated (in linkage disequilibrium) with that gene. This is why the resolution of conventional linkage analysis is relatively low in terms of defining the genome position of a causative gene/QTL. In addition, the size limitation of structured populations and occasionally inaccurate phenotypic data can make it difficult to define the location of genes or QTLs with small and moderate effects. Thus, if one desires to clone a gene/QTL controlling a trait of interest, conventional linkage analysis would be only the first step in the process. Despite the low resolution of gene mapping, many genes whose locations were identified via conventional genome-wide linkage analysis have been successfully cloned such as *fw2.2* (fruit size QTL) (Frary et al., 2000), and *OVATE* (fruit shape QTL) (Liu et al., 2002) in tomato (*Lycopersicon esculentum* L.), *Hd1* (heading time QTL) (Yano et al., 2000) in rice (Oryza sativa L.), and *tb1* (branch architecture QTL) (Doebley et al., 1997) in maize (*Zea mays* L.). Some of these genes have major genetic effects for the traits studied and others are qualitative in nature and control all phenotypic variation of the trait in question. However, despite these positive results, it remains difficult to clone the genes responsible for quantitative variation with moderate or small effect.

Association analysis is an alternative approach for gene or QTL discovery which, like linkage analysis, is also based upon the association of phenotypic variation with allelic variation of DNA marker loci. However, naturally existing populations such as germplasm collections or human populations can be used in an association analysis. The individuals included in such populations do not necessarily need to have defined genetic relationships such as would be the case in structured populations of RILs, NILs, or backcross-derived genotypes. Naturally occurring populations generally have much less extensive LD than structured populations since these populations take advantage of historical recombination over thousands of generations. Therefore, the size of the intact linkage block adjacent to a gene of interest is likely to be very small which results in high resolution in the positioning of genes/QTLs using association analysis. A case-control study is the simplest form of association analysis. In a case-control study, marker allele frequencies between case (patients with a specific disease or individuals with a certain trait) and control groups are compared. Case-control studies are designed such that the case and control groups are similar in every way except for the trait under examination. For example, in human case-control studies of a particular disease, the individuals making up the two groups would be similar in terms of age, sex, ethnicity, medical history etc. Then, the markers for which there were significant allele frequency differences between the case and control groups are considered to be associated with the gene or genes conditioning disease occurrence. One of the difficulties encountered in a case-control association study is “population structure” which is not a factor in conventional QTL analysis. Population structure is the result of allele frequency differences between different populations because of their genetic background differences which may be unrelated to the distinguishing phenotypic difference(s) being studied. The presence of population structure causes spurious associations between marker loci and the phenotypic trait under study. Thus, the selection of balanced individuals with similar genetic backgrounds is very important in association analysis.

Soybean is a one of the major crops grown in the United States and has a long history of development of new cultivars with superior agricultural traits such as seed with high nutritional value, high yield, or resistance to biotic and abiotic stress. Soybean germplasm collections provide a good resource of genetic materials to develop new cultivars because they contain a large pool of genetic diversity. Exotic germplasm from germplasm collections is generally employed in breeding programs based upon the presence of a desirable phenotypic trait such as disease resistance, seed composition, or a morphological trait of interest. There is sometimes little knowledge of the genetic control of the phenotype and in most cases no knowledge of the genome position of genes that control the traits. Association analysis would be a useful tool for the determination of the genome position of the gene(s) that controls a trait of interest in germplasm collections prior to the use of this germplasm in a breeding program.

Genetic Linkage Maps

The construction of a high-density genetic map that covers the whole genome of the species being analyzed is required for the implementation of whole-genome association analysis. Single nucleotide polymorphisms (SNPs) include single base differences and small insertions and deletions (indels) between two homologous DNA sequences or chromosomes. SNPs are the most abundant type of DNA polymorphism with extremely low mutation rates as compared to other genetic markers such as microsatellites (also referred to a simple sequence repeats or SSRs). The average rate of mutation of single DNA nucleotides (referred to as nucleotide substitutions) in higher plants is reported to be 6.0×10^{-9} per site per year (Wolfe et al., 1987) and 3 to 5×10^{-9} per site per year in animals (Kondrashov and Crow, 1993). In contrast, the average mutation rate of SSRs is 10^{-3} in humans (Xu et al., 2000), 7.7×10^{-4} in maize (Vigouroux et al., 2002), and 2×10^{-4} in soybean (Diwan and Cregan, 1997). The much lower mutation rate makes SNPs the appropriate choice for association analysis. In addition, the availability of high throughput SNP discovery and genotyping methods make the construction of high marker-density SNP-based genetic maps possible in a relatively short time. Current sequencing technology such as the Illumina's genome analyzer (Illumina Inc.) will permit the accumulation of more than a billion base pairs (bp) of sequence data per run, thus facilitating rapid SNP discovery. Likewise available SNP analysis approaches such as the Illumina Infinium assay have the capacity to genotype up to 295,000 data points in a 72 hour period (Fan et al., 2003; Hyten et al., 2008).

Currently, the 2008 Soybean Consensus Map contains a total of 5,503 genetic markers including 1,017 SSRs, 3,792 SNPs, and 694 other types of markers and almost all genetic markers have been associated with sequence scaffolds in the Williams 82 whole-genome sequence that was recently released by the U.S. Department of Energy, Joint Genome Institute, Walnut Creek, CA (http://www.phytozome.net/soybean). The availability of the genome sequence will accelerate the discovery of SNP markers for the construction of high-resolution genetic maps and for use in association analysis.

Association Analysis

Two different types of association analysis are commonly employed. The first is whole-genome association analysis while the second is the candidate gene approach. These approaches are applied depending on the extent of LD in a population under study and the density of the available markers on the genetic map. Whole-genome association analysis is applicable in populations with relatively high LD and/or with a sufficient number of markers that explain most of the genetic variability in the population under study. For populations with relatively low LD and when there is knowledge of the approximate genome position of the target gene, a candidate gene approach can be used to obtain a precise estimate of the position of the gene that controls the phenotypic difference in the trait. There are two ways to identify a gene of interest using association analysis; either direct or indirect methods. The direct method is the detection of the causative mutation that actually creates the phenotypic variation. In contrast, in the case of the indirect method, the causative mutation residing in the same linkage block or in LD with a marker locus is used to identify the approximate position of the causal mutation. The simplest form of association analysis is a case-control study. In human genetics, the case-control study has been used to find genes or variants causing disease susceptibility. The case group has been diagnosed with a specific disease and the control group is otherwise similar in terms of age, sex, ethnicity etc. but is disease free. In the case-control study, marker allele frequency differences between the two groups of individuals are compared using molecular markers throughout the genome or genomic regions in the vicinity of candidate genes. Markers with significant allele frequency differences between the case and control groups are putatively associated with the locus or loci that condition disease susceptibility. The other form of association analysis is based upon regression analysis. The regression model is often used when a trait under study has continuous variation. Regression analysis incorporates each marker in the model-based analysis to calculate the relationship between the allele frequency of markers and the phenotypic variation. Both approaches are applicable to whole-genome and to candidate gene association analyses.

In the past several years, a number of human genetics research teams have published reports of genes associated with disease occurrence using case-control studies in thousands of individuals with hundreds of thousands of markers. This research has been made possible by the large investments from the Human Genome Project (Sachidanandam et al., 2001) and the International HapMap Project (2003). These two large projects provided human genome sequence variation including more than 4.5 million SNPs through the Phase I (Sachidanandam et al., 2001) and Phase II (Frazer et al., 2007) HapMap projects, copy number variation (Redon et al., 2006), large scale indel polymorphism (Conrad et al., 2006), and all classes of structural variants (Khaja et al., 2006), all of which are publicly available. In addition, large reservoirs of genetic information on individuals with different disease symptoms at many different research centers and hospitals also assisted in the detection of genes associated with diseases. The extensive genome-wide variant database, the availability of high-throughput sequencing and genotyping technologies, and the clinical samples for which phenotypes are already determined, facilitated the founding of the Wellcome Trust Case Control Consortium. The Welcome Trust Case Control Consortium brought together more than 50 research groups to detect SNPs more likely to be associated with disease susceptibility using whole-genome case-control analysis. The products of all these efforts made genome-wide association analysis possible and facilitated genome scans of 17,000 individuals for seven diseases (Burton et al., 2007). Owing to these benefits, most of the recent case-control studies in humans have used more than several hundred thousand SNPs analyzed on thousands of individuals.

Rioux *et al*. (2007) performed genome-wide case-control association analysis to find loci associated with susceptibility to Crohn's disease. The first experiment was performed using 304,413 SNPs in 946 cases and 977 controls and detected significant associations with 27 SNPs from nine different genomic regions. The most significant 23 were analyzed in two unrelated populations, one that consisted of 530 father-mother child trios where the child was diagnosed with Crohn's disease and the other study with 353 case and 207 control individuals. Only five of the 23 SNPs were significantly associated with disease incidence in these studies. The results of the two association mapping studies and the one family-based linkage analysis identified two previously reported loci and three additional new loci associated with Crohn's disease. In the case of another disease, Type 2 diabetes, the occurrence of the disease is the consequence of the interactions of many genetic and environmental factors. Sladek *et al.* (2007) conducted a case-control study to find loci associated with susceptibility to Type 2 diabetes. They used a total of 392,935 SNPs to scan whole-genomes of the first case-control group, 661 individuals with Type 2 diabetes and 614 control individuals. They found 66 unique SNPs significantly associated with Type 2 diabetes from 44 unique loci. One of these loci was a previously reported locus and the others were newly detected loci. The most significant 57 loci were used to analyze 2,617 case and 2,894 contol individuals to test the fidelity of the results. In the second case-control study, only eight SNPs showed significant associations with Type 2 diabetes. The genome region of four of the eight SNPs had been previously discovered. Eeles *et al.* (2008) used whole-genome case-control analysis to find the location of common alleles causing prostate cancer using 541,129 SNP loci. The reliability of associations between SNP markers and the occurrence of prostate cancer was verified in two case-control groups, one with 1,854 cases and 1,894 controls and the other with 3,268 cases and 3,366 controls. The results of association analysis detected the previously reported common loci at 8q24 and 17q and three new loci. This result supported the conclusion that the genome regions 8q24 and 17q are the regions with common alleles associated with prostate cancer susceptibility.

The first reported association analysis in plants was conducted by Thornsberry *et al*. (2001) who detected nine SNPs from the *Dwarf8* locus in 92 maize inbred lines that were associated with a 7 to 11 day reduction in flowering time. Of these SNPs, two amino acid deletions were found that apparently were responsible for the early flowering phenotype. Other research has detected alleles or genes associated with flowering time in *Arabidopsis*. Stinchcombe *et al*. (2004) found that the alleles from the *FRI* region were associated with flowering time in 70 Northern European and Mediterranean ecotypes. In addition, Olsen *et al*. (2004) identified 103 SNPs in the *CRY2* locus using 95 *Arabidopsis* accessions. Ninety of the 103 SNPs were associated with the reduction of flowering time under short days. In maize, Palaisa *et al*. (2004; 2003) sequenced 10 low copy regions surrounding the *Y1* gene using 75 inbred lines including 3 orange endosperm lines, 38 yellow endosperm lines, and 34 white endosperm lines and detected a total of 168 SNPs associated with yellow endosperm color in 10 genomic regions. They found that SNPs located 600 Kbp upstream of *Y1* showed strong associations with the yellow endosperm phenotype. In this *Y1* region, the yellow endosperm lines showed higher LD than that of the white endosperm lines which also supported the conclusion that the yellow endosperm lines had undergone strong selection by maize breeders. In another association analysis in maize, Wilson *et al*. (2004) analyzed SNPs involved in kernel starch biosynthesis from six candidate genes: *amylose extender1*, *brittle endosperm2*, *shrunken1*, *shrunken2*, *sugary1*, and *waxy1* in 102 diverse maize inbred lines. They found that SNPs from the *bt2*, *sh1*, and *sh2* genes showed significant associations with the amount of starch in kernels.

Linkage Disequilibrium (LD)

LD is the nonrandom association of alleles at different loci in a population. LD is determined with frequency differences of alleles between two loci. Specific alleles at two loci in high LD are more likely to co-segregate than when in low LD. The extent of LD reflects the history of a population, for instance, recombination, mutation and out-crossing are factors that decrease LD (Flint-Garcia et al., 2003; Gaut and Long, 2003). Selfing, selection, population bottlenecks, and population admixture increase LD (Flint-Garcia et al., 2003; Gaut and Long, 2003; Gupta et al., 2005; Tishkoff and Verrelli, 2003). LD can be estimated by different statistics including r^2 (correlation based on LD) and Lewontin's *D'* (standardized disequilibrium coefficient). The value of r^2 ranges from 0 (no correlation or LD between alleles at two loci) to 1 (complete LD between two loci). The value of *D'* also varies from 0 to 1 but the *D'* between two loci is never less than 1 unless recombination between the loci is observed as indicated by the presence of four haplotypes (two parental and two recombinants) in the population. In the case of r^2, a value of less than 1 is not necessarily indicative of recombination between two loci because mutation can create a third haplotype without recombination. *D*' and r^2 provide measures that allow one to quantify LD. In particular, r^2 is the most useful measure of LD in assessing the density of markers required for a successful association analysis (Ardlie et al., 2002).

The HapMap project (The International HapMap Consortium, 2003) (http://www.hapmap.org/) was initiated to determine genome-wide variation in the human genome and the structure of haplotype blocks in populations from Africa, Europe, and Asia. A haplotype is the combination of particular alleles at physically linked loci which co-segregate as a block in a population (Cardon and Abecasis, 2003). The basic concept of the International Human HapMap project is that a common disease is caused by a common mutation. Thus, if all common haplotypes in the human genome were detected by the investigation of genome-wide SNPs, one could identify a haplotype associated with a common disease and eventually isolate the mutation causing disease susceptibility using LD between alleles at nearby loci.

A large-scale experiment to provide an estimate of LD in humans was conducted by Reich *et al*. (2001). The authors compared the "half-wide *D'* value (*D'*=0.5)" using 272 high allele frequency SNPs selected from 3,000 core SNPs spanning 19 genomic regions in 48 southern Swedes as a north-European sample and 96 Yorubans from Nigeria. The average half-wide *D'* of the north-European sample extended to 60 Kbp while it was less than 6 Kbp in the Yoruban population. Gabriel *et al*. (2002) reported the structure of 928 haplotype blocks selected from 54 genomic regions spanning 13.4 Mbp. They used a total of 3,738 SNPs in 275 individuals including 30 parent-offspring trios from Nigeria (Yoruban), 93 members of 12 multigenerational pedigrees of European ancestry, 42 unrelated individuals with Japanese and Chinese origin, and 50 unrelated African-Americans. The size of 928 haplotype blocks was estimated by confidence bounds on *D'*. In the Yoruban and African-American populations the average minimum size of haplotype blocks was 9 Kbp while the average size was 18 Kbp in the European and Asian populations. Hinds *et al*. (2005) also estimated the size of haplotype blocks using the HAP program with 1.6 million SNPs in 71 individuals including 24 European Americans, 23 African Americans, and 24 HanChinese. The average size of the haplotype blocks was 8.8 Kbp in the African-American population, 20.7 Kbp in the European-American population, and 25.2 Kbp in the HanChinese population. All of these results demonstrated that the European and Asian populations have experienced a severe population bottleneck and founder effects after they diverged from the African population.

Plant geneticists have also investigated the level of LD in the model species *Arabidopsis thaliana* (Kim et al., 2007) as well as in crop plants such as maize (*Zea mays* L.) (Jung et al., 2004; Remington et al., 2001; Wilson et al., 2004), soybean (*Glycine max* (L.) Merr.) (Hyten et al., 2007; Zhu et al., 2003), barley (*Hordeum vulgare* L.) (Caldwell et al., 2006; Rostoks et al., 2006), and potato (*Solanum tuberosum* L.) (Simko et al., 2006). In *Arabidopsis*, Kim *et al*. (2007) reported the level of LD based on the data from 341,602 SNPs across the whole genome in 19 accessions and found that average LD extended to 10 Kbp as indicated by r^2=0.2. In maize, an out-crossing species, extensive research has examined the extent of LD in many genes and different sets of genotypes. Tenaillon *et al*. (2002) investigated sequence diversity in 21 loci located on chromosome 1 in 35 individuals including 16 exotic landraces and nine U.S. inbred lines and found that intragenic LD declined within 100-200 bp. Remington *et al*. (2001) and Wilson *et al*. (2004) examined LD decline in ten different genes in maize inbred lines. Both studies examined more than 100 inbred lines and found that the extent of LD varied greatly depending on the genes under study. LD in *sh1, tb1, d3, and wx1* decayed within 0.5 Kbp (r^2=0.2) while in *su1* and *ae1* it extended to over 9 Kbp (r^2=0.2). All other genes had intermediate levels of LD. Extensive LD was reported in the maize *adh1* gene (Jung et al., 2004). Jung *et al*. (2004) examined LD in the *adh1* as well as in 14 loci in the vicinity of *adh1* using 192 inbred lines and found that LD around the *adh1* locus extended over 500 Kbp. In barley, a selfing species, Rostoks *et al*. (2006) used 612 SNPs throughout the genome in 91 European germplasm lines including 53 spring two-row and 38 winter lines to examine the level of LD. In winter barley, LD extended over 60 cM ($r^2 > 0.5$) in contrast to only 15 cM ($r^2 > 0.5$) in spring two-row barley. Caldwell *et al*. (2006) also examined four gene regions surrounding the hardness (*Ha*) locus, *hina*, *hinb*, *GSP*, and *PG2* loci spanning 215 Kbp in 74 cultivars, 23 landraces, and 34 wild barley lines. There was variability in the extent of LD depending on the gene and the genotype being analyzed. LD extended less than 1 Kbp in the *hina* gene while LD extended across the entire *hinb* gene spanning 3,000 Kbp. In addition, the average extent of LD in cultivated and landrace barley was similar while that of wild barley was much lower. Simko *et al*. (2006) determined the pattern of LD in potato which is an out-crossing autotetraploid species. The authors selected 67 DNA fragments spanning over 25 Kbp to estimate LD in 47 accessions, including 1 monoploid, 17 diploid, and 29 tetraploid accessions. The extent of LD was less than 1 Kbp (r^2=0.2) as would be anticipated in an out-crossing species. In soybean, Zhu *et al*. (2003) examined seven loci in a 12.5 cM region on linkage group (LG) G in 16 genotypes that showed extensive LD which decayed at a distance of 2.0-2.5 cM (r^2<0.1). Hyten *et al*. (2007) investigated the patterns of LD of 74 loci in three chromosomal regions on LGs A2, G, and J in four different populations including wild soybean, *Glycine soja* (Sieb. & Zucc.); Asian landraces, the ancestors of North American soybeans and elite North American cultivars. *Glycine soja* had much less extensive LD, which decayed to $r^2 \leq 0.1$ in 36 to 77 Kbp, than that in the three cultivated soybean populations. The LD level of the landraces, the elite North American cultivars, and the elite cultivars never decreased to $r^2=0.1$ in over 400 Kbp on LG A2 or G while on LG J it extended to 600 Kbp in the elite cultivars, 250 Kbp in the North American ancestors, and less than 100 Kbp in the landraces. The maximum difference in the extent of LD among species is approximately 1,200 fold from 0.5 Kbp in maize (Remington et al., 2001; Wilson et al., 2004) to 600 Kbp in soybean (Hyten et al., 2007) which may be partially explained by the different mating systems of the two species. Therefore, an understanding of the level of LD in a species may help to understand the population history of a species and will assist in determining how best to apply association analysis.

Population Structure

In addition to the extent of LD there are other factors influencing the success of association analysis including the presence and extent of population structure. Population structure is the allelic frequency differences between different subpopulations within a population that are caused by differences in ancestry among the subpopulations. Most species are not homogenously distributed but are subdivided into subpopulations that went through local colonization or geographical adaptation dependent on local conditions. Population subdivision resulting from local adaptation can cause changes of allele frequencies among subpopulations by natural selection, random genetic drift, or gene flow (Mitchell-Olds et al., 2007). The presence of population structure can cause spurious associations in association studies resulting from allele frequency differences between subpopulations that are not genetically associated with the phenotypic differences under study. Therefore, in planning association studies, to the greatest extent possible, it is important to select individuals with similar genetic backgrounds. It is also necessary to determine the existence of population structure and to know which statistical analysis is appropriate to adjust for population structure to reduce spurious associations as well as to increase the likelihood of detecting real associations.

There are several statistical analyses that have been developed to estimate the number of sub-groups in a population and to adjust for population structure to permit the identification of true associations. Devlin and Roeder (1999) introduced "genomic control" to adjust for population structure in which unlinked random markers are used to estimate the existence of population structure and apply an adjusted P-value to all loci to test for significant associations of markers with phenotypic variation. However, because population structure affects different regions of the genome differently, a single adjustment of significance level throughout the genome is probably not appropriate. To improve the genomic control procedure, Pritchard *et al*. (2000b) developed the STRUCTURE algorithm to detect the most likely number of sub-groups, 'K', that are present in a population. The subsequent statistical analysis of association is applied within each sub-population where allele frequencies are homogeneous, thereby facilitating a robust association analysis. However, this model is useful for the adjustment of admixture populations which are derived from individuals from different homogeneous populations such as in human populations. This adjustment method is not as useful for other complex populations such as plant populations which can have unknown ancestral relationships.

Thornsberry *et al*. (2001) used the STRUCTURE algorithm to estimate the number of sub-groups in 92 diverse inbred maize lines and the highest likelihood of sub-groups was expected when K=3. For those working in soybean, *Arabidopsis* populations may provide useful insights due to its selfing nature, its geographical distribution and local adaptation depending on climate and photoperiod response which are relatively comparable to soybean. Aranzana *et al*. (2005) verified the location of four previously discovered genes using SNPs at 100 Kbp intervals across the entire genome and additional SNP markers around four candidate genes in 96 ecotypes of *Arabidopsis* using association analysis. As would be expected, population structure was present in the *Arabidopsis* population which was causing spurious associations. They effectively reduced the rate of spurious association not only for qualitative traits but also quantitative traits using the STRUCTURE algorithm (Falush et al., 2003). In 2006, Yu and Buckler suggested a mixed-linear-model approach to account for population structure (Q model) as well as familial relatedness (K model) because, as the size of the population under study increases, population history is more complex, therefore, many sub-groups with different degrees of relatedness may be present. Yu *et al*. (2006) showed that the adjustment with the K model successfully controlled Type I error in human populations. In a maize population, the Q plus K mixed model most successfully controlled Type I error, although there was some variation depending on the characteristics of the genes being studied. Zhao *et al*. (2007) adopted the same mixed-linear-model as Yu *et al*.

(2006) to adjust for population structure in *Arabidopsis*. While Yu *et al*. (2006) used individual SNPs to estimate relative kinship (K matrix), Zhao *et al*. (2007) used haplotypes for the estimation of the K matrix and a principal components analysis (PCA) to estimate the Q matrix. These results suggested that haplotypes were more useful for determining the K matrix than was the use of single SNP alleles. Additionally, PCA provided a superior way to estimate the Q matrix than was the case with the STRUCTURE program. After the adjustment, they could eliminate many spurious associations and detected previously reported QTLs plus several new QTLs. These results indicated that although several statistical analyses were developed to adjust for population structure, those associations that remained after the adjustment might not necessarily be true associations. Therefore, for a successful association analysis, it is crucial to select the individuals in the test population with similar genetic backgrounds in order to minimize population structure, to estimate the presence of population structure, and then to apply the appropriate statistical analysis to effectively adjust for population structure, if it is present.

Qualitative Traits for the Examination of Association Analysis in Soybean

Glycine soja (Sieb. & Zucc), or wild soybean, is the species from which cultivated soybean (*Glycine max* (L.) Merr.) was domesticated approximately 3,000 – 5,000 years ago in Asia (Hymowitz, 2004). Almost all *Glycine soja* accessions have purple flowers except for accessions which may be out-crosses with *Glycine max* (Chen and Nelson, 2004). Approximately 35% of *Glycine max* accessions from the USDA Soybean Germplasm Collection are white flowered (Hymowitz, 1970). These white flowered soybean plants might have been selected along with several other agronomic traits by ancient farmers during domestication.

Six loci controlling flower color variation have been reported in soybean, *W1, W2, W3, W4, Wp*, and *Wm* (Hegstad et al., 2000; Palmer et al., 2004; Xu and Palmer, 2005). Only the *W1* locus, however, has been cloned (Zabala and Vodkin, 2007). Flower color is determined by anthocyanins in the petals which are the byproduct of the flavonoid biosynthesis pathway. Groose *et al*. (1988) identified the *W4* locus from the unstable mutable line with the near-white, purple, and purple sectors on the near-white petal which could revert to the stable form near-white flower. Another locus, the *Wp* was identified by Hegstad *et al*. (2000). The authors reported that the recessive *Wp* allele in the presence of the W1 allele on linkage group (LG) D1b caused the alteration of the flower color to pink. Zabala and Vodkin (2005) found that flavanone 3-hydroxylase had reduced expression in pink flower buds versus in purple flower buds. They identified a 5.7 Kbp insertion, which is a transposable element member, in the flavanone 3-hydroxylase cDNA from the *Wp* mutant line. Those *Wp* mutant lines with the 5.7 Kbp insertion produce immature transcripts in the flower buds and seed coats. Flavanone 3-hydroxylase is the key enzyme initiating flower pigmentation metabolism. Thus, the authors concluded that the flavanone 3-hydroxylase gene encodes the *Wp* locus. Xu and Palmer (2005) discovered another mutant line at the *W4* locus which is the key locus controlling the production of anthocyanin. This mutant line contained a transposable element at the *W4* locus that produced “pale” flowers. Takahashi *et al*. (2007) reported the *Wm* locus on LG F which produces magenta flowers as a result of a single G deletion in the flavonol synthase gene. This deletion causes a truncated protein of flavonol synthase which prevents the production of the flavonols resulting in flowers that are pink in color. The *W1* locus encodes a flavonoid 3'5'-hydroxylase (*F3'5'H*) which maps to LG F. This locus conditions the flower color alteration from purple to white. Purple flower coloration is the result of the activation of the delphinidin pathway. The *F3'5'H* gene plays a key role in producing the precursor of the delphinidin pathway so that, if the *F3'5'H* gene is inactivated, flower color will be altered to white, pink, or magenta. Zabala and Vodkin (2007) used isolines which have a different allele at the *W1* locus and found that a 65 bp insertion in the third exon of the *F3'5'H* gene was completely associated with the flower coloration from purple to white. This 65 bp insertion produced a premature stop codon in the *F3'5'H* gene which resulted in an alteration of the delphinidin pathway and which produced plants with white flower rather than purple flower.

Pubescence (trichome hairs) color of soybean is controlled by the *T* locus. A mutation in the *T* allele produces pubescence color alteration from tawny to gray. Toda *et al*. (2002) identified a deletion in the flavonoid 3'-hydroxylase (*F3'H*) gene that co-segregates with the *T* locus on LG C2. They compared cDNA of two near-isogenic lines which had a different *T* allele and identified a single C deletion in the cDNA of the *F3'H* gene that produced an early stop codon and altered pubescence color from tawny to gray. Zabala and Vodkin (2005) verified that plants with gray pubescence have a one base-pair deletion in the 3' end of the *F3'H* cDNA sequence and that this deletion co-segregated with the *T* locus leading to the conclusion that the *F3'H* gene could be the *T* locus.

Quantitative Traits for the Examination of Association Analysis in Soybean

Quantitative traits are defined as traits that are controlled by multiple genes each segregating in a Mendelian fashion, that show continuous variation, and that are affected by external environmental factors. Seed protein concentration in soybean, like other agronomically important traits such as seed yield, seed oil concentration and seed size behave as quantitative traits. *Glycine soja* (Sieb. & Zucc), the wild progenitor of soybean, has high seed protein concentration (up to 55% on a dry weight basis), very small seed size (1 to 8 g/100 seeds), and low oil concentration (11% on a dry weight basis) (Bao, 1989; Xu, 1985). This contrasts to cultivated soybean that has large seed size (10 to 50 g/100 seed), average protein of 40%, and oil concentration of 20% on a dry weight basis. Seed size was probably an obvious trait for selection during domestication by ancient farmers. This selection may have unintentionally resulted in lower seed protein concentration in cultivated soybean and in a negative correlation between seed protein concentration and both seed size and seed oil concentration (Hurburgh et al., 1990; Wilcox and Guodong, 1997). A number of studies have been conducted to identify and map soybean seed protein QTLs using molecular markers in various populations. Diers *et al*. (1992) reported protein QTLs linked with eight RFLP markers using a population derived from crossing *Glycine max* and *Glycine soja*. Their results showed that seed protein concentration was controlled by several major genes with relatively large effects. A number of subsequent studies have reported the detection of seed protein QTLs including those by Mansur *et al*. (1996), Lee *et al*. (1996), Brummer *et al*. (1997), Qiu *et al*. (1999), Orf *et al*. (1999b), Sebolt *et al*. (2000), Specht *et al*. (2001), Csanadi *et al*. (2001), Chung *et al*. (2003) , Fasoula *et al*. (2004), Hyten *et al*. (2004), Kebelka *et al*. (2004), Panthee *et al*. (2005), and Nichols *et al*. (2006). In addition, numerous seed protein QTLs from these and other studies are reported in SoyBase (the USDA Soybean Genome Database, http://soybase.org). These results indicate that protein concentration in soybean seed is controlled by several QTLs. Although the presence, position and magnitude of a QTL can be affected by many factors, there are several seed protein concentration QTLs in similar or identical positions reported at least thrice in the aforementioned literature (Table 2-1). Seed protein QTLs on linkage group (LG) A1 (Mansur et al., 1996; Orf et al., 1999b; Specht et al., 2001) and on LG E (Brummer et al., 1997; Fasoula et al., 2004; Lee et al., 1996) were reported thrice. The other seed protein QTL on LG M (Csanadi et al., 2001; Hyten et al., 2004; Orf et al., 1999a; Specht et al., 2001) and on LG I (Brummer et al., 1997; Chung et al., 2003; Diers et al., 1992; Sebolt et al., 2000) were reported four times. Those seed protein QTLs each identified a number of times at similar genome locations in different studies using different sources of high seed protein suggested that these QTLs may occur at a reasonably high frequency in soybean germplasm.

Chapter 2: Association Analysis in Soybean

Introduction

Association analysis is an alternative to conventional family-based methods to detect the location of gene(s) or quantitative trait loci (QTLs). Association analysis uses the correlation between DNA marker alleles and the phenotypic expression of a trait of interest to detect genes or QTLs. A number of advantages of association analysis *versus* conventional QTL analysis have motivated plant geneticists to attempt to use association analysis for the discovery of gene(s) or QTLs associated with agricultural traits. Association analysis should provide relatively high resolution in terms of defining the genome position of a gene or QTL. It also has the advantage that it can be applied to naturally occurring populations such as human populations or germplasm collections without the requirement of knowledge of the genetic relationship among individuals. One form of association analysis is the case-control analysis. Genome-wide case-control analysis has been applied by human geneticists as a tool to find the gene(s) or genetic factor(s) underlying disease susceptibility (Gudmundsson et al., 2007; Rioux et al., 2007; Sladek et al., 2007). In the case-control study, allele frequency differences are compared between two groups, those diagnosed with a disease (the case group) and an otherwise similar group but without any disease symptoms (the control group). A marker with significant allele frequency difference between the two groups putatively identifies a genome region that contains the DNA variant or gene underlying the disease.

Linkage disequilibrium (LD) is the nonrandom association of alleles at different loci in a population. The extent of LD is one factor determining the success of whole-genome association analysis and the number of markers required for a whole-genome scan. The structure of LD partially explains the population history of a population. For instance, recombination and the degree of out-crossing are factors that decrease LD (Flint-Garcia et al., 2003; Gaut and Long, 2003) while selfing, selection, population bottlenecks, migration, and population admixture increase LD (Flint-Garcia et al., 2003; Gaut and Long, 2003; Gupta et al., 2005; Tishkoff and Verrelli, 2003). The combination of these factors may affect the level of LD in small parts of the genome or throughout the entire genome. There are two commonly used measures of the extent of LD, *D'* (standardized disequilibrium coefficient) and r^2 (correlation based on LD). The values of *D'* and r^2 range from 0 (no correlation or LD between two loci) to 1 (complete LD between two loci). The value of *D'* cannot be less than 1 unless recombination is observed between two loci which is indicated by the presence of four different haplotypes. However, the value of r^2 can be less than one without recombination because mutation can produce a third haplotype without recombination, in which case, the frequency of alleles also reflects the period of time since the mutation occurred. For instance, when the value of *D'* between two markers is 1, a higher r^2 value between the loci suggests that the mutation occurred more recently while a lower r^2 value indicates that the mutation occurred further in the past. In human genetics, a haplotype, that is the linear order of alleles on a chromosome, has been used to quantify the level of LD in different populations (The International HapMap Consortium, 2003). A haplotype block is a group of physically linked loci in strong LD. A number of discrete haplotypes may be present in a haplotype block. Tag SNPs are a selected sub-set of SNPs that together account for all of the haplotype variation in a haplotype block. A knowledge of haplotype blocks and tag SNPs that define them would help to reduce the number of SNPs required for whole-genome scans to evaluate the association of common alleles and common phenotypes. A population with a high level of LD would be anticipated to have haplotype blocks of relatively larger size. In this situation, it would not be possible to position a causal mutation with high resolution. In contrast, in a population with a low level of LD, many markers would be needed to completely define haplotype variation making whole-genome scans impractical. In the latter situation a candidate gene approach is more appropriate where high-density markers would only be positioned in close proximity to candidate genes. Such an analysis would allow the fine mapping of a causative genetic variant. However, the candidate gene approach requires a knowledge of the approximate genome position(s) of genes or quantitative trait loci (QTLs) that are likely to control the phenotype being studied.

There are several plant species for which the level of LD has been estimated. In *Arabidopsis thaliana* (L.) Heynh, a 99% selfing species, LD (r^2=0.2) decayed within 10 Kbp based upon the examination of 341,602 SNPs in 19 accessions (Kim et al., 2007). In maize (*Zea mays* L.), an out-crossing crop species, LD (r^2) declined within 100-200 bp which was determined via the analysis of 21 genic loci averaging 687 bp in length in 25 individuals (Tenaillon et al., 2001). Subsequent studies in maize have indicated higher levels of LD (r^2) depending on the gene being studied from 500 bp in the case of *sh2*, *tb1*, *d3*, and *wx1* (Wilson et al., 2004) to 500 Kbp at the *adh1* locus (Jung et al., 2004). In the selfing crop plant, barley (*Hordeum vulgare* L.), Caldwell *et al*. (2006) examined *Hinb*, *Hina*, *GSP*, and *PG*, which are genes in a 215 Kbp region surrounding *Ha*, the grain hardness locus, and compared the extent of LD (r^2) in barley cultivars, landraces, and wild barley (*Hordeum spontaneum* L.). In the *Hinb* gene, the level of LD extended to 3,000 bp in all three populations, however, the landraces had a lot of low LD and the wild barley had many intermediate and low LD than that in the cultivars. In the case of the *Hina* gene, LD extended up to 900 bp in the cultivars, less than 900 bp with many intermediate LD in the landraces, and completely declined within 1,100 bp in the wild barley. In the two other genes, *GSP* and *PG*, an insufficient number of markers made it difficult to compare the level of LD in the three populations. In soybean (*Glycine max* L. Merr.) which is a selfing species, LD (r^2) structure was analyzed in three independent genome regions ranging in length from 336 to 574 Kbp ($r^2 \leq 0.1$) in 26 wild soybeans (*Glycine soja* Sieb. & Zucc.), 52 Asian landraces, 17 ancestors of North American soybeans, and 25 elite North American cultivars (Hyten et al., 2007). *Glycine soja* had the least extensive LD (ranging from 36 to 77 Kbp) as compared to that in the landraces (100-400 Kbp), the North American ancestors (250-400 Kbp), and the elite cultivars (400-600 Kbp) in all three genomic regions. The different level of LD between selfing and out-crossing species can be mainly attributed to the much higher recombination rate in outcrossers. In addition, there are a number of factors that increase the level of LD of a domesticated crop as compared with that of the wild progenitor of the crop. These would include domestication as a result of selection and population bottlenecks and population admixture resulting from migration or movement of the domesticate from region to region.

There have been several studies in plant species using association analysis to identify specific genes controlling traits of interest with markers developed in the vicinity of candidate genes. In maize (*Zea mays* L.), which is an out-crossing species, association analysis was performed to detect alleles associated with the reduction of flowering time (Thornsberry et al., 2001). The authors detected nine SNPs from the *Dwarf8* gene in 92 inbred lines and six nucleotide deletions that were associated with a 7 to 11 day reduction in flowering time. Palaisa *et al*. (2003) found a locus strongly associated with yellow endosperm using the sequence of 10 low copy regions in close proximity to the *Y1* gene in 75 orange, yellow, and white endosperm lines in maize. Of the 168 SNPs detected in the 10 genomic regions, SNPs developed from a region 600 Kbp upstream of the *Y1* gene showed significant associations with yellow endosperm. Wilson *et al*. (2004) attempted to find SNPs involved in kernel starch biosynthesis from the six candidate genes, *ae1*, *bt2*, *sh1*, *sh2*, *sugary1*, and *waxy1* via the analysis of 102 diverse maize inbred lines. Three of the genes, *bt2*, *sh1*, and *sh2* contained SNP markers showing significant associations with kernel starch composition. In *Arabidopsis*, the model selfing species, alleles associated with flowering time were identified by Olsen *et al*. (2004) and Stinchcombe *et al*. (2004). Olsen *et al*. (2004) analyed 103 SNPs in the *CRY2* locus in 95 accessions and 90 SNPs were strongly associated with the reduction of flowering time under short days. Stinchcombe *et al*. (2004) detected the most significant associations with alleles from the *FRI* locus with flowering time in 70 Northern European and Mediterranean ecotypes. The first whole-genome association analysis in a plant species was reported in *Arabidopsis* by Aranzana *et al.* (2005). The authors verified the previously determined location of loci controlling flowering time (*FRI*) and disease resistance (*Rpm1*, *Rps5*, and *Rps2*) using haplotypes estimated from SNPs developed from more than 976 fragments across the genome in 95 *Arabidopsis* accessions.

Population structure is another confounding factor that influences the success of association analysis. Population structure is the result of allele frequency differences between different subpopulations in a population. The allele frequency difference between different populations is the result of one or more events that have occurred in the population including selection, migration, local adaptation, geographical isolation, or genetic draft. The presence of population structure causes spurious associations between markers and the trait under study not because of markers are genetically associated with the trait of interest but because the allele frequency difference between subpopulations. Several statistical procedures have been developed to determine the presence of population structure in a population being studied and to eliminate or reduce the undesirable effects of population structure which often cause spurious associations or mask true associations. Pritchard *et al*. (2000a) developed the STRUCTURE algorithm which has been used to adjust for population structure in human populations resulting from admixture, which is a population derived from more than two different homogeneous groups of individuals between which some degree of intermating has occurred for a number of generations (Darvasi and Shifman, 2005). The linkage model in the STRUCTURE algorithm estimates the correlation within markers across the genome as well as along each chromosome in which markers with high correlation may be the alleles derived from the same ancestry (Falush et al., 2003). The output of the structure analysis is a matrix which can be incorporated into association analysis to adjust for population structure. However, when one uses the linkage model, this method cannot consider LD between very tightly linked loci, resulting in lack of power to adjust for population structure. To provide additional analysis of population structure, Hardy and Vekemans (2002) introduced the kinship matrix. Kinship is indicative the degree of identity by state of two homologous loci between individuals in a population. When two individuals have the same allele in homologous loci, this allele is a condition of identity by state. Thus, the adjustment for structure using the kinship matrix increases the accurate estimate of genetic distance between individuals. Yu *et al*. (2006) suggested a mixed-linear-model which incorporated the structure matrix (Falush et al., 2003) and the kinship matrix (Hardy and Vekemans, 2002) to adjust for population structure in plant species. The mixed-linear-model analysis reduced the rate of Type I error in maize (Yu et al., 2006) and *Arabidopsis* (Zhao et al., 2007) association analyses. However, in the case of association analysis in *Arabidopsis*, several new associations in addition to known loci were detected which led the authors to question the veracity of the analysis (Zhao et al., 2007). Soybean has been domesticated for more than 3,000 years (Hymowitz, 1970) and thus population history is unknown and might be very complex. Therefore, for an association analysis, the individuals in a population should be carefully selected to have similar genetic backgrounds in terms of the origin, the time of maturation, and all other phenotypic and morphological traits in order to reduce the confounding effects caused by population structure.

In soybean, the loci controlling the qualitative traits, flower color and pubescence color, have been identified using linkage analysis. Almost all *Glycine soja*, the wild soybean, have purple flowers. White flowered soybean is produced by a mutation at the *W1* locus that encodes the flavonoid 3'5'-hydroxylase (*F3'5'H*) (Zabala and Vodkin, 2007) on linkage group (LG) F. Zabala and Vodkin (2007) identified a 65 bp insertion at the 3' end of the *F3'5'H* gene that causes an alteration in flower color from purple (*W1_*) to white (*w1w1*). In the case of pubescence color, tawny (*T_*) or gray (*tt*) color is determined by the allele substitution at the *T* locus that encodes a flavonoid 3'-hydroxylase (*F3'H*) on LG C2. Toda *et al*. (2002) reported that a single base-pair deletion in the *F3'H* gene co-segregated with the *T* locus which supported the conclusion that the *F3'H* was the *T* locus. However, unlike genes controlling qualitative traits, there are no specific genes controlling quantitative traits that have been identified in soybean. Seed protein concentration is a quantitative trait which is determined by the interaction among many genes with small and moderate genetic effects and the environment. There are 72 seed protein QTLs that have been reported in a number of studies over the last 15 years using QTL analysis (SOYBASE, the USDA Soybean Genome Database). Of these, a number of QTLs were identified more than two times in the identical or very similar positions in different populations (Table 2-1). This suggests that these QTL regions, that have been reported several times in the same genome position are likely to contain seed protein QTLs with relatively large genetic effects and that occur relatively frequently, and would thus be

Table 2-1. Linkage group, marker associations, position in the molecular linkage group for seed protein QTLs, R^2 that have been reported more than three times in the literature.

LG	DNA marker	cM position	Reference	R^{2*}
A1	B170_1	95	Specht et al., 2001	0.05
A1	T155_1	93	Mansur et al., 1996	0.09
A1	T155_1	93	Orf et al., 1999b	0.15
E	A454_1	30	Fasoula et al., 2004	0.12
E	A454_1	31	Lee et al., 1996	0.09
E	B174_1	31	Brummer et al., 1997	0.11
E	Satt384	20	Tajuddin et al., 2003	0.07
I	A144	32	Brummer et al., 1997	0.28
I	A144_1	32	Diers et al., 1992	0.24
I	A144_1	32	Sebolt et al., 2000	0.44
I	Satt496 ~ Satt239	36	Chung et al., 2003	0.28
I	Satt239 ~ ACG9b	36	Nichols et al., 2006	0.12
M	R079_1	39	Orf et al., 1999	0.06
M	Satt540	35.8	Hyten et al., 2004	0.13
M	Satt567	33	Csanadi et al., 2001	0.07
M	Satt567	33	Specht et al., 2001	0.27

*R^2 is the proportion of the total genetic variability explained by the QTL

good candidate regions to examine the ability of association analysis to detect the presence of seed protein QTL.

Soybean is the most important legume crop in the world and a main source of protein for livestock in the U.S.A. In 2007, soybean was planted in 25.4 M ha in the U.S and production was 70.8 M MT (National Agricultural Statistics Service, 2007). For a stable supply of high quality soybean products, the development of new superior cultivars with beneficial agricultural traits such as high yield, high protein concentration, resistance to biotic and abiotic stress has been the objective of breeding programs for many years. An important initial step in the breeding of improved crop cultivars would be the identification of new beneficial alleles. However, in the case of soybean, a recent study indicated that currently cultivated soybeans in the U.S contain about 72% of the sequence diversity of the Asian landraces from which they derive (Hyten et al., 2006). However, approximately 79% of alleles with minor allele frequency of less than 0.1 in the Asian landraces are absent in this same group of modern cultivars (Hyten et al., 2006). Thus, these Asian landraces may contain much genetic variation that could be used to improve modern U.S. cultivars. The USDA Soybean Germplasm Collection (U.S. Department of Agriculture, Agriculture Research Station, University of Illinois, Urbana, IL) has in excess of 19,000 soybean accessions collected mostly in Asia over the past 80+ years. These germplasm accessions should provide an excellent source of genetic variability for soybean improvement. Association analysis would be an appropriate approach to identify the location of genes/QTLs in the Soybean Germplasm Collection that could be used for the improvement of a number of traits including seed protein concentration.

The objectives of this study were 1) to provide a preliminary estimate of the level of LD in soybean germplasm, 2) to assess the presence or absence of population structure and to adjust for population structure if it is determined to be present, and 3) to apply whole-genome association analysis for the detection of the genes controlling the qualitative traits, flower color and pubescence color, and the quantitative trait, seed protein concentration.

Materials and Methods

Plant materials

The germplasm accessions were selected based upon seed protein concentration as reported in the GRIN (Germplasm Resource Information Network, U.S. Department of Agriculture, Agricultural Research Service, http://www.ars-grin.gov/npgs/index.html) database. The 319 germplasm accessions were selected to create two groups, a case group including 159 accessions with high seed protein concentration ranging from 45.9 to 51.4% and a control group including 160 accessions with normal seed protein concentration ranging from 40 to 43.4% (Table 2-2). These high and normal seed protein groups were similar in terms of their maturity group (II, III & IV), origin (China, Korea, Japan), growth habit, seed coat color, and other phenotypic traits but seed size, seed oil concentration, and seed protein concentration were highly variable. Among the 319 germplasm accessions, 159 were white flowered *vs.* 160 that were purple flowered. One hundred and thirty

Table 2-2. Germplasm collection used in this study and phenotypic traits.

PI	MG[a]	Origin	Flower Color	Pubescence color	Seed Protein Conc. (%)		Seed coat color
GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	From this study	GRIN[b]
PI 507353	II	JAPAN	WHITE	GRAY	40.7	42.25	YELLOW
PI 507297	II	JAPAN	PURPLE	GRAY	40.2	41.90	YELLOW
PI 438226	II	CHINA	PURPLE	TAWNY	47.1	45.12	YELLOW
PI 437931	II	CHINA	PURPLE	TAWNY	47.8	NA	BLACK
PI 437877 B	II	CHINA	WHITE	GRAY	47.5	45.88	YELLOW
PI 437873	II	CHINA	WHITE	GRAY	40.7	41.90	YELLOW
PI 437902 C	II	CHINA	WHITE	TAWNY	46.6	44.57	YELLOW
PI 507552	II	JAPAN	WHITE	TAWNY	41.2	43.15	YELLOW
PI 437715	II	CHINA	WHITE	GRAY	40.8	42.43	YELLOW
PI 437711 A	II	CHINA	PURPLE	TAWNY	47.0	45.02	YELLOW
PI 507063	II	JAPAN	WHITE	TAWNY	41.8	43.44	YELLOW
PI 437899	II	CHINA	PURPLE	TAWNY	46.9	44.28	YELLOW
PI 507516	II	JAPAN	WHITE	TAWNY	47.7	NA	GREEN
PI 507164	II	JAPAN	WHITE	GRAY	41.3	43.43	YELLOW
PI 506881	II	JAPAN	WHITE	GRAY	41.4	43.83	YELLOW
PI 507162	II	JAPAN	PURPLE	GRAY	41.9	43.52	YELLOW
PI 464941	II	CHINA	WHITE	GRAY	41.0	43.10	YELLOW
PI 464922	II	CHINA	PURPLE	GRAY	40.7	42.87	YELLOW
PI 506825	II	JAPAN	PURPLE	TAWNY	40.2	42.03	YELLOW
PI 438070	II	CHINA	WHITE	GLABROUS	46.1	43.88	YELLOW
PI 437890 B	II	CHINA	PURPLE	TAWNY	46.8	45.74	YELLOW
PI 437908	II	CHINA	PURPLE	TAWNY	41.4	41.27	YELLOW
PI 437817	II	CHINA	WHITE	GRAY	46.6	NA	BUFF
PI 437845 B	II	CHINA	PURPLE	GRAY	47.0	44.32	YELLOW
PI 437882 A	II	CHINA	PURPLE	GRAY	46.9	45.62	YELLOW
PI 437743	II	CHINA	WHITE	GRAY	46.4	44.80	YELLOW
PI 437722	II	CHINA	WHITE	GRAY	40.7	43.30	YELLOW
PI 437904	II	CHINA	PURPLE	TAWNY	41.0	41.94	YELLOW
PI 424201	II	CHINA	WHITE	GRAY	46.0	43.04	YELLOW
PI 437568	II	CHINA	PURPLE	TAWNY	41.9	NA	BROWN
PI 506942	II	JAPAN	WHITE	GRAY	40.7	43.53	YELLOW
PI 438144	II	CHINA	PURPLE	GRAY	47.4	45.75	YELLOW
PI 437718	II	CHINA	PURPLE	TAWNY	47.2	NA	BLACK
PI 437699	II	CHINA	WHITE	GRAY	41.4	43.30	YELLOW
PI 437698	II	CHINA	WHITE	GRAY	41.4	42.07	YELLOW
PI 423932	II	JAPAN	PURPLE	TAWNY	46.7	NA	YELLOW
PI 437592	II	CHINA	WHITE	GRAY	41.3	41.45	YELLOW
PI 437647	II	CHINA	PURPLE	GRAY	41.0	43.12	YELLOW
PI 437685 B	II	CHINA	PURPLE	GRAY	46.1	45.33	YELLOW
PI 417040 B	II	CHINA	WHITE	GRAY	47.9	44.35	YELLOW
PI 417487	II	JAPAN	WHITE	GRAY	48.2	45.70	YELLOW
PI 427088 C	II	CHINA	PURPLE	TAWNY	47.9	44.17	YELLOW
PI 417304	II	JAPAN	WHITE	GRAY	48.3	46.04	YELLOW
PI 430597	II	CHINA	WHITE	TAWNY	46.8	NA	BROWN
PI 417029	II	JAPAN	PURPLE	GRAY	48.4	45.10	YELLOW
PI 430596	II	CHINA	WHITE	GRAY	46.6	45.87	YELLOW

Table 2-2. Cont.

PI	MG[a]	Origin	Flower Color	Pubescence color	Seed Protein Conc. (%)		Seed coat color
GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	From this study	GRIN[b]
PI 416941	II	JAPAN	PURPLE	GRAY	47.6	44.53	YELLOW
PI 416773	II	JAPAN	PURPLE	TAWNY	48.0	44.98	YELLOW
PI 423948 A	II	JAPAN	PURPLE	GRAY	49.9	46.28	YELLOW
PI 407717	II	CHINA	WHITE	GRAY	47.8	44.40	YELLOW
PI 417349	II	JAPAN	PURPLE	GRAY	46.7	44.00	YELLOW
PI 417268	II	JAPAN	WHITE	TAWNY	46.5	43.55	YELLOW
PI 407655 B	II	CHINA	PURPLE	TAWNY	47.6	NA	GRAY
PI 417151	II	JAPAN	WHITE	GRAY	47.6	45.68	YELLOW
PI 417174	II	JAPAN	WHITE	GRAY	47.9	43.98	YELLOW
PI 417452	II	JAPAN	WHITE	TAWNY	46.9	44.64	YELLOW
PI 291288	II	CHINA	PURPLE	TAWNY	48.7	44.87	YELLOW
PI 291302 C	II	CHINA	PURPLE	GRAY	46.8	44.00	YELLOW
PI 361053	II	CHINA	PURPLE	GRAY	46.8	44.25	YELLOW
PI 416749	II	JAPAN	PURPLE	GRAY	46.4	41.88	YELLOW
PI 416986	II	JAPAN	PURPLE	GLABROUS	48.8	43.84	YELLOW
PI 407719	II	CHINA	WHITE	GRAY	47.0	43.78	YELLOW
PI 297528	II	CHINA	WHITE	GRAY	46.9	43.93	YELLOW
PI 291309 B	II	CHINA	WHITE	TAWNY	48.4	NA	BROWN
PI 291302 B	II	CHINA	WHITE	GRAY	46.0	44.52	YELLOW
PI 398296	II	KOREA	PURPLE	GRAY	46.7	44.30	YELLOW
PI 291310 C	II	CHINA	WHITE	TAWNY	47.2	43.60	YELLOW
PI 391585	II	CHINA	WHITE	GRAY	47.7	44.98	YELLOW
PI 416904 A	II	CHINA	PURPLE	GRAY	47.1	44.22	YELLOW
PI 291286	II	CHINA	PURPLE	TAWNY	46.9	45.13	YELLOW
PI 227321	II	JAPAN	PURPLE	GRAY	40.8	44.08	YELLOW
PI 200552	II	JAPAN	WHITE	TAWNY	41.2	45.98	YELLOW
PI 200482	II	JAPAN	PURPLE	GRAY	40.6	45.63	YELLOW
PI 227684	II	JAPAN	WHITE	TAWNY	40.3	NA	GREEN
PI 200596	II	CHINA	PURPLE	GRAY	41.5	44.57	YELLOW
PI 86463	II	JAPAN	PURPLE	GRAY	41.0	43.44	YELLOW
PI 84965	II	JAPAN	PURPLE	GRAY	40.7	43.85	YELLOW
PI 92571	II	CHINA	WHITE	GRAY	40.4	44.47	YELLOW
PI 92677	II	CHINA	WHITE	GRAY	41.2	44.17	YELLOW
PI 437572	II	CHINA	PURPLE	TAWNY	46.8	NA	BLACK
PI 89170	II	CHINA	PURPLE	TAWNY	40.3	43.90	YELLOW
PI 88442	II	CHINA	PURPLE	GRAY	41.3	45.48	YELLOW
PI 84928	II	KOREA	PURPLE	GRAY	40.5	44.32	YELLOW
PI 68696	II	CHINA	PURPLE	GRAY	41.5	43.25	YELLOW
PI 68718	II	CHINA	PURPLE	TAWNY	40.2	41.72	YELLOW
PI 70224	II	CHINA	WHITE	TAWNY	40.5	46.22	YELLOW
PI 68709	II	CHINA	PURPLE	TAWNY	41.1	43.15	YELLOW
PI 72337	II	CHINA	PURPLE	GRAY	40.2	43.63	YELLOW
PI 70197	II	CHINA	PURPLE	TAWNY	40.6	46.04	YELLOW
PI 68748	II	CHINA	WHITE	GRAY	41.4	42.92	YELLOW
PI 86443	II	JAPAN	PURPLE	GRAY	41.0	45.24	YELLOW
PI 68457	II	CHINA	WHITE	GRAY	41.1	42.07	YELLOW
PI 69500	II	CHINA	WHITE	TAWNY	40.4	44.40	YELLOW

Table 2-2. Cont.

PI	MG[a]	Origin	Flower Color	Pubescence color	Seed Protein Conc. (%)		Seed coat color
GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	From this study	GRIN[b]
PI 68694	II	CHINA	PURPLE	TAWNY	41.8	44.52	YELLOW
PI 68712	II	CHINA	WHITE	GRAY	41.7	44.53	YELLOW
PI 54619	II	CHINA	WHITE	GRAY	40.1	43.18	YELLOW
PI 54607	II	CHINA	WHITE	TAWNY	41.1	43.83	YELLOW
FC 19976	II	JAPAN	WHITE	GRAY	46.3	45.42	YELLOW
PI 68421	II	CHINA	PURPLE	TAWNY	40.6	43.27	YELLOW
PI 68454	II	CHINA	PURPLE	TAWNY	40.3	43.56	YELLOW
PI 68600	II	CHINA	PURPLE	GRAY	41.6	45.18	YELLOW
PI 47131	II	CHINA	PURPLE	TAWNY	41.1	NA	BLACK
PI 507197 A	III	JAPAN	PURPLE	GRAY	46.0	44.82	YELLOW
PI 490769	III	CHINA	WHITE	TAWNY	41.5	NA	BLACK
PI 475785	III	CHINA	PURPLE	GRAY	40.4	42.13	YELLOW
PI 506572	III	JAPAN	WHITE	TAWNY	46.3	NA	GREEN
PI 506721	III	JAPAN	WHITE	GLABROUS	41.7	42.50	YELLOW
PI 506787	III	JAPAN	WHITE	TAWNY	41.5	41.68	YELLOW
PI 464920 B	III	CHINA	WHITE	GRAY	40.0	42.87	YELLOW
PI 445845	III	CHINA	WHITE	GRAY	50.4	46.98	YELLOW
PI 468919	III	CHINA	PURPLE	TAWNY	46.8	NA	BLACK
PI 417482	III	JAPAN	PURPLE	TAWNY	46.2	NA	BROWN
PI 437770	III	CHINA	WHITE	GRAY	46.3	NA	BLACK
PI 417485	III	JAPAN	WHITE	TAWNY	41.4	41.83	YELLOW
PI 437563	III	CHINA	WHITE	GRAY	46.7	44.17	YELLOW
PI 417328	III	JAPAN	WHITE	GRAY	47.8	44.92	YELLOW
PI 417309 A	III	JAPAN	PURPLE	TAWNY	47.5	NA	GREEN
PI 417248	III	JAPAN	PURPLE	TAWNY	46.5	44.10	YELLOW
PI 417291	III	JAPAN	WHITE	GRAY	47.5	44.13	YELLOW
PI 417100	III	JAPAN	WHITE	TAWNY	46.7	NA	BLACK
PI 506723	III	JAPAN	WHITE	GLABROUS	41.6	42.38	YELLOW
PI 416868 B	III	JAPAN	WHITE	GRAY	47.4	45.05	YELLOW
PI 407810	III	KOREA	PURPLE	TAWNY	46.9	NA	BLACK
PI 417075	III	JAPAN	PURPLE	GLABROUS	47.0	43.92	YELLOW
PI 416868 A	III	JAPAN	WHITE	TAWNY	47.8	45.26	YELLOW
PI 196164	III	JAPAN	PURPLE	TAWNY	40.6	NA	BLACK
PI 229336	III	JAPAN	PURPLE	GLABROUS	40.9	42.80	YELLOW
PI 416750	III	JAPAN	WHITE	GRAY	47.7	43.32	YELLOW
PI 171450	III	JAPAN	PURPLE	TAWNY	46.3	45.38	YELLOW
PI 398620	III	KOREA	PURPLE	NEAR-GRAY	48.0	44.95	YELLOW
PI 339995	III	KOREA	PURPLE	GRAY	47.7	43.68	YELLOW
PI 291306 B	III	CHINA	PURPLE	GRAY	46.4	42.20	YELLOW
PI 361101	III	KOREA	WHITE	GRAY	46.0	41.37	YELLOW
PI 243532	III	JAPAN	WHITE	TAWNY	47.8	45.38	YELLOW
PI 360843	III	JAPAN	WHITE	GRAY	48.4	43.78	YELLOW
PI 187152	III	JAPAN	PURPLE	GLABROUS	40.2	43.30	YELLOW
PI 200485	III	JAPAN	PURPLE	GLABROUS	40.7	42.70	YELLOW
PI 417244	III	JAPAN	PURPLE	GRAY	46.3	NA	GREEN
PI 393538	III	JAPAN	WHITE	TAWNY	46.3	44.05	YELLOW
PI 417066	III	JAPAN	PURPLE	GLABROUS	46.4	43.35	YELLOW

Table 2-2. Cont.

PI	MG[a]	Origin	Flower Color	Pubescence color	Seed Protein Conc. (%)		Seed coat color
GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	From this study	GRIN[b]
PI 91730	III	CHINA	PURPLE	GRAY	41.0	43.82	YELLOW
PI 91151	III	CHINA	WHITE	GRAY	46.5	44.85	YELLOW
PI 229333	III	JAPAN	PURPLE	GLABROUS	40.5	42.62	YELLOW
PI 91725 -4	III	KOREA	WHITE	GRAY	47.7	44.26	YELLOW
PI 90499 -1	III	CHINA	PURPLE	TAWNY	41.5	44.03	YELLOW
PI 88353	III	CHINA	WHITE	GRAY	41.2	42.82	YELLOW
PI 89133	III	KOREA	PURPLE	GRAY	41.4	42.18	YELLOW
PI 88349	III	CHINA	WHITE	GRAY	41.2	42.30	YELLOW
PI 86457	III	JAPAN	PURPLE	GLABROUS	40.4	42.78	YELLOW
PI 87634	III	JAPAN	PURPLE	TAWNY	41.5	42.77	YELLOW
PI 86114	III	JAPAN	WHITE	TAWNY	41.7	41.84	YELLOW
PI 506872	III	JAPAN	WHITE	TAWNY	41.5	43.53	YELLOW
PI 86111	III	JAPAN	PURPLE	TAWNY	41.0	42.24	YELLOW
PI 88287	III	CHINA	PURPLE	TAWNY	41.9	NA	BROWN
PI 404186	III	CHINA	PURPLE	GRAY	45.9	43.48	YELLOW
PI 404196 B	III	CHINA	PURPLE	GRAY	47.2	45.30	YELLOW
PI 86456	III	JAPAN	WHITE	GRAY	41.7	41.98	YELLOW
PI 86445	III	JAPAN	PURPLE	GRAY	46.5	44.95	YELLOW
PI 84610	III	KOREA	PURPLE	TAWNY	41.8	NA	BLACK
PI 85630	III	KOREA	PURPLE	GRAY	41.9	43.23	YELLOW
PI 89130	III	KOREA	WHITE	GRAY	40.6	41.72	YELLOW
PI 87615	III	KOREA	WHITE	TAWNY	40.8	42.40	YELLOW
PI 81761	III	JAPAN	WHITE	GRAY	41.9	41.90	YELLOW
PI 86081	III	JAPAN	PURPLE	GLABROUS	41.1	42.88	YELLOW
PI 80470	III	JAPAN	PURPLE	TAWNY	41.1	NA	GREEN
PI 80480	III	JAPAN	WHITE	NEAR-GRAY	40.6	41.26	YELLOW
PI 80831	III	JAPAN	WHITE	GRAY	41.9	42.07	YELLOW
PI 507448	IV	JAPAN	WHITE	NEAR-GRAY	40.6	41.40	YELLOW
PI 495017 B	IV	CHINA	PURPLE	GRAY	46.5	NA	GREEN
PI 506663	IV	JAPAN	PURPLE	TAWNY	40.5	NA	GREEN
PI 458282	IV	KOREA	PURPLE	GRAY	41.3	44.08	YELLOW
PI 506681	IV	JAPAN	PURPLE	TAWNY	40.3	42.03	YELLOW
PI 507026	IV	JAPAN	PURPLE	TAWNY	40.4	42.63	YELLOW
PI 507160	IV	JAPAN	PURPLE	GRAY	47.2	44.83	YELLOW
PI 423843	IV	KOREA	PURPLE	GRAY	46.0	43.25	YELLOW
PI 423833 A	IV	KOREA	WHITE	TAWNY	47.5	NA	GREEN
PI 417382	IV	CHINA	PURPLE	GRAY	46.7	NA	GREEN
PI 437845 D	IV	CHINA	PURPLE	GRAY	46.3	44.15	YELLOW
PI 506937	IV	JAPAN	WHITE	TAWNY	40.5	40.72	YELLOW
PI 438304 B	IV	KOREA	PURPLE	TAWNY	47.7	NA	BLACK
PI 483082 A	IV	KOREA	WHITE	GRAY	40.9	NA	GREEN
PI 507021	IV	JAPAN	PURPLE	TAWNY	40.5	41.26	YELLOW
PI 504812	IV	KOREA	PURPLE	GRAY	41.3	41.05	YELLOW
PI 408083 A	IV	KOREA	WHITE	TAWNY	46.2	NA	GREEN
PI 404174	IV	CHINA	WHITE	GRAY	45.9	44.02	YELLOW
PI 423836	IV	KOREA	WHITE	TAWNY	46.2	NA	GREEN
PI 423902	IV	JAPAN	WHITE	TAWNY	40.9	40.98	YELLOW

Table 2-2. Cont.

PI	MG[a]	Origin	Flower Color	Pubescence color	Seed Protein Conc. (%)		Seed coat color
GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	From this study	GRIN[b]
PI 430598 B	IV	CHINA	PURPLE	TAWNY	45.9	NA	GREEN
PI 417217	IV	JAPAN	WHITE	GRAY	40.5	40.22	YELLOW
PI 423979	IV	JAPAN	WHITE	GRAY	47.7	43.58	YELLOW
PI 417176	IV	JAPAN	WHITE	GRAY	46.8	42.58	YELLOW
PI 417243	IV	CHINA	WHITE	TAWNY	46.5	NA	BLACK
PI 486354 A	IV	KOREA	PURPLE	GRAY	40.4	41.40	YELLOW
PI 507312	IV	JAPAN	PURPLE	GRAY	46.5	45.28	YELLOW
PI 458227	IV	KOREA	PURPLE	GRAY	46.1	44.85	YELLOW
PI 437749	IV	CHINA	PURPLE	TAWNY	46.6	NA	GREEN
PI 437711 B	IV	CHINA	PURPLE	TAWNY	47.6	44.90	YELLOW
PI 424275	IV	KOREA	WHITE	TAWNY	46.1	NA	GREEN
PI 424258	IV	KOREA	PURPLE	TAWNY	47.2	NA	BROWN
PI 424005	IV	KOREA	WHITE	TAWNY	46.3	NA	GREEN
PI 458226	IV	KOREA	WHITE	TAWNY	48.0	NA	GREEN
PI 458184	IV	KOREA	PURPLE	TAWNY	47.5	NA	BLACK
PI 445844	IV	CHINA	WHITE	GRAY	46.1	44.02	YELLOW
PI 437916	IV	CHINA	PURPLE	TAWNY	48.9	NA	BLACK
PI 424592	IV	KOREA	WHITE	TAWNY	46.7	NA	GREEN
PI 417022	IV	JAPAN	WHITE	GRAY	46.7	46.00	YELLOW
PI 416983	IV	JAPAN	WHITE	GRAY	46.0	42.82	YELLOW
PI 423850	IV	KOREA	PURPLE	GRAY	46.7	46.28	YELLOW
PI 424583	IV	KOREA	WHITE	TAWNY	47.6	NA	GREEN
PI 507569	IV	JAPAN	WHITE	GRAY	46.5	NA	BUFF
PI 416857	IV	JAPAN	PURPLE	TAWNY	46.1	43.75	YELLOW
PI 408200 A	IV	KOREA	WHITE	GRAY	47.0	44.58	YELLOW
PI 424581	IV	KOREA	PURPLE	TAWNY	41.7	NA	BROWN
PI 423877	IV	JAPAN	PURPLE	TAWNY	40.2	41.63	YELLOW
PI 417298	IV	JAPAN	PURPLE	GRAY	46.8	44.25	YELLOW
PI 417135 A	IV	JAPAN	PURPLE	TAWNY	46.1	43.72	YELLOW
PI 408287	IV	KOREA	PURPLE	GRAY	48.0	45.33	YELLOW
PI 407805 A	IV	KOREA	PURPLE	GRAY	47.0	44.55	YELLOW
PI 417096	IV	JAPAN	PURPLE	TAWNY	46.0	NA	BLACK
PI 407796	IV	KOREA	PURPLE	TAWNY	46.5	NA	BLACK
PI 407736	IV	CHINA	WHITE	GRAY	47.5	42.75	YELLOW
PI 407773 A	IV	KOREA	WHITE	TAWNY	47.5	44.90	YELLOW
PI 408097	IV	KOREA	PURPLE	TAWNY	46.2	NA	BROWN
PI 408020 C	IV	KOREA	WHITE	GRAY	41.7	42.04	YELLOW
PI 407729	IV	CHINA	WHITE	TAWNY	46.9	NA	BLACK
PI 407658 B	IV	CHINA	WHITE	NEAR-GRAY	43.3	NA	GRAY
PI 407918 B	IV	KOREA	PURPLE	TAWNY	46.9	44.52	YELLOW
PI 408125 A	IV	KOREA	WHITE	GRAY	46.1	43.97	YELLOW
PI 407947	IV	KOREA	PURPLE	GRAY	46.4	44.68	YELLOW
PI 407914 C	IV	KOREA	PURPLE	TAWNY	47.1	44.78	YELLOW
PI 408280	IV	KOREA	PURPLE	GRAY	46.5	43.88	YELLOW
PI 408201 A	IV	KOREA	PURPLE	TAWNY	46.3	43.48	YELLOW
PI 424367	IV	KOREA	WHITE	TAWNY	46.3	NA	GREEN
PI 404185	IV	CHINA	WHITE	TAWNY	46.5	45.48	YELLOW

Table 2-2. Cont.

PI	MG[a]	Origin	Flower Color	Pubescence color	Seed Protein Conc. (%)		Seed coat color
GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	From this study	GRIN[b]
PI 398706	IV	KOREA	WHITE	TAWNY	47.9	44.78	YELLOW
PI 407658 C	IV	CHINA	WHITE	TAWNY	47.2	45.52	YELLOW
PI 417249	IV	JAPAN	WHITE	TAWNY	47.2	42.87	YELLOW
PI 424321	IV	KOREA	WHITE	TAWNY	46.0	NA	GREEN
PI 398830	IV	KOREA	PURPLR	TAWNY	46.2	43.23	YELLOW
PI 404173 A	IV	CHINA	WHITE	GRAY	46.2	44.36	YELLOW
PI 398705	IV	KOREA	PURPLR	TAWNY	46.2	45.78	YELLOW
PI 404176	IV	CHINA	WHITE	GRAY	45.9	43.78	YELLOW
PI 446893	IV	CHINA	PURPLR	GRAY	40.4	42.48	YELLOW
PI 219787	IV	JAPAN	PURPLR	TAWNY	40.2	41.70	YELLOW
PI 253653 B	IV	CHINA	PURPLR	GRAY	42.1	44.06	YELLOW
PI 404177	IV	CHINA	PURPLR	GRAY	51.4	45.77	YELLOW
PI 404183	IV	CHINA	WHITE	GRAY	46.2	43.25	YELLOW
PI 253654	IV	CHINA	WHITE	GRAY	46.6	47.65	YELLOW
PI 407731	IV	CHINA	WHITE	GRAY	46.2	43.10	YELLOW
PI 89154 -2	IV	KOREA	WHITE	TAWNY	40.3	42.12	YELLOW
PI 393540	IV	JAPAN	WHITE	TAWNY	46.0	45.08	YELLOW
PI 385942	IV	JAPAN	PURPLR	GRAY	46.4	42.98	YELLOW
PI 179826	IV	CHINA	WHITE	NEAR-GRAY	40.2	NA	BROWN
PI 96808	IV	KOREA	WHITE	GRAY	41.8	43.34	YELLOW
PI 97155	IV	KOREA	WHITE	GRAY	41.7	42.93	YELLOW
PI 157398	IV	KOREA	PURPLR	TAWNY	41.8	41.77	YELLOW
PI 92713	IV	CHINA	WHITE	TAWNY	41.1	41.83	YELLOW
PI 253656 B	IV	CHINA	PURPLR	TAWNY	46.5	NA	BROWN
PI 87561	IV	KOREA	PURPLR	GRAY	41.3	44.43	YELLOW
PI 253663	IV	CHINA	WHITE	GRAY	46.6	45.30	YELLOW
PI 248515	IV	JAPAN	WHITE	GRAY	40.3	41.23	YELLOW
PI 243522	IV	JAPAN	PURPLR	TAWNY	40.5	40.27	YELLOW
PI 248513	IV	JAPAN	WHITE	TAWNY	40.6	41.47	YELLOW
PI 91702	IV	KOREA	PURPLR	GRAY	41.0	42.65	YELLOW
PI 90760	IV	CHINA	WHITE	GRAY	40.4	41.80	YELLOW
PI 407658 A	IV	CHINA	WHITE	TAWNY	47.8	44.65	YELLOW
PI 246365	IV	JAPAN	WHITE	GRAY	40.4	41.02	YELLOW
PI 266807 D	IV	CHINA	PURPLR	GRAY	41.3	42.93	YELLOW
PI 253666 A	IV	CHINA	WHITE	GRAY	47.9	47.73	YELLOW
PI 157462	IV	KOREA	PURPLR	GRAY	41.7	NA	GREEN
PI 157453	IV	KOREA	PURPLR	TAWNY	41.9	NA	GREEN
PI 157405	IV	KOREA	PURPLR	GRAY	41.7	NA	GREEN
PI 96333	IV	KOREA	PURPLR	GRAY	41.0	NA	GREEN
PI 157437	IV	KOREA	WHITE	GRAY	41.7	41.67	YELLOW
PI 229349	IV	JAPAN	PURPLR	TAWNY	40.3	40.38	YELLOW
PI 398299	IV	KOREA	PURPLR	GRAY	48.5	NA	GREEN
PI 398446	IV	KOREA	PURPLR	TAWNY	47.4	NA	BLACK
PI 171432	IV	CHINA	WHITE	GRAY	40.8	NA	GREEN
PI 96280	IV	KOREA	PURPLR	GRAY	40.5	42.48	YELLOW
PI 92707 -2	IV	CHINA	WHITE	TAWNY	41.0	42.04	YELLOW
PI 91100 -4	IV	CHINA	WHITE	GRAY	40.8	41.15	YELLOW

Table 2-2. Cont.

PI	MG[a]	Origin	Flower Color	Pubescence color	Seed Protein Conc. (%)		Seed coat color
GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	GRIN[b]	From this study	GRIN[b]
PI 253665 C	IV	CHINA	PURPLE	GRAY	40.5	43.07	YELLOW
PI 88452	IV	CHINA	WHITE	GRAY	40.5	41.63	YELLOW
PI 157435	IV	KOREA	PURPLE	NEAR-GRAY	41.1	NA	BROWN
PI 90763	IV	CHINA	PURPLE	TAWNY	40.5	NA	BLACK
PI 88499	IV	CHINA	WHITE	GRAY	41.3	43.48	YELLOW
PI 87011	IV	KOREA	PURPLE	GRAY	40.4	44.22	YELLOW
PI 86972 -2	IV	KOREA	PURPLE	GRAY	40.5	42.20	YELLOW
PI 90221	IV	KOREA	PURPLE	GRAY	40.9	43.08	YELLOW
PI 88444	IV	CHINA	WHITE	GRAY	40.3	41.82	YELLOW
PI 89061 -3	IV	CHINA	WHITE	TAWNY	40.4	41.90	YELLOW
PI 89769	IV	CHINA	WHITE	GRAY	40.3	44.47	YELLOW
PI 86904 -1	IV	KOREA	PURPLE	TAWNY	41.0	42.48	YELLOW
PI 86903 -3	IV	KOREA	WHITE	GRAY	41.4	42.33	YELLOW
PI 79870 -4	IV	CHINA	PURPLE	TAWNY	41.2	41.28	YELLOW
PI 84669 N	IV	KOREA	WHITE	GRAY	41.4	42.02	YELLOW
PI 86136	IV	JAPAN	WHITE	TAWNY	40.9	NA	BLACK
PI 87588	IV	KOREA	WHITE	GRAY	41.7	43.72	YELLOW
PI 84751	IV	KOREA	WHITE	TAWNY	40.5	NA	BLACK
PI 83892	IV	KOREA	PURPLE	GRAY	40.2	42.03	YELLOW
PI 84628	IV	KOREA	WHITE	GRAY	40.5	41.88	YELLOW
PI 83945 -4	IV	KOREA	WHITE	TAWNY	41.0	43.40	YELLOW
PI 87631 -3	IV	JAPAN	WHITE	TAWNY	46.8	46.12	YELLOW
PI 84985	IV	JAPAN	WHITE	GRAY	40.4	40.67	YELLOW
PI 72227	IV	CHINA	WHITE	GRAY	47.6	44.82	YELLOW
PI 82312 N	IV	KOREA	PURPLE	GRAY	41.7	42.43	YELLOW
PI 79825 -1	IV	CHINA	WHITE	GRAY	40.6	41.93	YELLOW
PI 84646 -2	IV	KOREA	PURPLE	GRAY	40.8	42.63	YELLOW
PI 80466 -2	IV	JAPAN	WHITE	GRAY	47.2	45.98	YELLOW
PI 71444	IV	CHINA	WHITE	TAWNY	47.7	45.85	YELLOW
PI 79870 -6	IV	CHINA	WHITE	TAWNY	40.2	41.67	YELLOW
PI 69507 -1	IV	CHINA	PURPLE	TAWNY	41.3	42.80	YELLOW
PI 70467	IV	CHINA	PURPLE	GRAY	41.3	42.55	YELLOW
PI 61947	IV	CHINA	WHITE	GRAY	41.6	42.42	YELLOW
PI 417254	IV	JAPAN	PURPLE	GRAY	45.9	42.95	YELLOW
PI 68644	IV	CHINA	PURPLE	GRAY	41.5	42.40	YELLOW
PI 79696	IV	CHINA	WHITE	TAWNY	40.6	NA	GREEN
PI 64698	IV	KOREA	WHITE	NEAR-GRAY	41.6	NA	BLACK
PI 61944	IV	CHINA	WHITE	GRAY	41.6	42.23	YELLOW

a MG: Maturity Group
b GRIN: data from the GRIN database

had tawny pubescence *vs*. 179 with gray pubescence. In the case of pubescence color, 12 of 319 accessions were glabrous (no or very low pubescence) and were not included in the analysis of pubescence color. In addition, seven gray accessions were determined to have “near-gray pubescence”. Near-gray pubescence is controlled by the epistatic interaction between the *Td* and *T* loci (*TT TdTd*, tawny; *TT tdtd*, light tawny or near-gray; *tt TdTd* or *tt tdtd*, gray) (Iwashina et al., 2006). The near gray accessions were also eliminated from the analysis of pubescence color. In addition to the 319 accessions from the USDA Soybean Germplasm Collection, four control lines, one with normal seed protein concentration, Williams, and three with high seed protein concentration, U97-207209, U99-310435, and U97-304721 were included in the analysis. Genomic DNA from all lines was extracted from bulked young leaf tissue grown in the green house using the CTAB method (Keim et al., 1988).

Field trials and Evaluation of Protein Concentration

Field tests were conducted using a randomized complete block design with four replicates of hillplots at Beltsville, MD and two replicates at Lincoln, NE in 2003. The flower color and pubescence color of each plot was recorded. Seed protein concentration was determined with three to five sub-samples from each plot by near-infrared transmission spectroscopy (Tecator Infratec 1255 analyzer, Tecator AB, Sweden) (Nowotan et al., 2006). The seed protein analysis was determined only for those entries with yellow seed coat color for which the near-infrared transmission spectroscopy can be applied.

SNP Marker Development within the vicinity of candidate gene regions

SNP development within the candidate genes controlling qualitative traits

SNP markers were developed within the *GmF3'5'H* gene (AY117551) (Zabala and Vodkin, 2007) for flower color and the *F3'H* gene (AB191404) (Toda et al., 2002) for pubescence color. SNP discovery followed the previously described protocol used by Zhu *et al*. (2003). Briefly, PCR primers were designed to the sequence of AY117551 and AB191401 and used to amplify genomic DNA of the soybean cultivar Archer. The amplified products were analyzed on an agarose gel to identity primer sets that produced a single amplicon. The primers that amplified a single amplicon were used to amplify genomic DNA of the six soybean genotypes, Archer, Minsoy, Evans, Noir 1, Peking, and PI 209332. The DNA sequence of the amplified genome fragments of the six genotypes was determined using Sanger sequencing on the ABI3730 (Applied Biosystems, Foster City, CA). The sequence data were analyzed with Phred and aligned with with Phrap followed by analysis with PolyBayes SNP discovery software (Marth et al., 1999) using a machine learning algorithm as described by Matukumalli *et al*. (2006). The alignment for SNP predictions was viewed with the Consed viewer (Gordon et al., 1998). The fragments showing DNA sequence variation among the six genotypes were used for the genotyping via re-sequencing of the 319 germplasm accessions plus the four control genotypes.

SNP development in close proximity to previously discovered protein QTL using BAC-end sequence and end sequence of BAC subclones

For the development of SNP markers in close proximity to the previously reported seed protein QTL, BAC clones associated with simple sequence repeat

(SSR) markers, BARC-Satt496, BARC-Satt239, BARC-Sat_174, and BARC-Sat_219 in the vicinity of the protein QTL on linkage group (LG) I (Chung et al., 2003; Nichols et al., 2006) were used. BAC DNA was isolated with the Qiagen plasmid midi kit (Qiagen, Hilden, Germany). Each clone was digested separately with the four restriction endonuclease AluI, RsaI, HindIII and XmnI and then ligated into the pUC19 plasmid vector which was used to transform XL2-Blue *E. coli* competent cells. Transformed cells were identified with blue-white colony selection using X-galactosidase and isopropyl-D-thiogalactopyranoside on Luria Broth (LB) media. Plasmid-containing cells were picked into 96 well-plates containing LB liquid media and directly amplified with primers designed based on the 5' and 3' end sequence of the pUC19 plasmid after dialysis with 0.1% Tween 20. Inserts with sizes between 400 and 800 bp were selected for sequence analysis and the discovery of SNP markers. The SNP discovery procedure was conducted using sequence analysis and the alignment of amplicons from the soybean genotypes, Archer, Minsoy, Evans, Noir 1, Peking, and PI 20933 as described by Zhu *et al*. (2003). Fragments having at least one SNP were used for the genotyping of the 319 germplasm accessions.

Genotyping

A total of 3,072 SNPs which were selected from sequence tagged sites (STSs) developed from three sources of sequences, expressed sequence tags, BAC-end sequences, and BAC subclone sequences, were genotyped in the 319 germplasm accessions using the Illumina GoldenGate assay following the protocol described by Fan *et al*. (2006) and Hyten *et al*. (2008). Of these SNP markers, 2,652 (86.3%) produced successful assays with good cluster separation as described by Hyten *et al*.

(2008) and 2,579 markers were integrated on the 20 soybean LGs while the remaining 73 markers with successful assays were unlinked with markers on the 20 LGs. In addition, 41 markers, 19 SNPs developed from the protein QTL region on LG I, 19 SNPs from the *F3'H* gene and, and three SNPs from the *F3'5'H* gene were genotyped by direct sequencing as described by Zhu *et al*. (2003). Five loci were genotyped by a single-base-extension method using the Luminex flow cytometer as described by Choi *et al*. (2007). The physical distances of these markers across the whole soybean genome were determined using the recently released 7.23× whole-genome assembly of Williams 82 soybean by the U.S. Department of Energy, Joint Genome Institute, Walnut Creek, CA (http://www.phytozome.net/soybean).

LD and Haplotype Block Estimation

For the estimation of the level of LD, a total of 2,313 loci with MAF > 0.10 and the number of missing data points less than 96 (30%) were used. Heterozygous alleles were treated as missing data. For the estimation of the extent of LD, genetic distance between physically linked marker loci was used. However, because the distance between genetic markers located within a few cM of each other is not a reliable estimate, only physically linked markers within distances of 1,000 Kbp for which distance was determined using the 7.23× whole-genome assembly of Williams 82 soybean were used to estimate the extent of LD. A total of 216 sequence scaffolds, associated with 1,747 markers (MAF > 0.10) that contained at least two SNPs were used to estimate the LD between loci within 1,000 Kbp of each other. Haploview 4.0 was used for all pair-wise comparisons of the alleles to calculate r^2 (coefficient based on LD) and *D'* (standardized disequilibrium coefficient), and to estimate haplotype block size (Barrett et al., 2005). For the estimation of the average size of a haplotype block, three different methods were used. Firstly, the confidence interval method that defines haplotype block based on *D’* value from 0.7 to 0.98 between the pair of SNPs (Gabriel et al., 2002). Secondly, the four gamete rule defines the haplotype block when there is no evidence of recombination between loci (Wang et al., 2002). Thirdly, the solid spine of LD method was used to estimate haplotype block as a pairwise *D’* value of greater than 0.8 between SNPs (Barrett et al., 2005).

Statistical Analysis

An analysis of variance was conducted from which the variance components were estimated to calculate the heritability of seed protein concentration. The variance among location, replications within location, accessions, and accessions × locations interaction were determined using the SAS procedure GLM using Statistical Analysis System programs (SAS institute, Inc., Cary, NC). Genetic variance and environmental variances were estimated based on expected mean squares. Both genotypes and locations were considered to be random effects.

The heritability of seed protein concentration was defined as

$$h_B^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}, (1)$$

where σ_g^2 is the genetic variance among accessions, and σ_e^2 is the environmental variance.

To reduce Type I error caused by multiple comparisons, the Bonferroni correction was used. To estimate the presence of population structure (Q-matrix) using the most likely number of sub-groups in the 319 germplasm accessions, a total of 2,622 SNPs were analyzed with the STRUCTURE program (version 2.1) (Falush et al., 2003). The linkage model of STRUCTURE was used to treat all alleles in the intact-linkage-block as derived from the same ancestral population. The length of the burn-in period which is the number of simulations before the collecting data was 50,000 and the number of Markov Chain Monte Carlo was 50,000 which explains the number of runs after the burn-in period to estimate accurate ln Pr($X|K$) with different values of K from 5 to 15. The variability of estimates obtained from a different K as well as the posterior probability calculated based on Bayesian rule (Latch et al., 2006) was compared to select the most likely number of K. The kinship coefficient (K-matrix) that explained the most probable number of identity by state of each allele between random individuals was estimated with SPAGeDi v1.2 (Hardy and Vekemans, 2002). For whole-genome case-control analysis, a Chi-square test was used as a naïve test for flower color and pubescence color. Regression analysis was used for seed protein concentration QTL detection using SAS. To compare the power of the structure adjustment of the different methods, logistic regression incorporating the Q-matrix, general-linear-model incorporating the Q-matrix, and a mixed-linear-model incorporating both the Q- and K-matrixes were applied using the TASSEL program (Yu et al., 2006) for flower color and pubescence color. PROC MLM was used in SAS to detect QTLs for seed protein concentration.

Results

Linkage Disequilibrium in Soybean Germplasm Accessions

A total of 3,115 SNPs across the soybean-genome were genotyped in 319 soybean germplasm accessions. Of these markers, 2,698 (86.6%) produced successful golden gate assays and 2,622 were positioned on the most recent version of the Soybean Consensus Map of 20 soybean LGs. After filtering rare and monomorphic SNPs, a total of 2,313 loci with a minor allele frequency (MAF) of > 0.10 were used for the estimation of LD. This was an average density of one SNP every 0.99 centimorgan (cM). The extent of LD measured by *D'* and r^2 was calculated between all physically linked pairs of markers based on genetic distance (Fig. 2-1). The *D'* value declined to 0.35 within 5 cM and the r^2 value decreased to less than 0.1 within 5 cM. Each linkage group showed a similar pattern of LD decline (Fig. 2-2). To obtain a better estimate of LD decline, the *D'* and r^2 values for loci within 1,000 Kbp (≅ 2.2 cM) based on their physical distance were plotted against distance (Fig. 2-3). Physical distance was determined with 2,259 SNP markers associated with sequence scaffolds of Williams 82 soybean. The 41 markers developed from the targeted genes and candidate seed protein QTL regions were also associated with sequence scaffolds. LD, as measured by r^2, declined very rapidly to less than 0.8 within 50 Kbp and slowly decreased to 0.2 over 1,000 Kbp. In those regions with sufficiently high marker density, *D'* was used to estimate the level of haplotype block structure and length. Haplotype block size was estimated with a total of 1,970 markers (MAF > 0.10) associated with the 216 sequence scaffolds in the

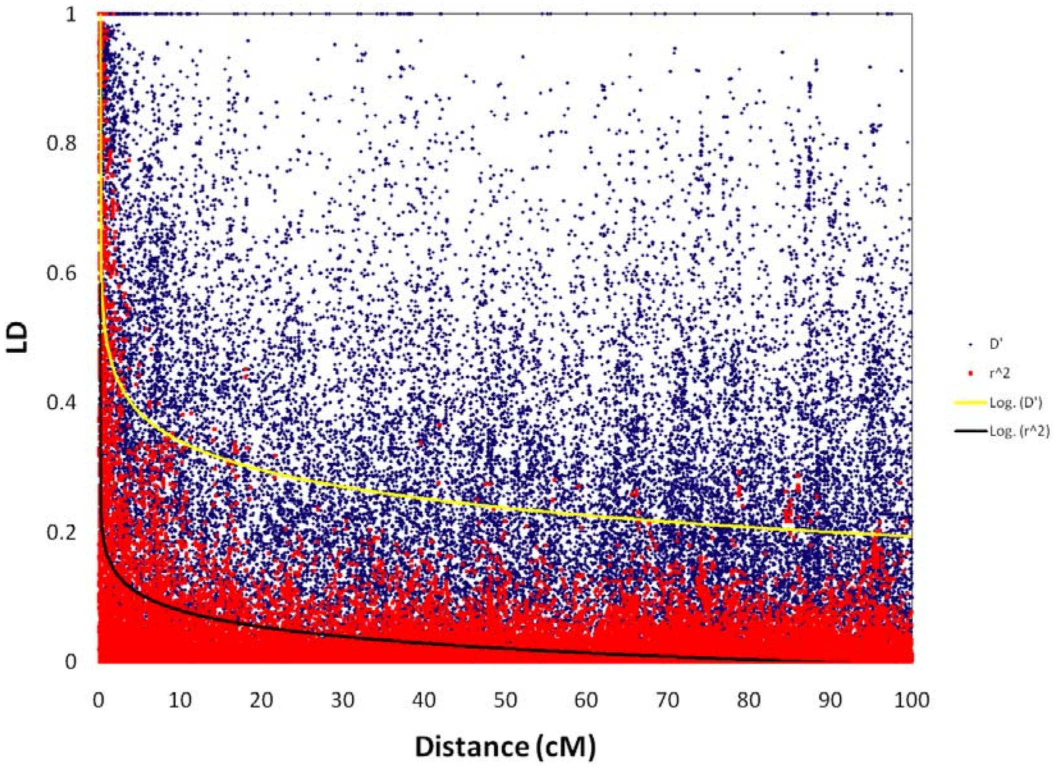
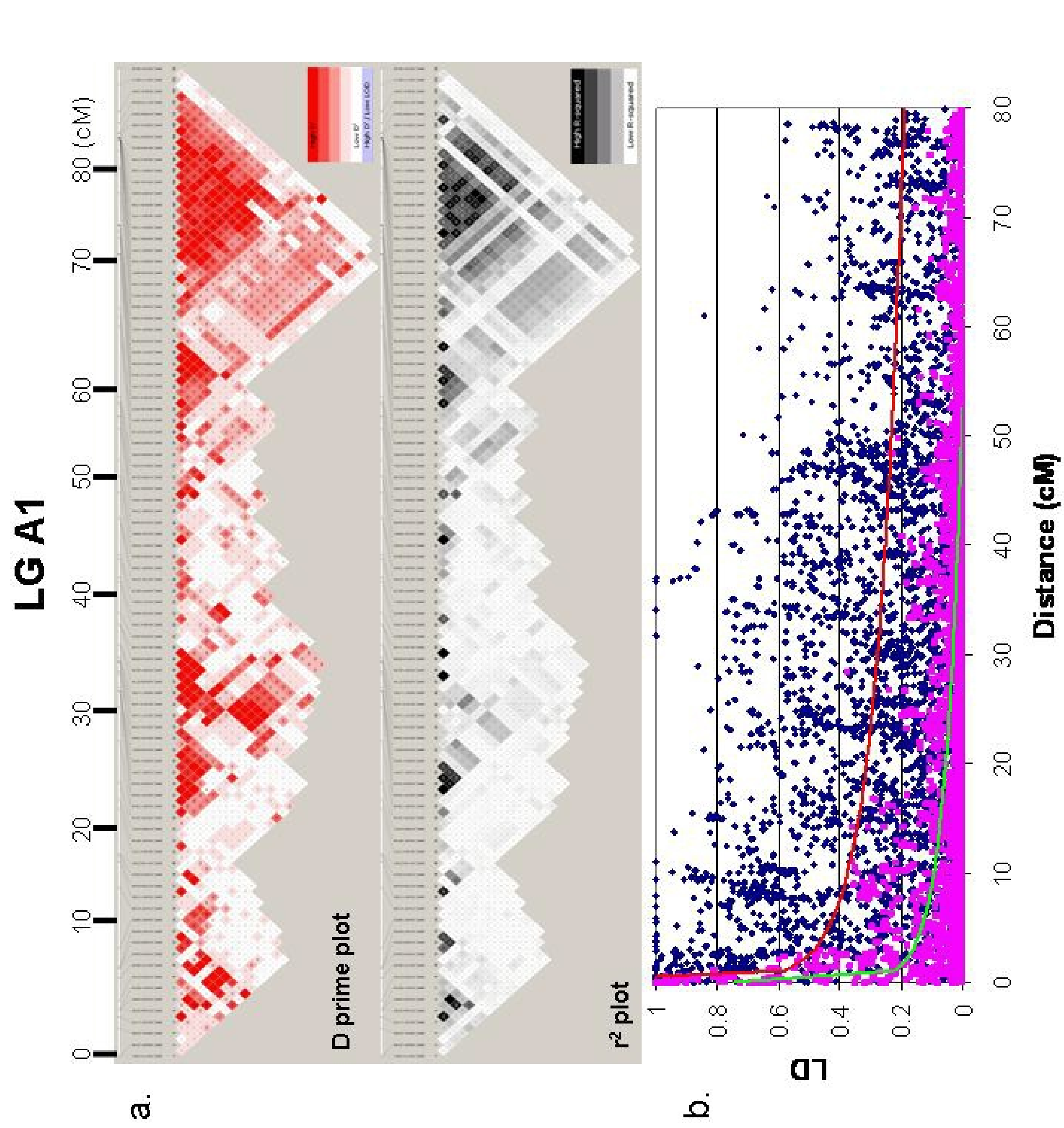
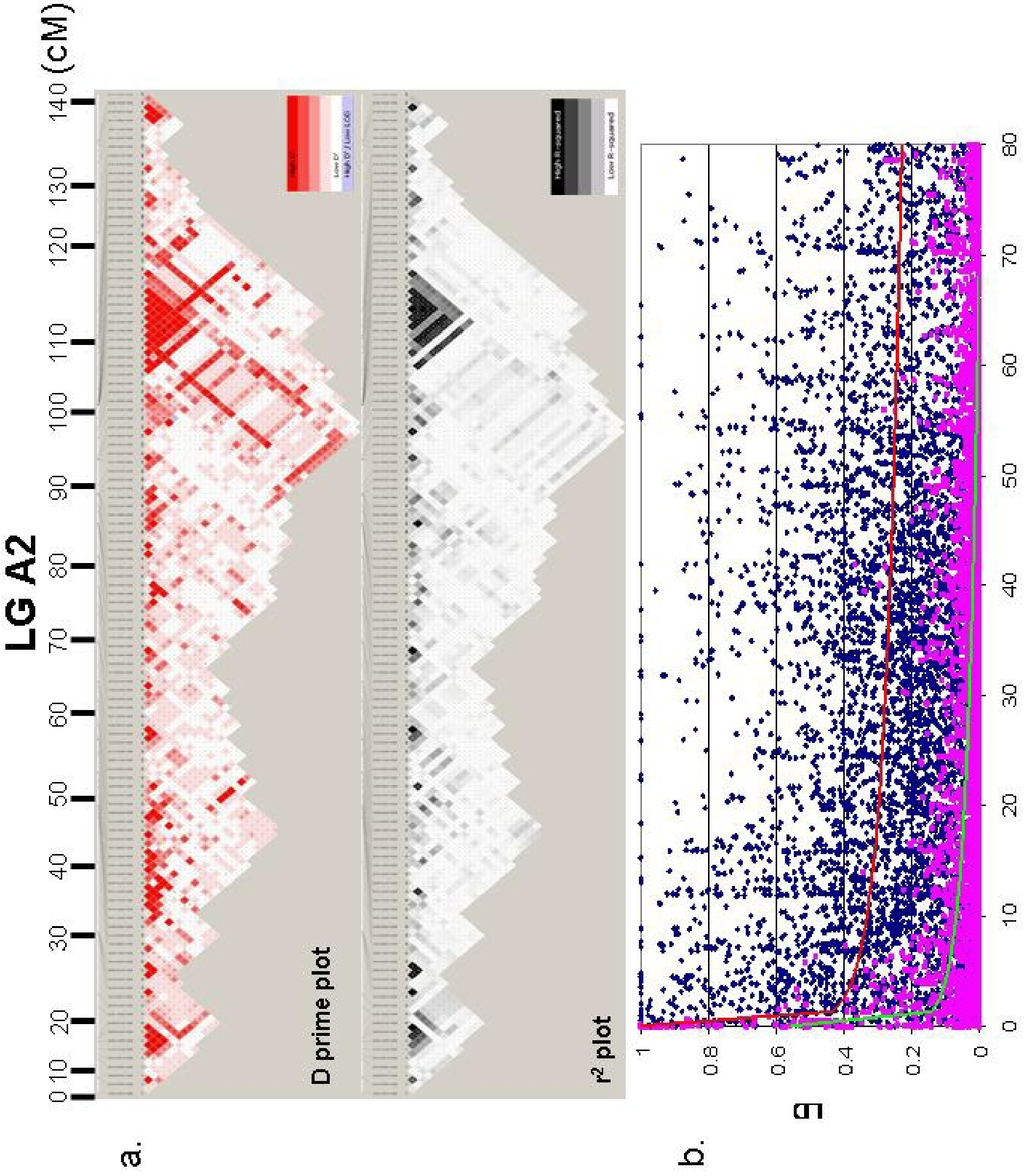
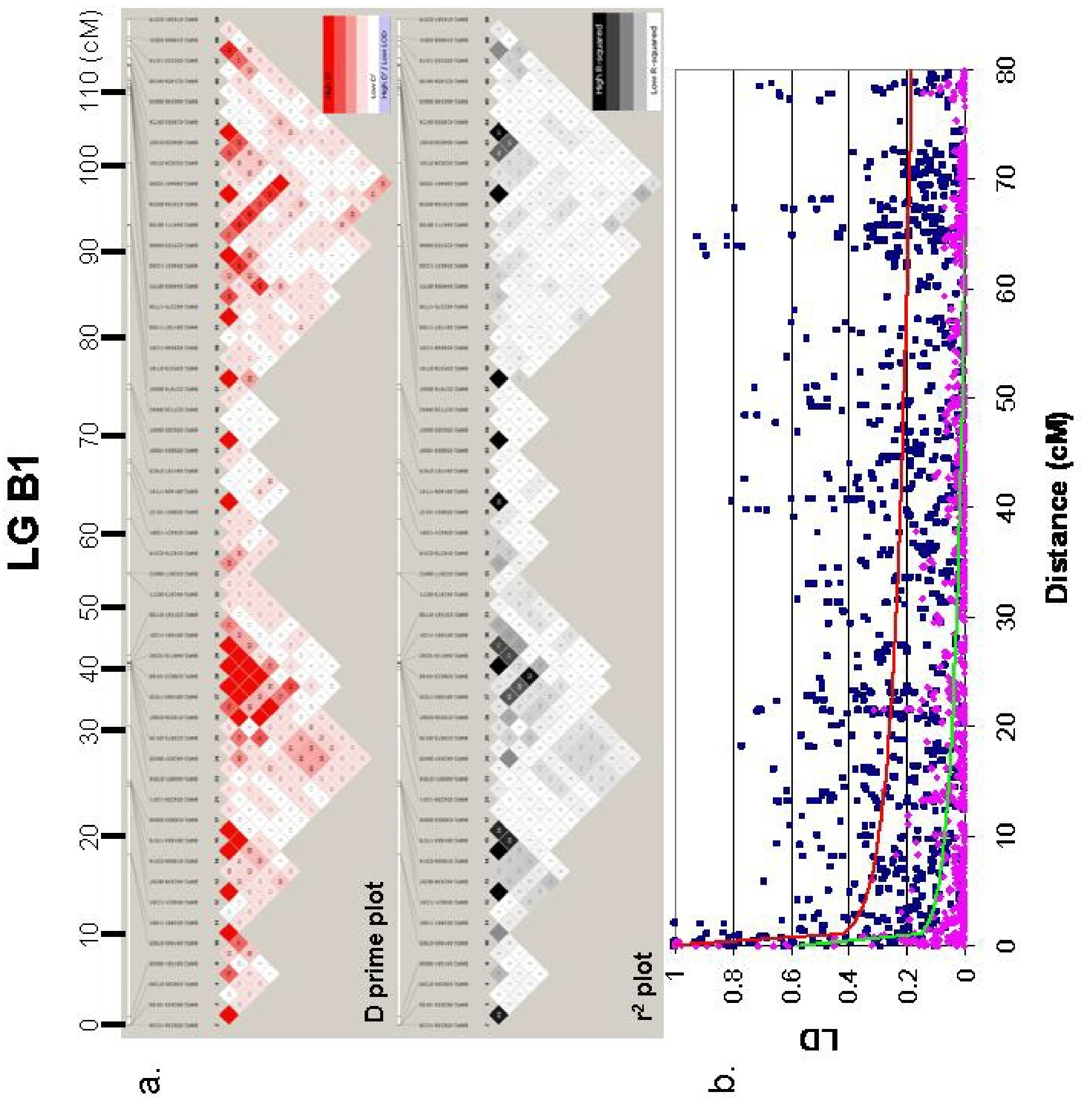
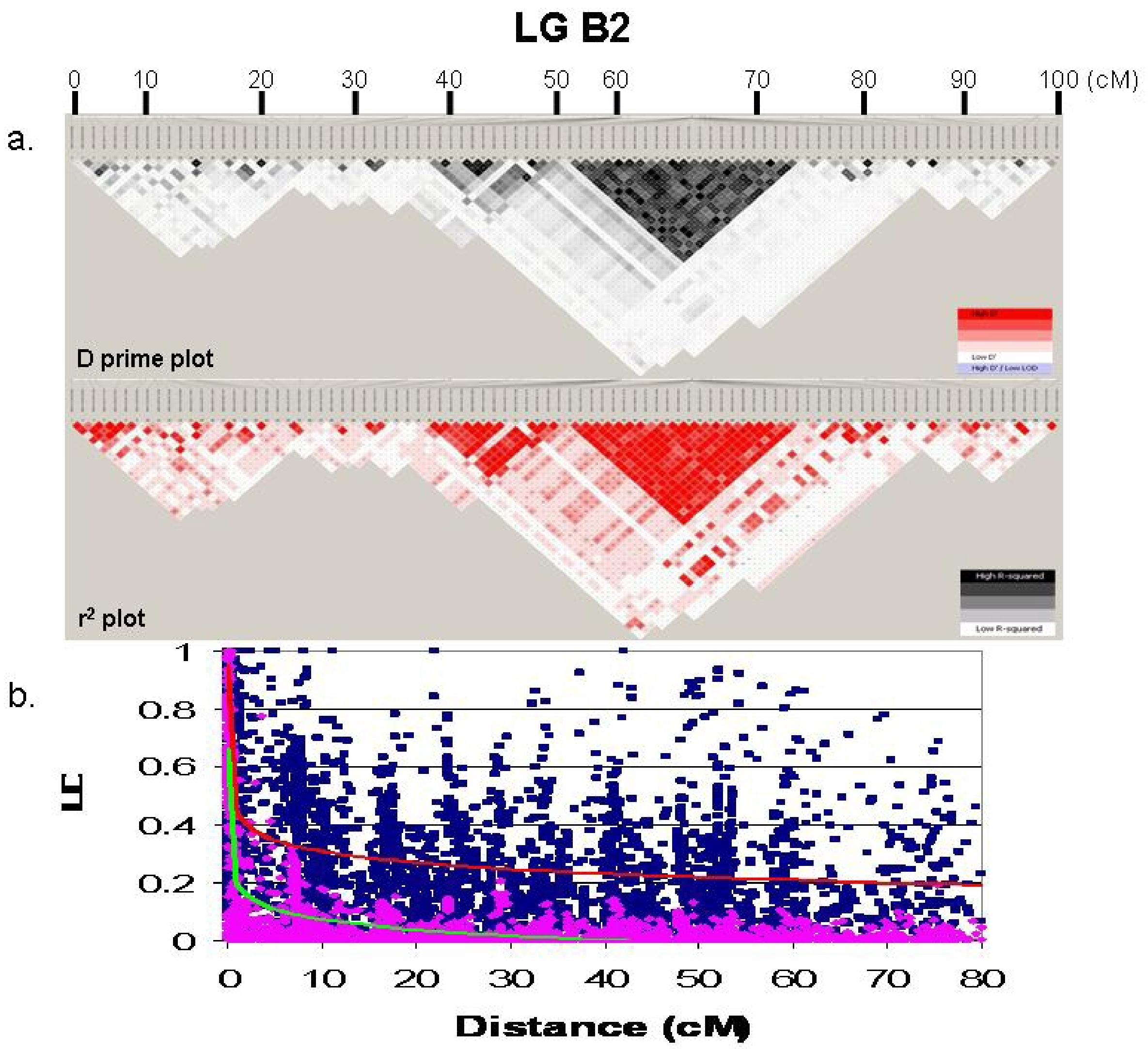


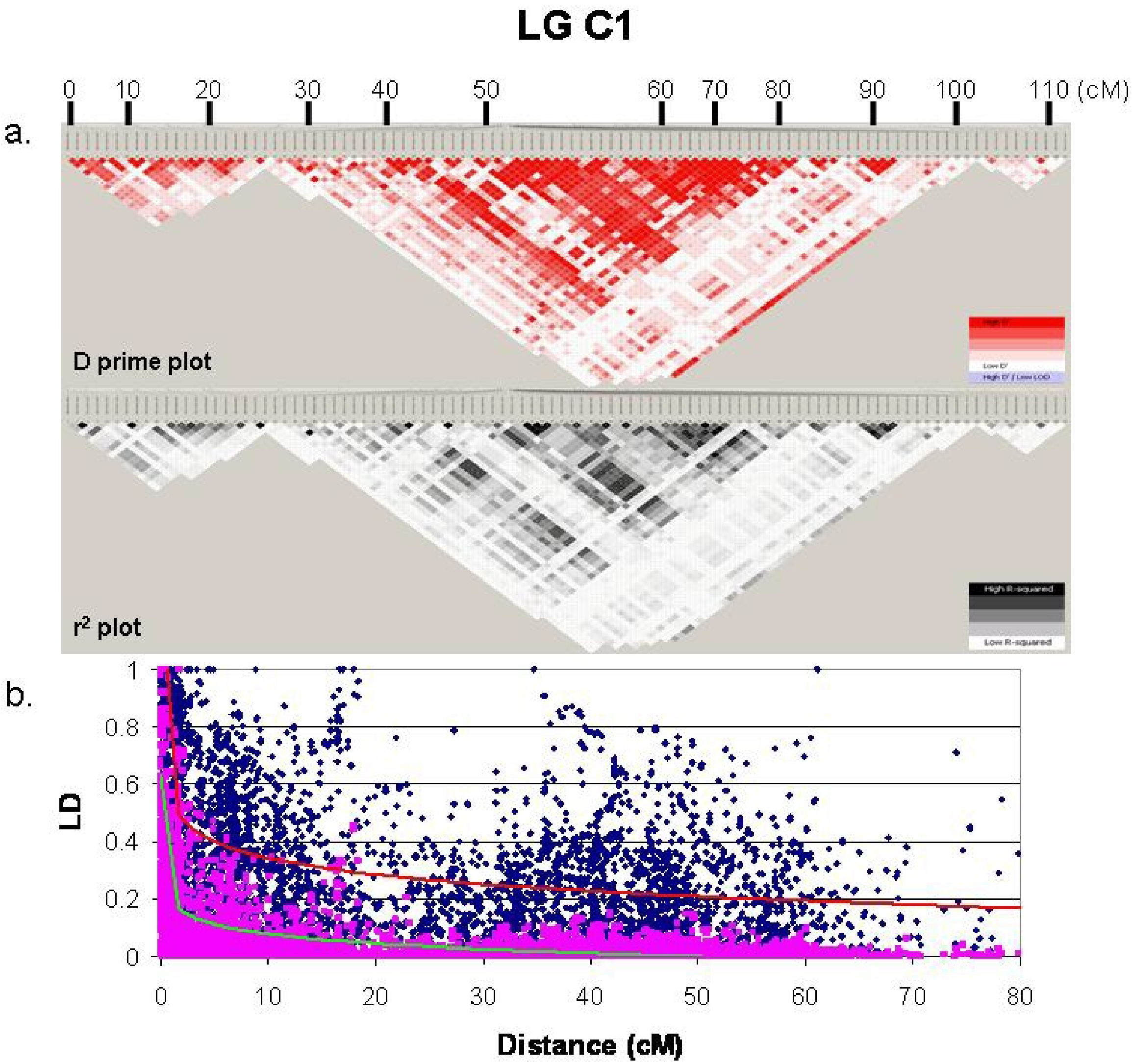
Figure 2-1. Linkage disequilibrium (LD) in the 319 soybean germplasm accessions. A total of 2,313 SNP loci (MAF > 0.10) across the soybean genome were analyzed in the 319 germplasm accessions. *D'* and r^2 are used to show the level of LD as a function of distance between pairs of physically linked loci. Blue dots are the *D'* value and red dots are the r^2 value of pairwise comparisons. The logarithmic trend lines of *D'* and r^2 are indicated in yellow and black, respectively.

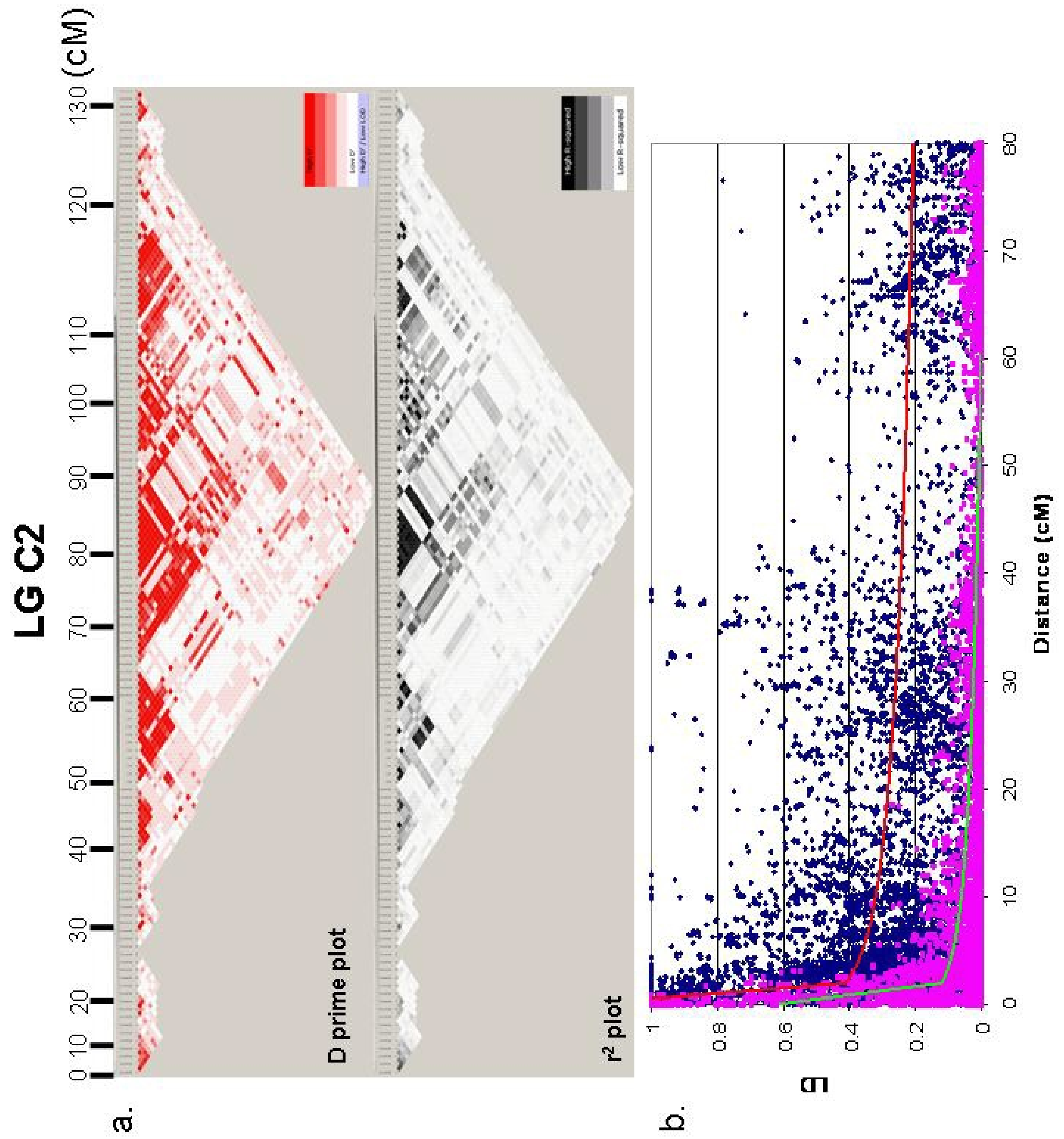

LG A1
a.
0
10
20
30
40
50
60
70
80 (cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r^2 plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

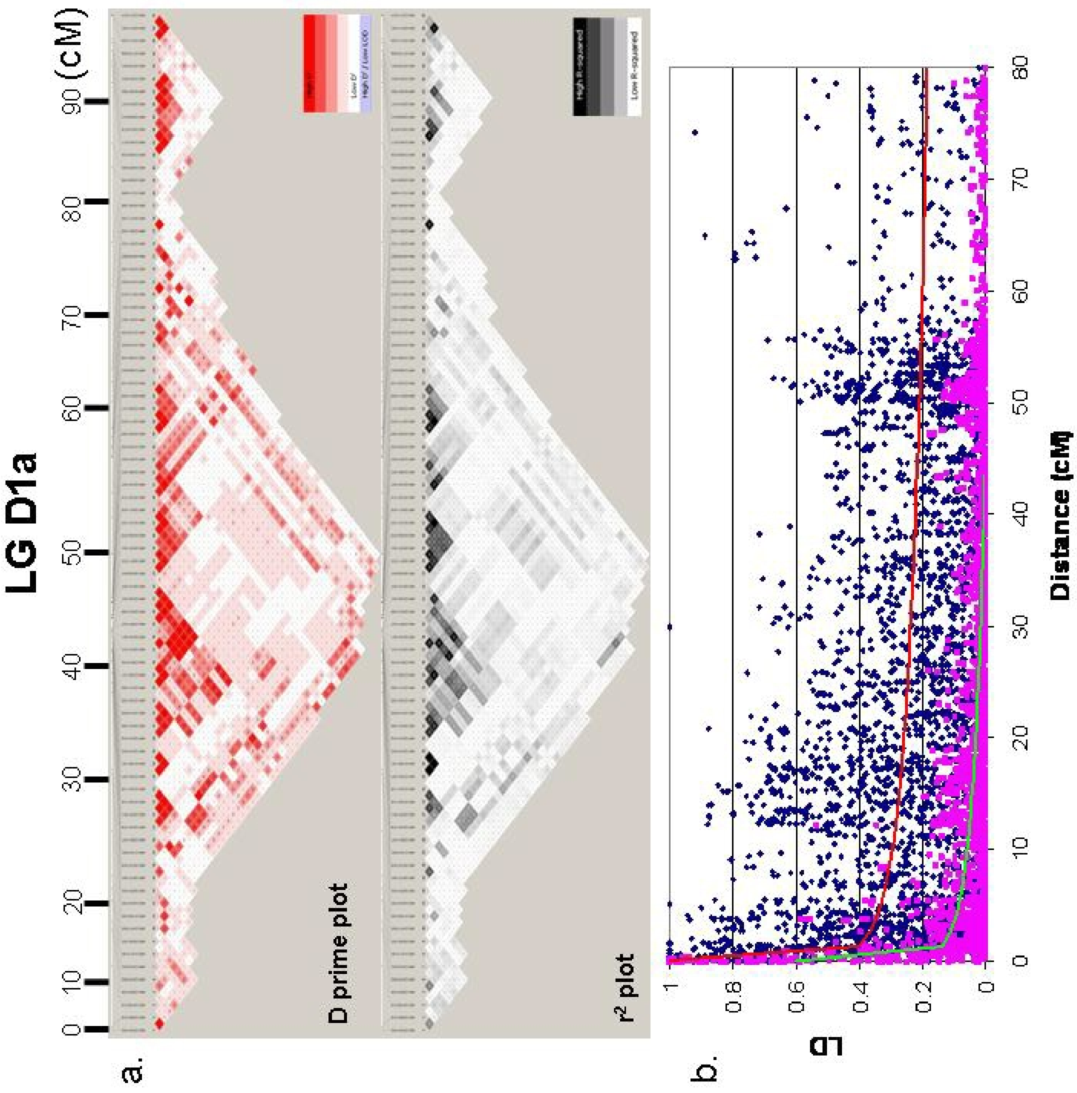

LG A2
(cM)
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
a.
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

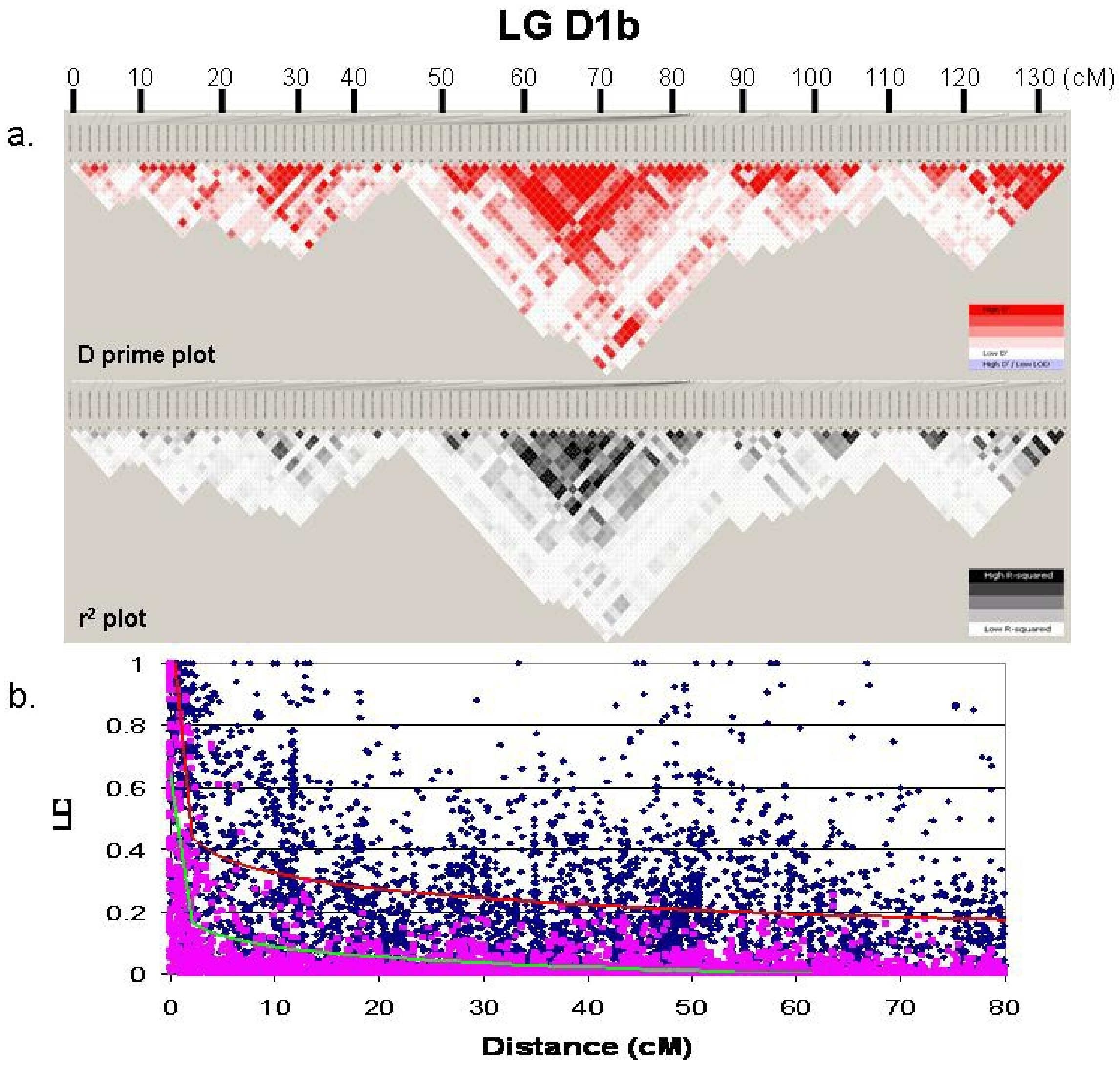

LG B1
a.
0
10
20
30
40
50
60
70
80
90
100
110 (cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)


LG B2
0
10
20
30
40
50
60
70
80
90
100 (cM)
a.
D prime plot
High D'
Low D'
High D' / Low LOD
r2 plot
High R-squared
Low R-squared
b.
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)


LG C1
a.
0
10
20
30
40
50
60
70
80
90
100
110 (cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r2 plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

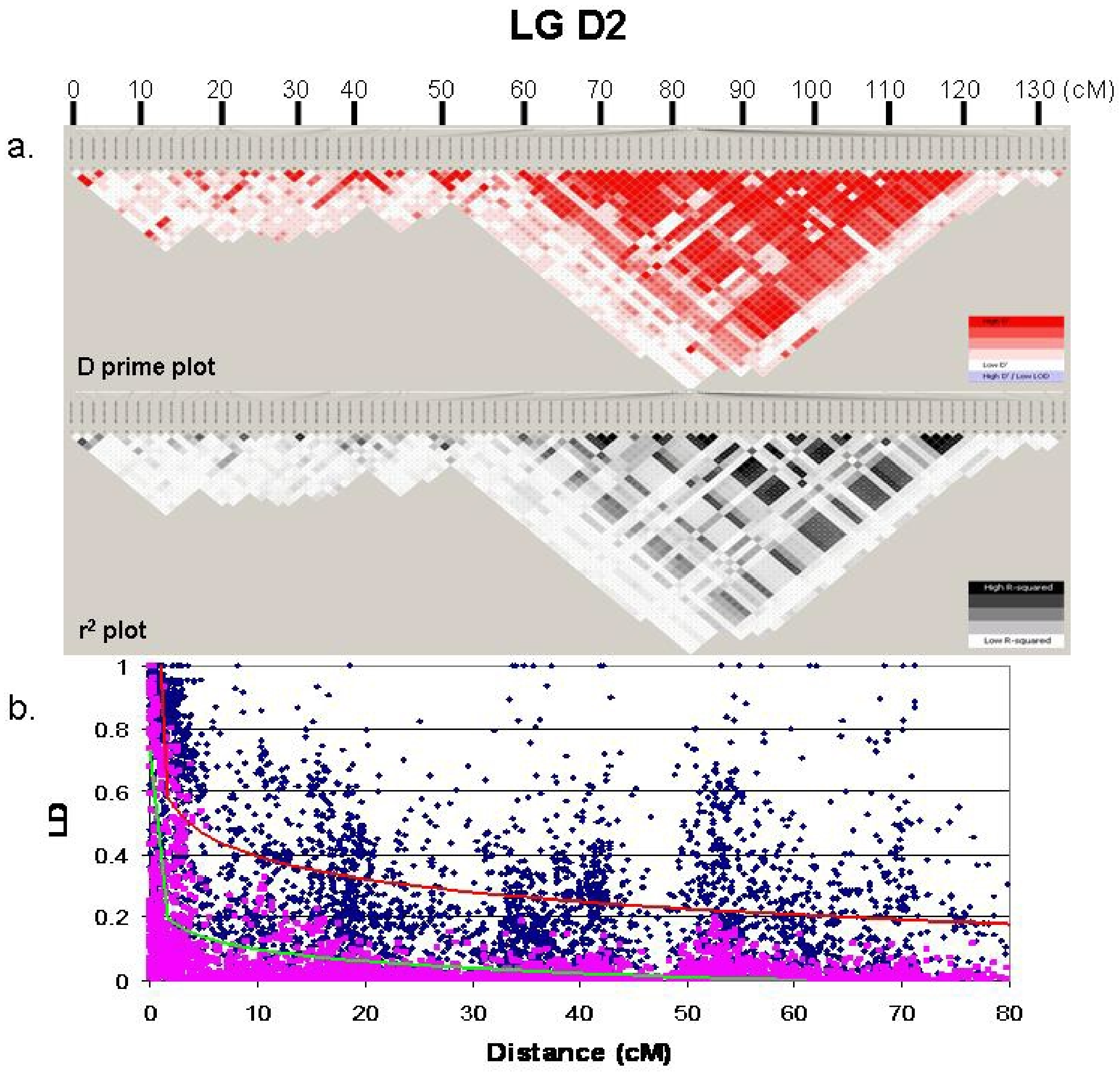

a.
LG C2
0 10 20 30 40 50 60 70 80 90 100 110 120 130 (cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
r²
1
0.8
0.6
0.4
0.2
0
0 10 20 30 40 50 60 70 80
Distance (cM)


LG D1a
a.
0
10
20
30
40
50
60
70
80
90
(cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

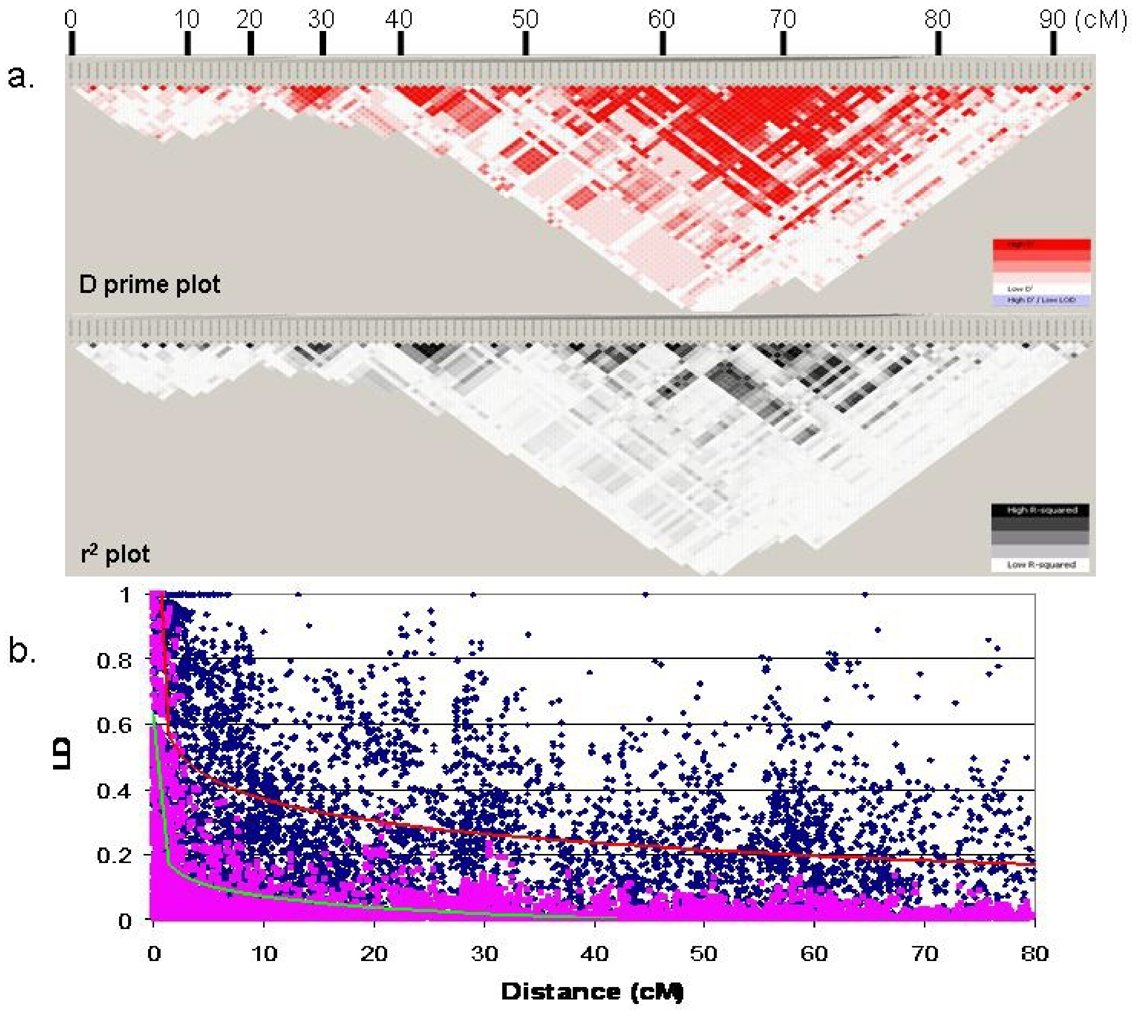


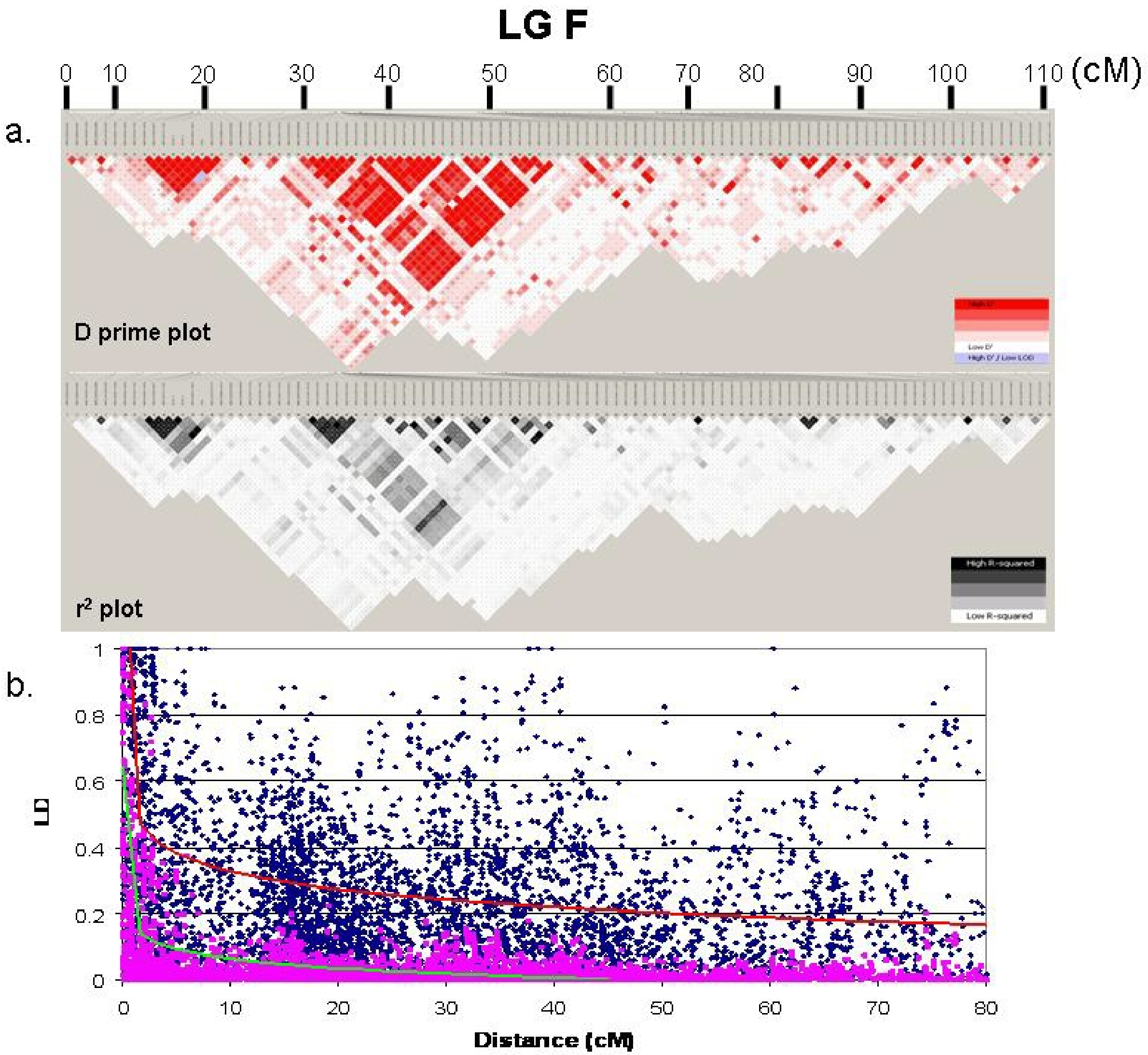
LG D1b

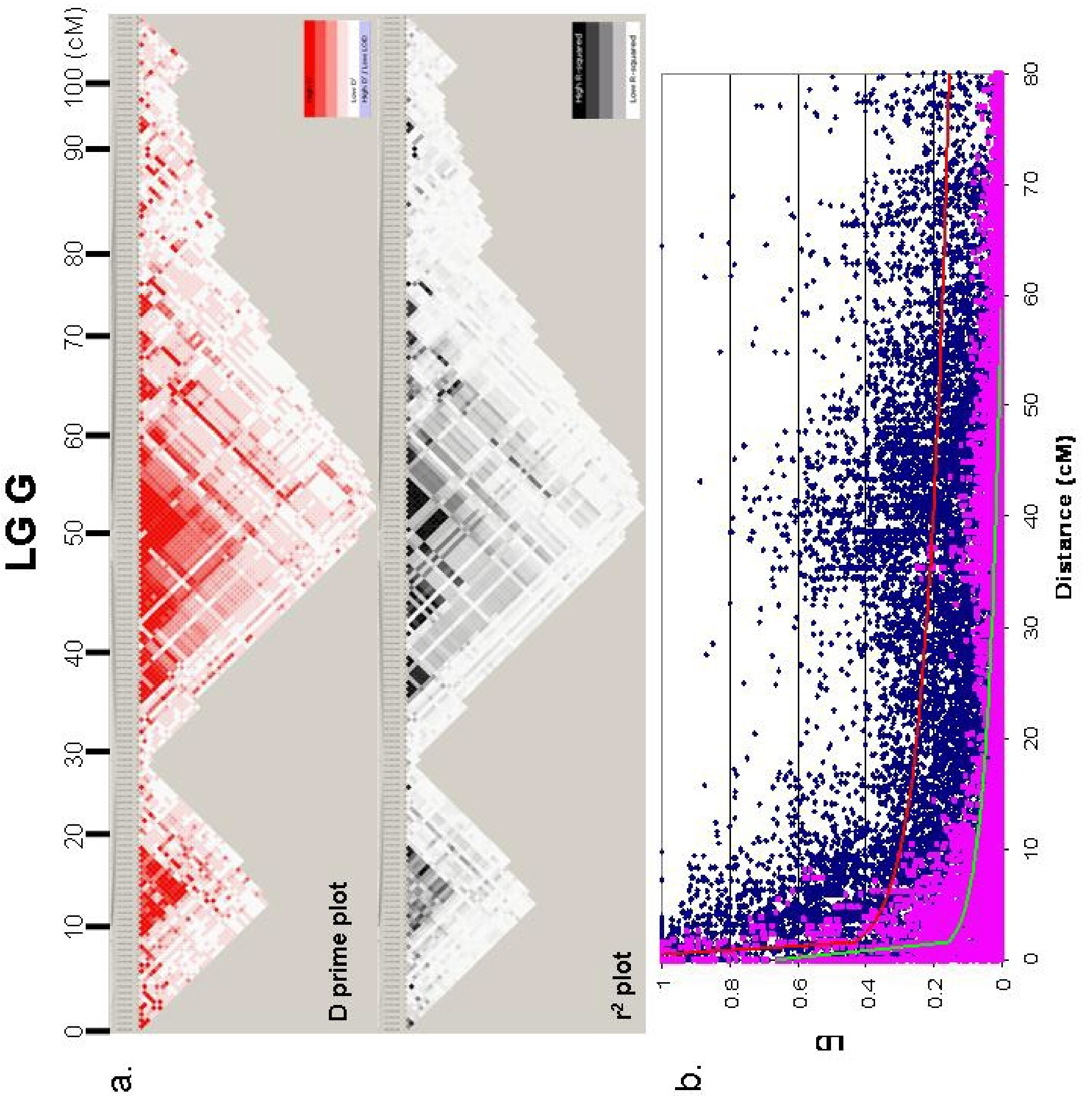
a.
0
10
20
30
40
50
60
70
80
90
100
110
120
130 (cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

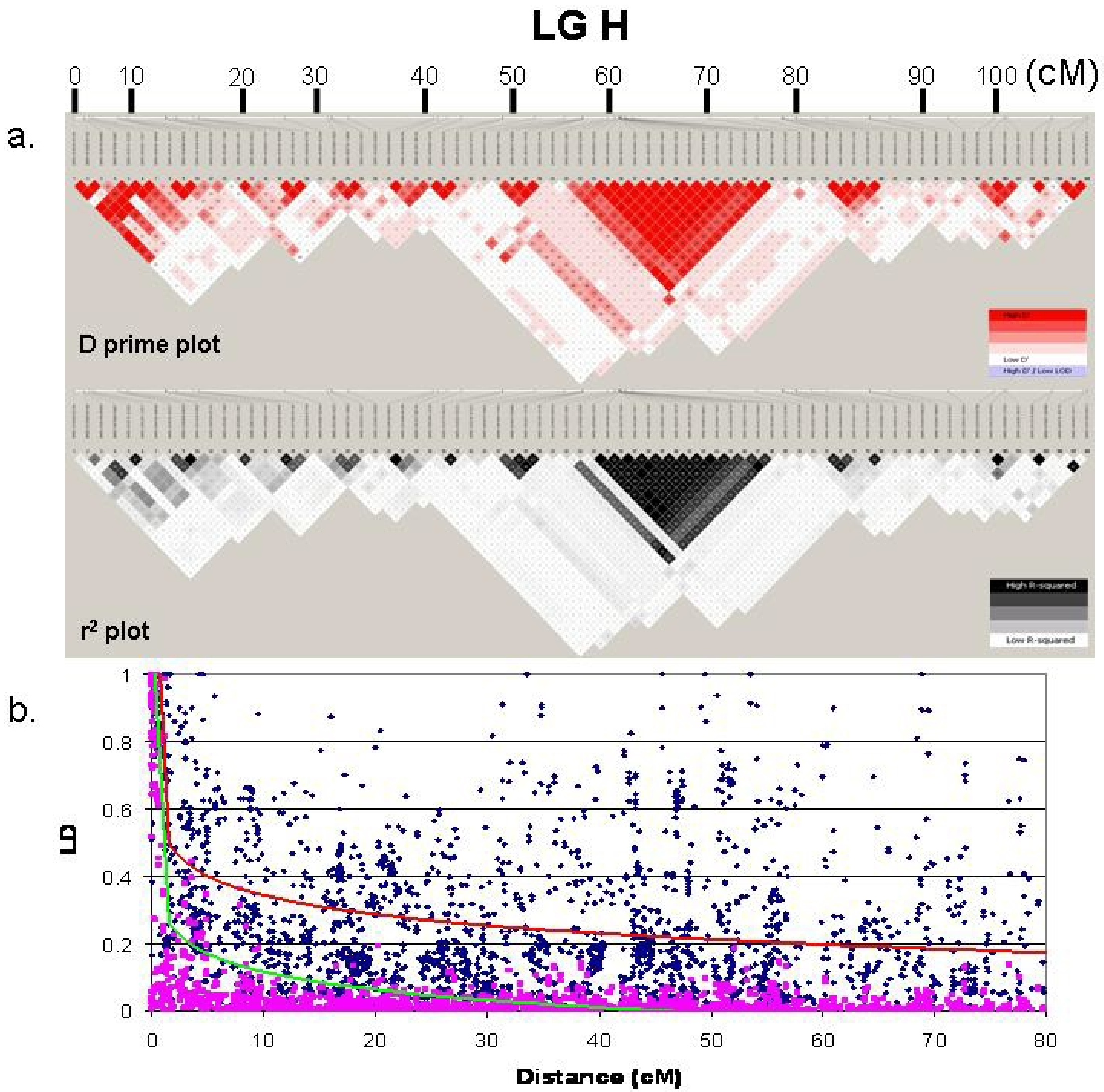

LG D2
a.
0
10
20
30
40
50
60
70
80
90
100
110
120
130 (cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r2 plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

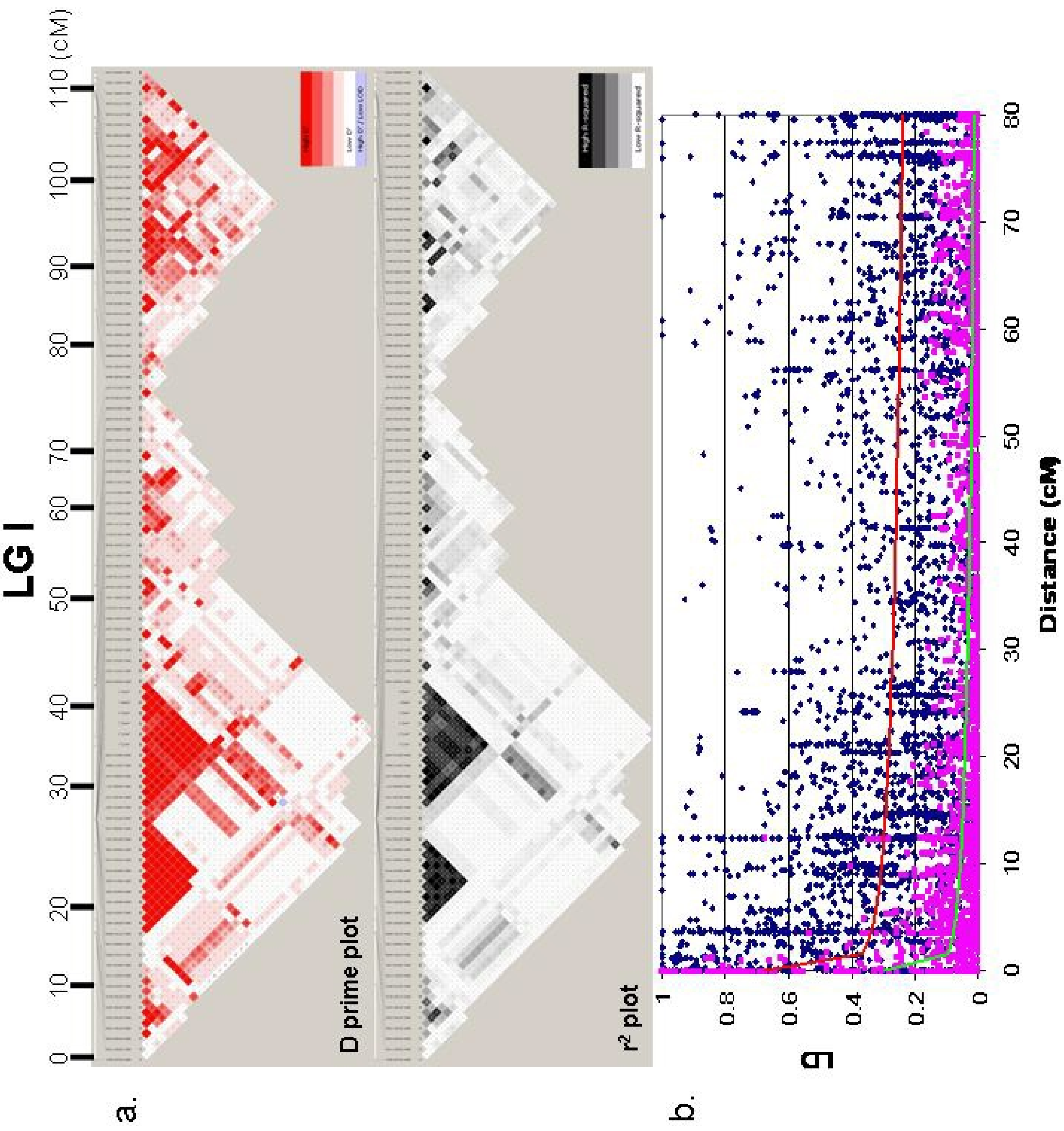
LG E

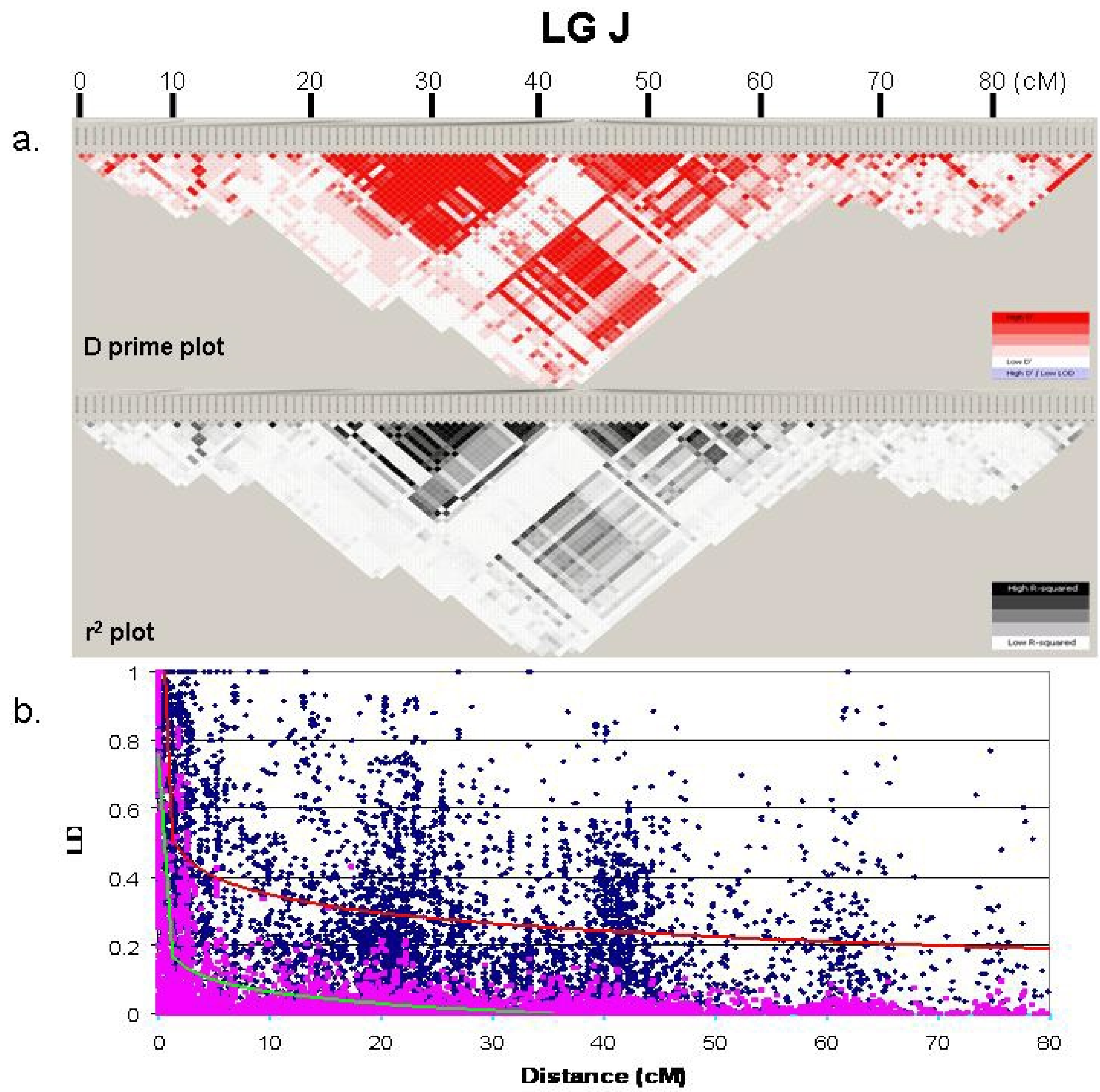

a.
0
10
20
30
40
50
60
70
80
90 (cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

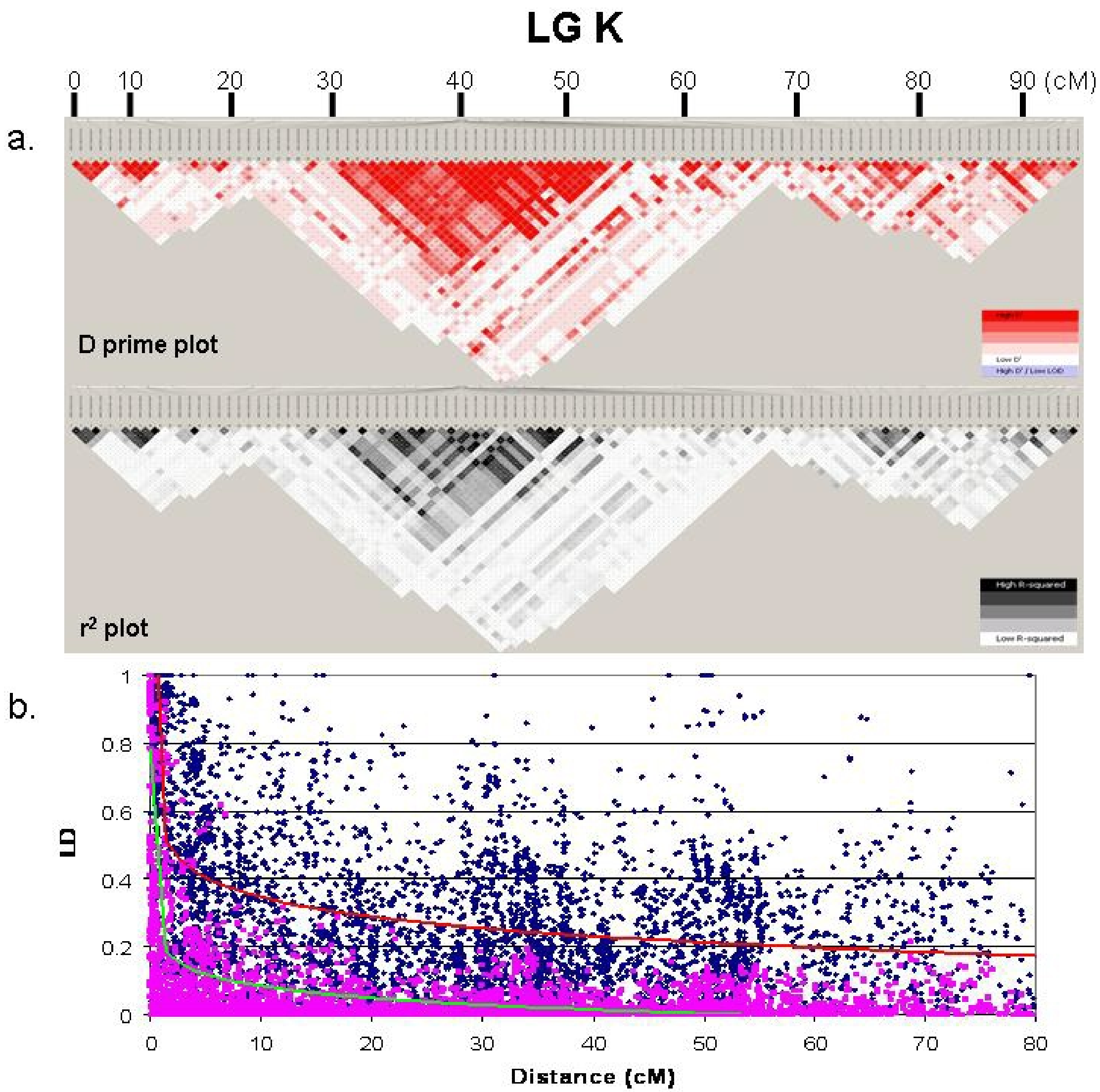

LG F
0
10
20
30
40
50
60
70
80
90
100
110
(cM)
a.
D prime plot
High D'
Low D'
High D' / Low LOD
r^2 plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

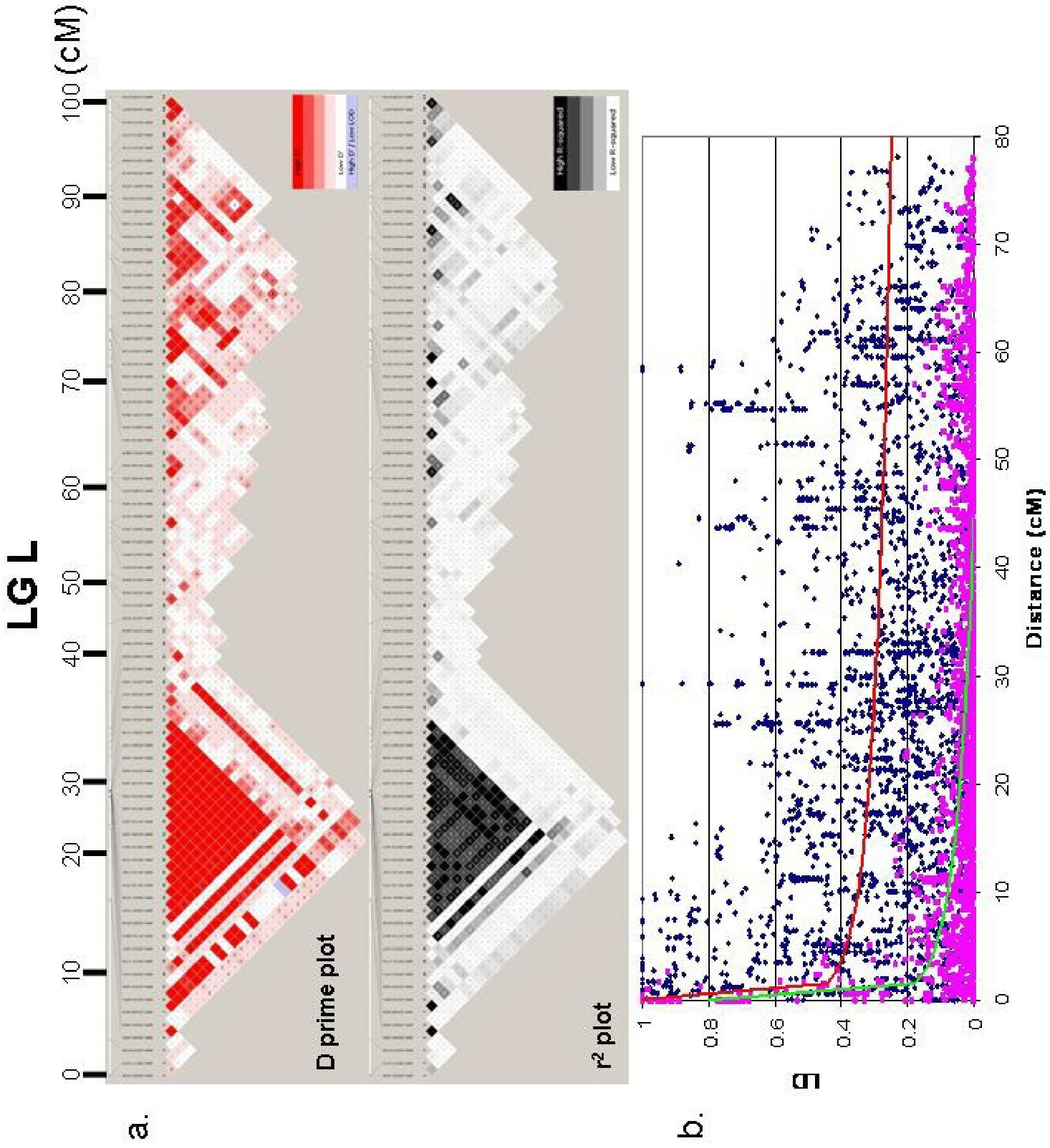

LG G
a.
0
10
20
30
40
50
60
70
80
90
100 (cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
r²
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

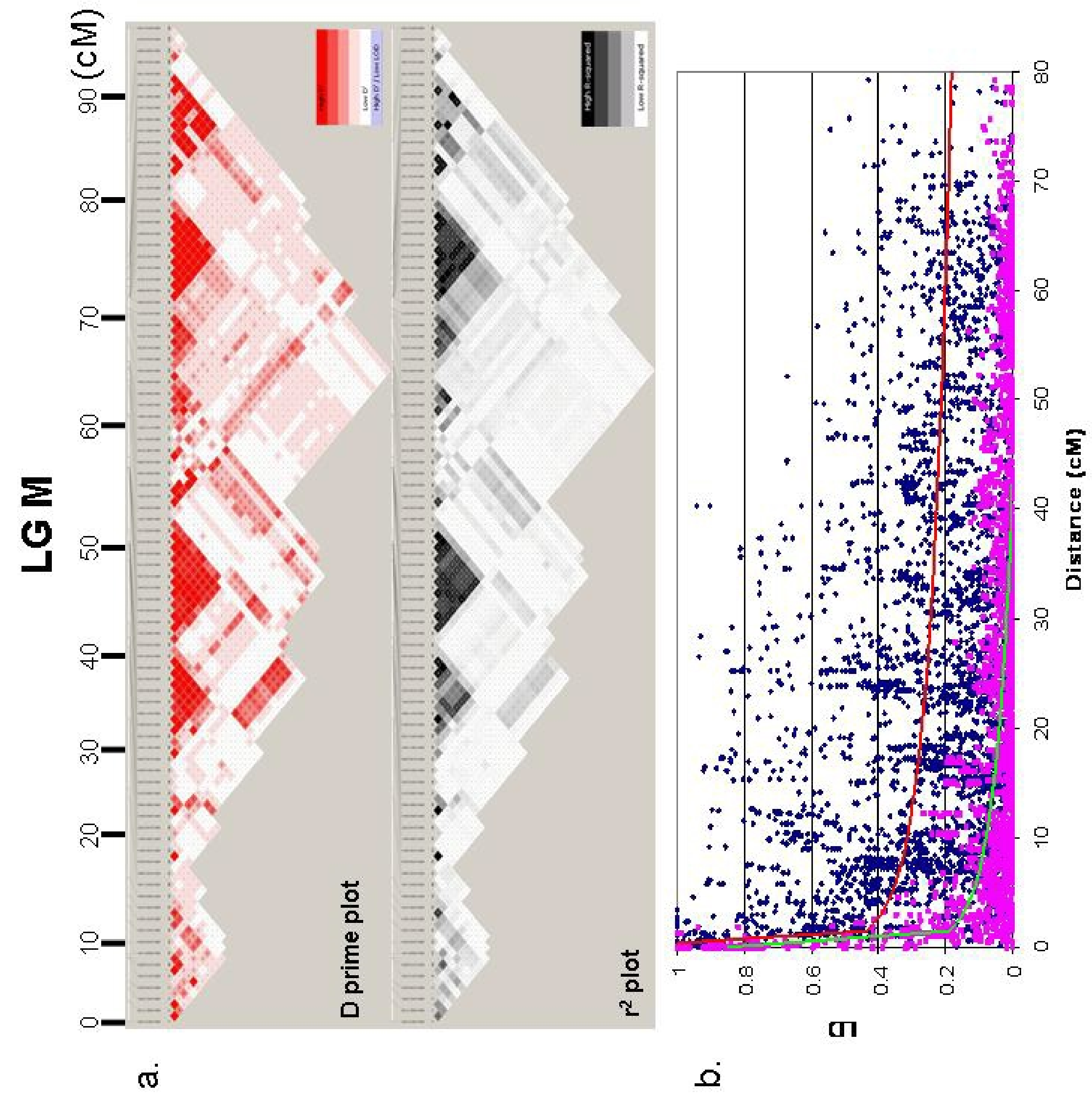

LG H
0
10
20
30
40
50
60
70
80
90
100
(cM)
a.
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

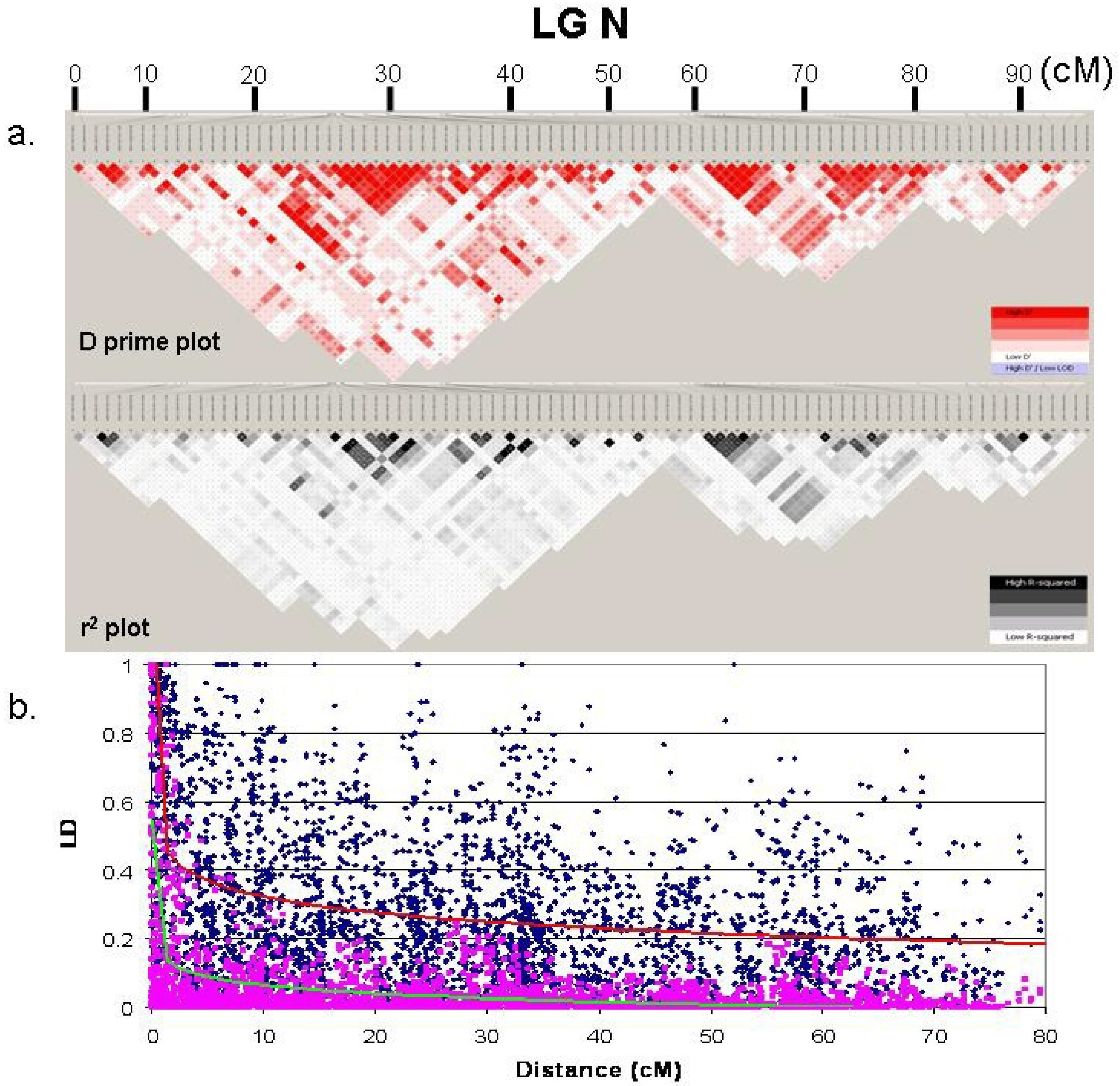

a.
b.
LG I
0
10
20
30
40
50
60
70
80
90
100
110 (cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
r²
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)


LG J
0
10
20
30
40
50
60
70
80 (cM)
a.
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)


LG K
a.
0
10
20
30
40
50
60
70
80
90 (cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r2 plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)


LG L
a.
0
10
20
30
40
50
60
70
80
90
100
(cM)
D prime plot
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)


LG M
a.
0
10
20
30
40
50
60
70
80
90
(cM)
D prime plot
High D'
Low D'
High D' / Low LOD
r^2 plot
High R-squared
Low R-squared
b.
LD
0
0.2
0.4
0.6
0.8
1
0
10
20
30
40
50
60
70
80
Distance (cM)


LG N
0
10
20
30
40
50
60
70
80
90
(cM)
a.
D prime plot
High D'
Low D'
High D' / Low LOD
r2 plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

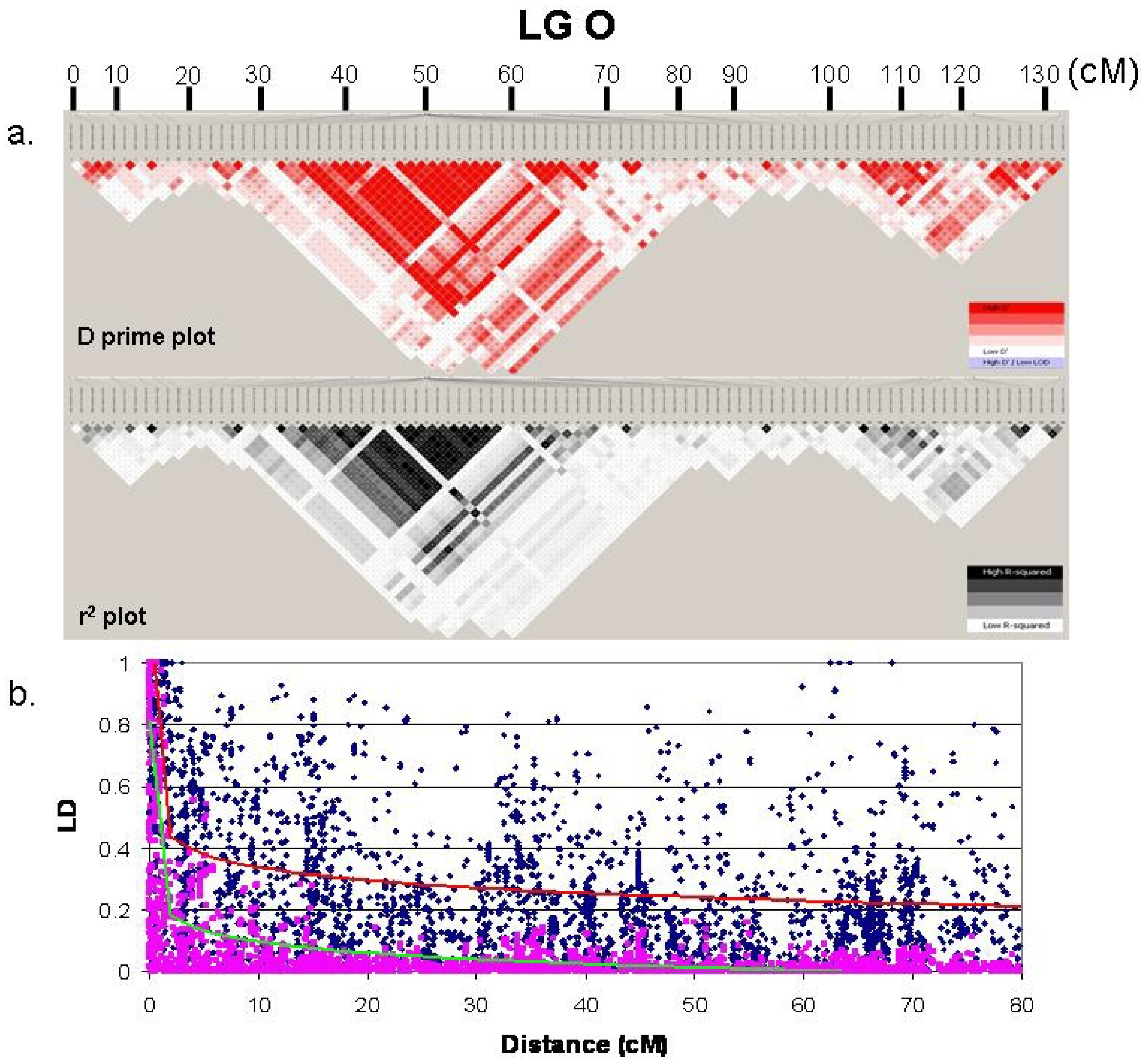

LG O
0
10
20
30
40
50
60
70
80
90
100
110
120
130
(cM)
a.
D prime plot
High D'
Low D'
High D' / Low LOD
r² plot
High R-squared
Low R-squared
b.
LD
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
Distance (cM)

Figure 2-2. The decline of LD on 20 linkage groups (LGs). a) The plots of *D'* and r^2 are constructed by a comparison of physically linked markers located within 20 cM of each other on each LG. The position of markers is indicated on the top of the LD plot. b) The decay of LD against the distance between two polymorphic sites. Red and blue dots indicate r^2 and *D'* values, respectively, between pairs of physically linked markers. Yellow and green lines indicate the logarithmic trend lines of *D'* and r^2, respectively

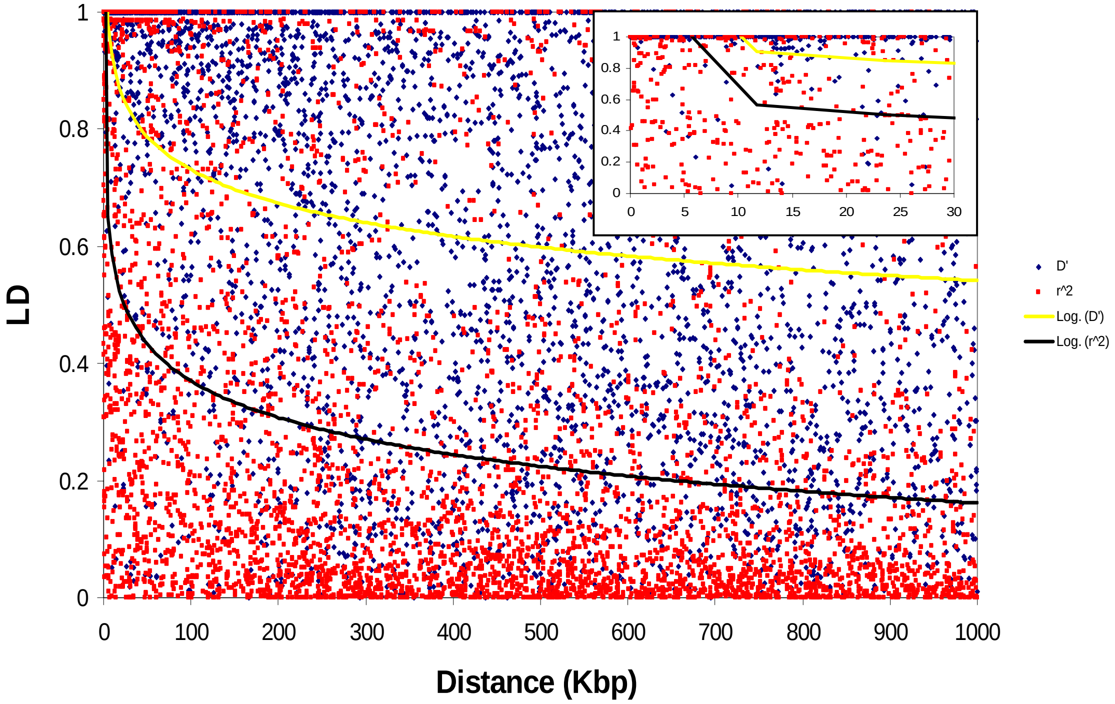


Figure 2-3. The level of LD within 1,000 Kbp in physical distance between pairs of loci. *D'* and r^2 are plotted against the distance between two markers. Blue dots indicate the D' values and red dots indicate r^2 values. The yellow and black lines are the logarithmic trend of D' and r^2, respectively. The small box on the top shows the LD decline (r^2) within 30 Kbp.

Williams 82 genome sequence spanning 609.5 Mbp which covered approximately 55.4% of the 1.1 Gbp of the soybean genome. The number of SNPs contained in scaffolds varied from 2 to 72 and the size of haplotype blocks was also highly diverse, ranging from 1 to 13,926 Kbp (Fig. 2-4). The Confidence Interval method yielded the longest estimate of average haplotype block size and the Four Gamete method gave the shortest estimate (Table 2-3). Of the 609.5 Mbp of genome scaffolds in which more than two SNPs were present, the genome regions for which haplotype block variation was defined ranged from 20.9% to 43.4% depending on the method used.

Population Structure

A total of 2,622 SNPs incorporated into 20 LGs of the 2008 Soybean Consensus Map were used to estimate the most likely number (*K*) of groups in the 319 germplasm accessions using the STRUCTURE program (Falush et al., 2003). The most likely *K* value was *K*=8 which was used for the analysis of the 319 germplasm accessions. Figure 2-5a shows the 319 individuals clustered into eight groups. The individuals within the eight groups generally had similar geographical origins. Most of the individuals within group one, two, three, five were of Chinese origin, the individuals within group four were of Korean origin, and the individuals within group seven were from Japan, while group six and eight were a mixed population of Chinese and Japanese origin. Groups one, two, three, and five, which were from China, were divided by maturity group, stem termination characteristics (determinate *versus* indeterminate), seed coat color, hilum color, and flower color

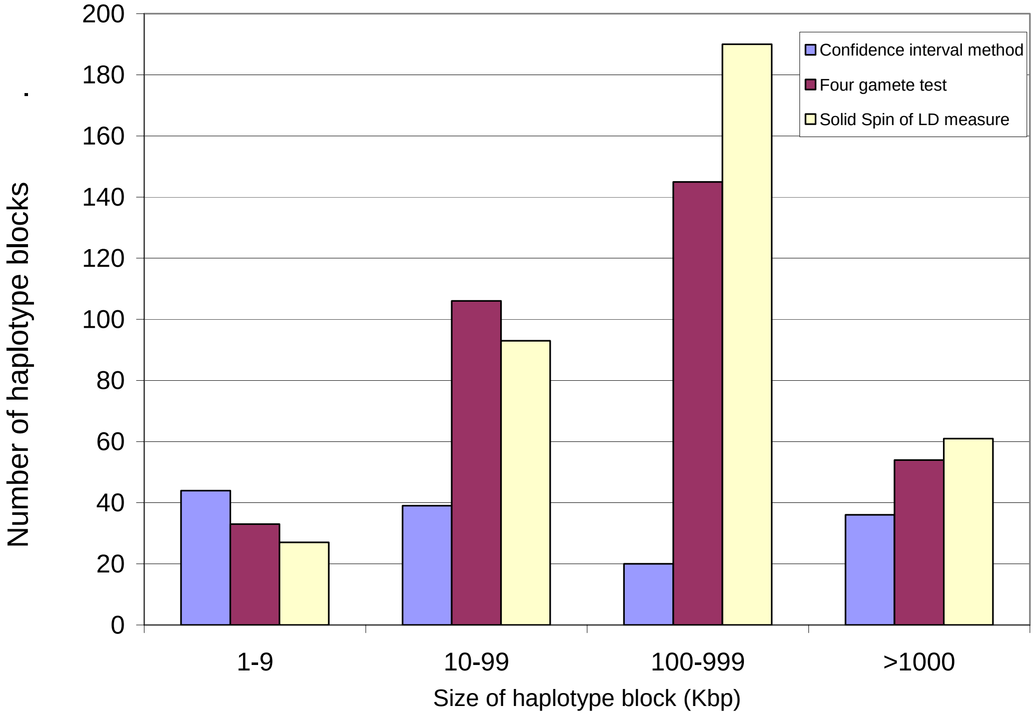


Figure 2-4. The average number and size of haplotype blocks estimated by different methods including the Confidence Interval method, the Four Gamete Rule method, and the Solid Spine of LD method. X-axis indicates the size of haplotype blocks estimated with the different estimators. Y-axis indicates the number of haplotype blocks determined by each method.

Table 2-3. Haplotype blocks in soybean germplasm accessions defined by three methods.

	Number of blocks	Block size (Kbp)	Percentage of genome covered by haplotype blocks (%)	Number of markers used (MAF>0.1)	Average block size (Kbp)	Average marker density (Kbp/SNP)
Confidence intervals	187	127,498	20.9	701	682	182
Four gamete rule	370	198,956	32.6	1,323	538	150
Solid spine of LD	396	264,656	43.4	1,324	668	200

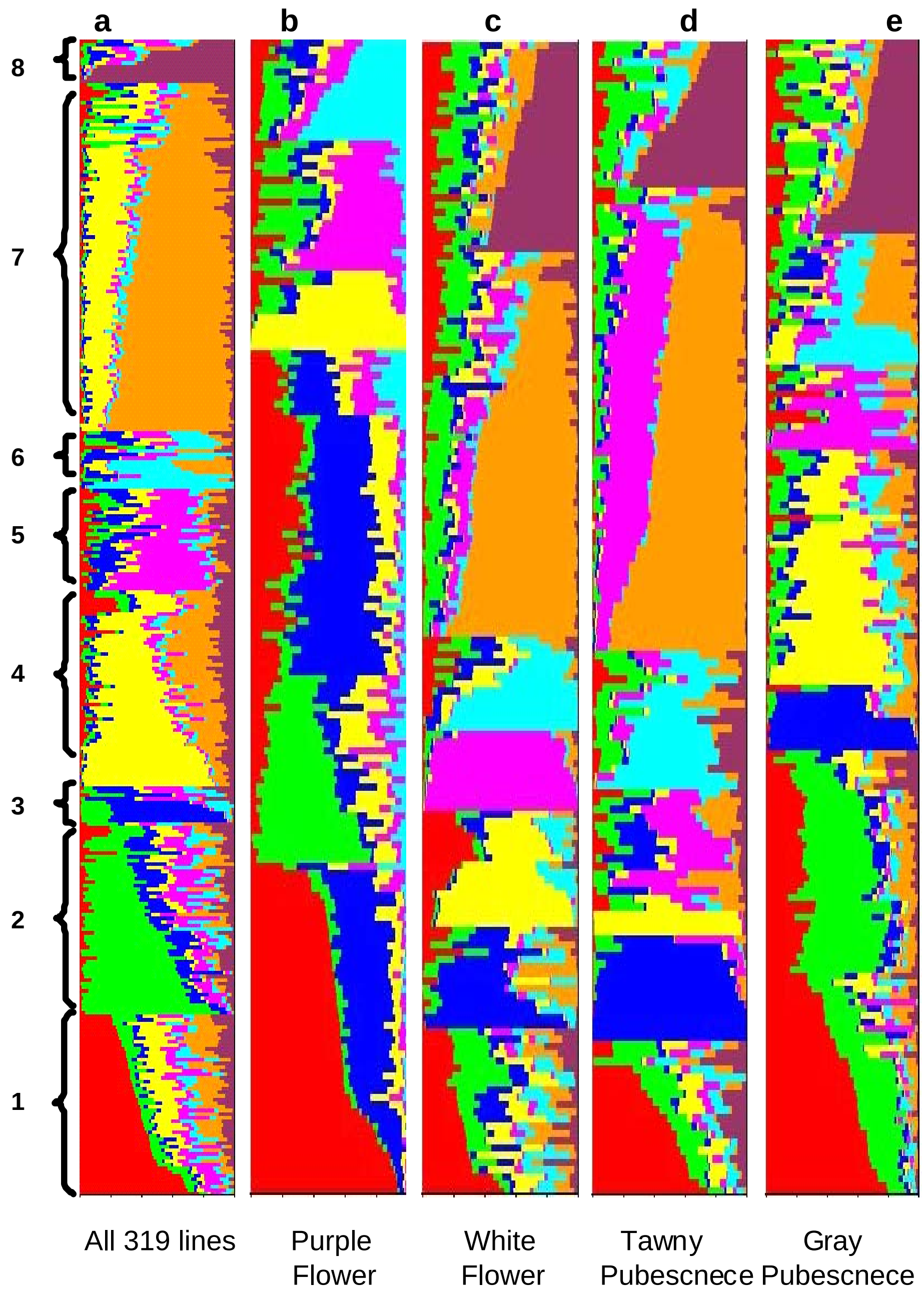


Figure 2-5. The presence of population structure in the 319 soybean germplasm accessions and in the groups with the different phenotypic traits: purple flower, white flower, tawny pubescence color, and gray pubescence color. 1~8 on the left side of the block a. indicates the most likely number of the sub-groups present in the 319 accessions. Blocks b-e show the presence of population structure in the sub-samples of the 319 germplasm lines of the different phenotypes.

(Fig. 2-5a). Figure 2-5b, c, d, and e showed population structure in groups with different phenotypes such as those accessions with purple flowers (Fig. 2-5b), white flowers (Fig. 2-5c), tawny pubescence (Fig. 2-5d), and gray pubescence (Fig. 2-5d). These results indicate the presence of highly diverse population structure.

Statistical Analysis to Reduce Type I error Caused by the Confounding Effects of Population Structure

The selection of an appropriate statistical analysis to reduce false associations caused by population structure is important for the success of association analysis. The simplest method to reduce Type I error resulting from the multiple comparisons used in association analysis is the Bonferroni correction. In this study, a total of 2,313 markers (MAF > 0.1) were analyzed with a Chi-square test to detect markers significantly associated with flower color and pubescence resulting in an adjusted P-value of 4.32×10^{-6} when the cutoff value for significance was set at 0.01. However, when the Bonferroni correction was applied to the results of the Chi-square test, there were 232 and 179 markers significantly associated with flower color and pubescence color, respectively. These results indicated that the Bonferroni correction alone was not appropriate to adjust for Type I error in this association analysis. Adjustment for population structure is a second approach to reduce Type I error resulting from the effects of population structure. There are several statistical analyses developed to adjust for population structure including logistic regression (LR) and the general-linear-model (GLM) which are incorporated with the structure matrix (Q), and the mixed-linear-model (MLM) (Yu and Buckler, 2006) which is incorporated with the structure (Q) and kinship matrix (K). To select the most appropriate method to adjust for population structure, the cumulative P-value was plotted against the expected P-value that resulted from association analysis for flower color and pubescence color from each procedure used to adjust for population structure. Figure 2-6 showed the cumulative distribution of P-values of the results of association analysis. By definition, when the rate of false positives is reduced to zero (null hypothesis), a cumulative P-value should be the same value as the expected P-value and would result in the straight line (Black line in Fig. 2-6). With less than complete adjustment, the plot deviates from the straight line (Schweder and Spjotvoll)1982). In Fig. 2-6a and b, the distributions of cumulative P values were strongly skewed toward high significance in the Chi-square test (naïve test) while all of the other three adjustment methods, LG+Q, GLM, and MLM, showed different degrees of improvement *vs*. the naïve test in terms of reducing Type I error rate. The different adjustments for population structure reduced the false positive rate, but none completely adjusted for the effects of population structure.

Whole-Genome Case-Control Analysis for the Detection of Genes Controlling Flower Color and Pubescence Color

A whole-genome case-control analysis was attempted in an effort to detect the single major genes that underlie flower color and pubescence color. The result of whole-genome case-control analysis for the gene controlling flower color produced many associations when the naïve test (Chi-square test) was applied (Fig. 2-7a). However, after the adjustment for population structure using a mixed-linear-model, only nine SNPs on LG F remained strongly associated (Fig. 2-7b). Of these markers

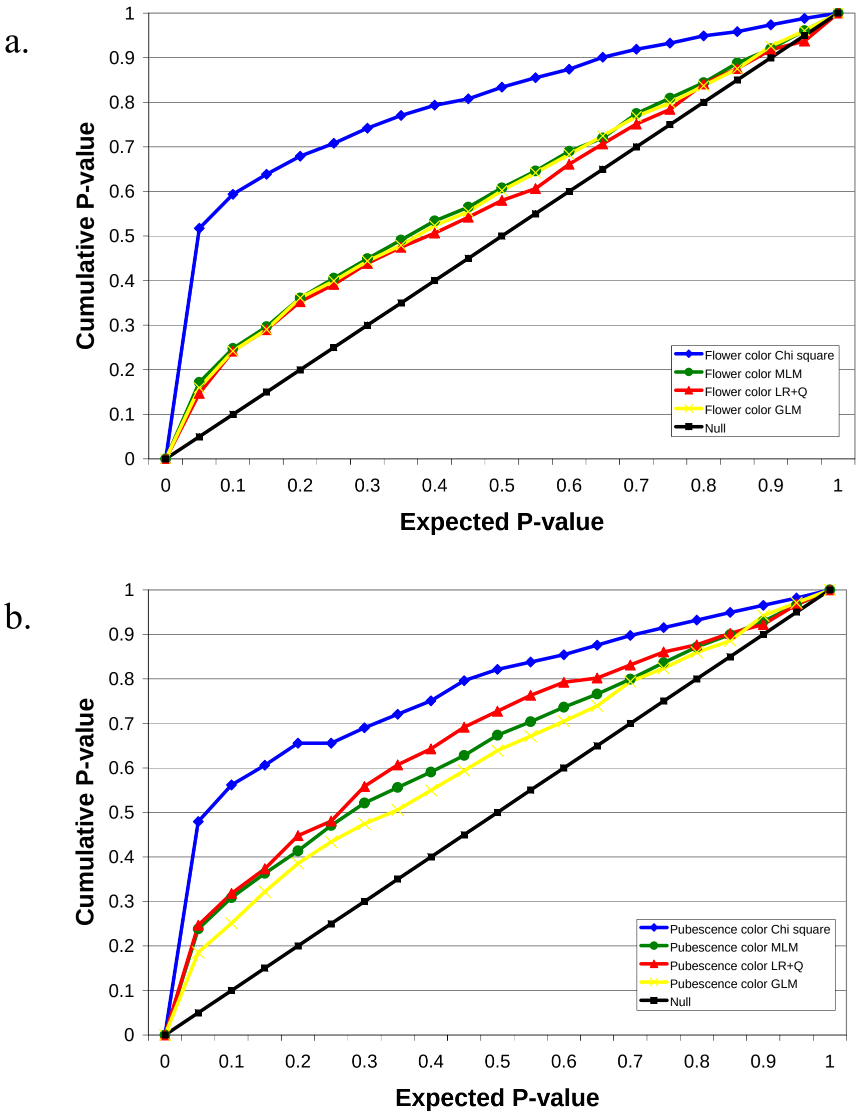


Figure 2-6. The cumulative distribution of P-value calculated with the result of genome-wide association analysis for flower color (upper panel) and pubescence color (lower panel). X-axis indicates the expected P-value and Y-axis indicates the cumulative P-value. The lines with the different colors are the result of the association analysis with a different statistics. Blue line: naïve test, green line: a mixed-linear model incorporated with structure and kinship matrixes, red line: logistic regression with structure matrix, yellow line: generalized liner model with structure matrix, and black line: the hypothetical line when type I error is eliminated.

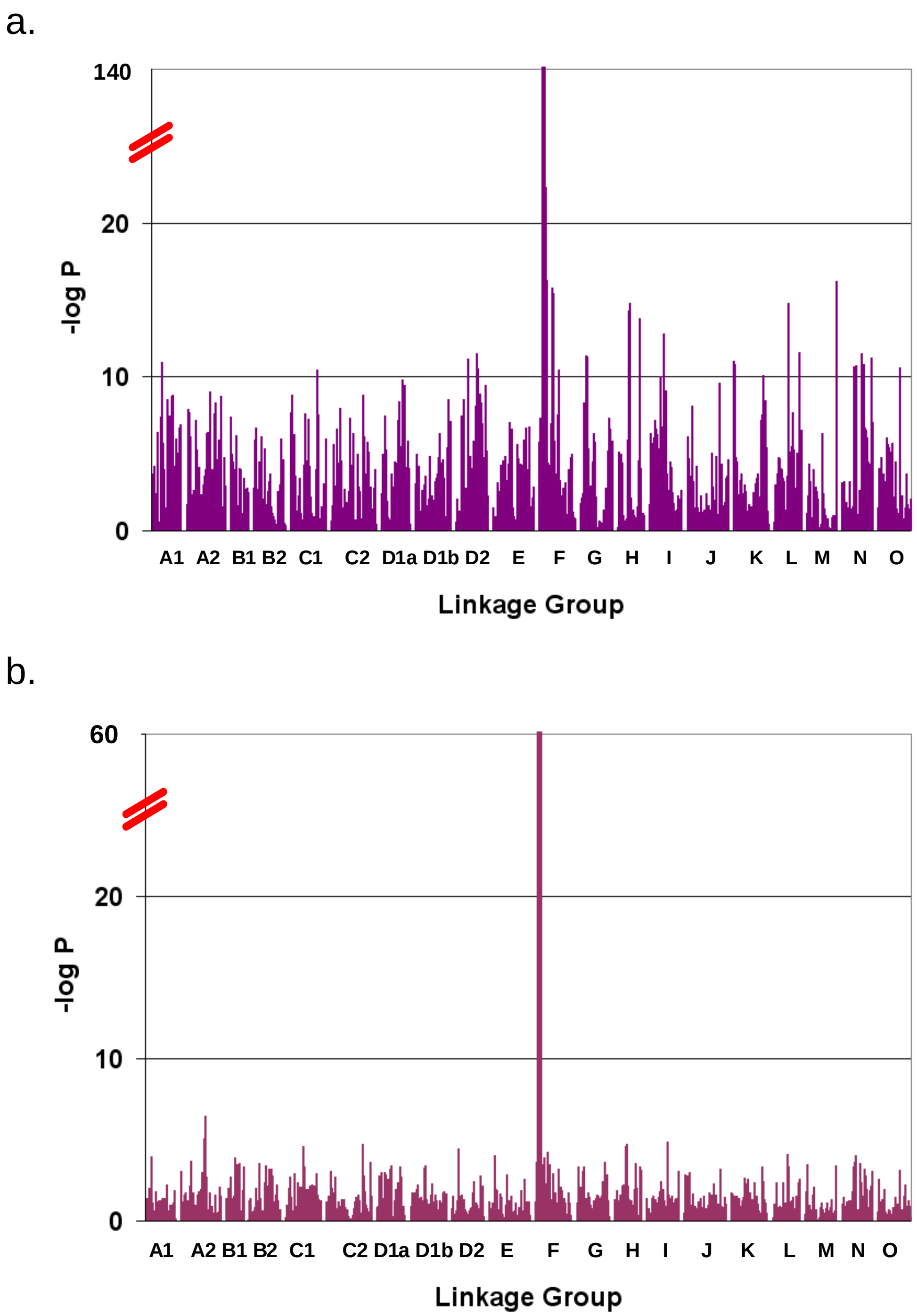
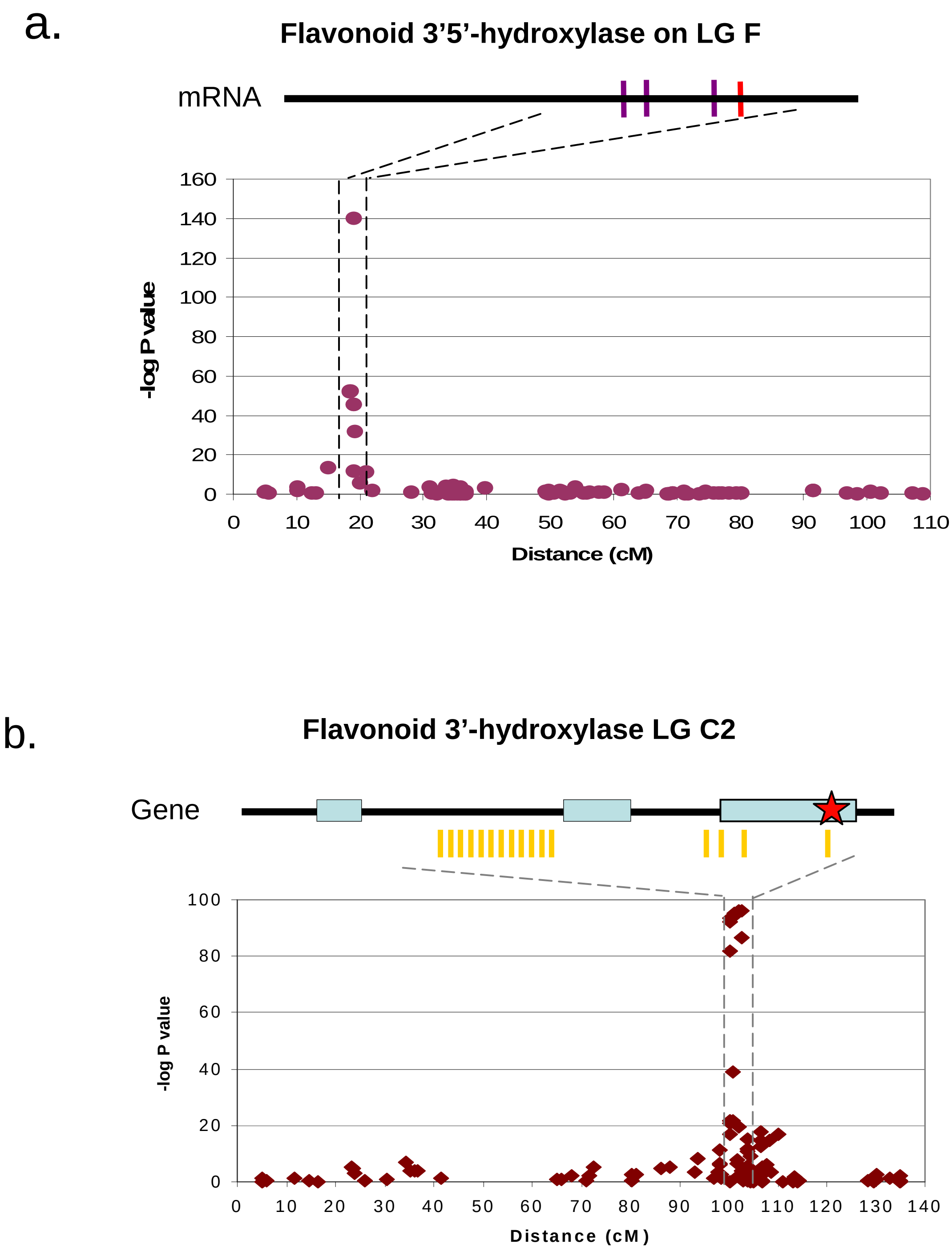


Figure 2-7. Genome-wide association analysis for genes controlling flower color. a is the result of naïve tests (no adjustment for population structure) (Chi-square test) for flower color and b is the result after the structure adjustment using a mixed-linear-model. The height of the peak indicates the degree of association. X-axis indicates all the markers on each linkage group and Y-axis indicates the negative log P-values.

that were significantly associated with flower color, three markers were developed from BAC-end sequence that included BARC-064051-18538, BARC-051237-11031, and BARC-051955-11307 and two markers, BARC-016463-02617 and BARC-043267-08567, were developed from expressed sequence tags. In addition, four markers developed from the *GmF3'5'H* (AY117551) showed significant associations with flower color including 66191_4 , BARC-066191-19815, BARC-035375-07174, and a 65 bp insertion (Fig. 2-8a and fig. 2-9a). However, none of other markers showed the significant level as high as the 65 bp insertion. The 65 bp insertion in the *GmF3'5'H* gene (AY117551) truncates translation prematurely resulting in the white flowered phenotype (Zabala and Vodkin, 2007). All of the nine markers that showed significant association with flower color were mapped to the 18-20 cM region on LG F where the *F3'5'H* gene is located (Fig. 2-8a) and they were associated with sequence scaffold, super_176 (http://www.phytozome.net/soybean). Four markers (BARC-066191-19815, 66191-4, BARC-035375-07174, and 65 bp insertion) developed from the *GmF3'5'H* gene were in high LD as indicated by *D'* but the 65 bp insertion was in relatively low LD with the three SNPs as measured by r^2 (Fig. 2-9a). Although they were located within 3 Kbp of the 65 bp insertion (Fig. 2-9b), the lack of associations between the three SNP markers and flower color was easily explained based on the low level of r^2 between them and the 65 bp insertion.

In the case of pubescence color, the adjustment for population structure also reduced the rate of spurious associations (Fig. 2-10b) compared with the results of the naïve test (Fig. 2-10a) in the whole-genome case-control analysis.


a.
Flavonoid 3’5’-hydroxylase on LG F
mRNA
-log P value
0
20
40
60
80
100
120
140
160
0
10
20
30
40
50
60
70
80
90
100
110
Distance (cM)
b.
Flavonoid 3’-hydroxylase LG C2
Gene
-log P value
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
Distance (cM)

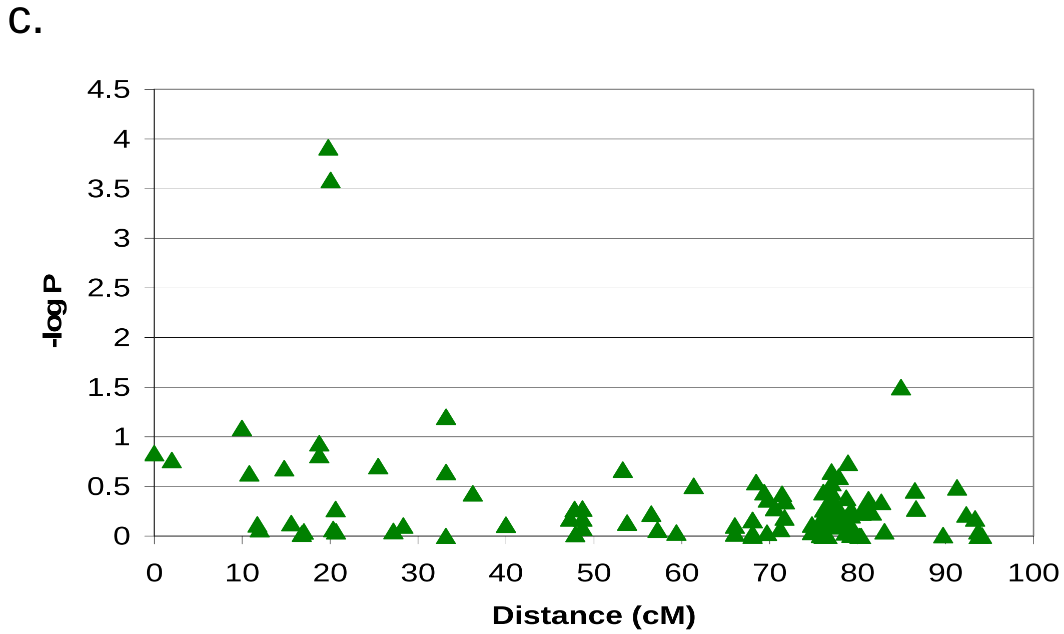


Figure 2-8. Association analysis of flower color, pubescence color and seed protein. The most significantly associated markers with flower color on LG F (a), pubescence color on LG C2 (b), and seed protein concentration on LG E (c). Red bar and dot in a. indicates the 65 bp insertion and star in b. indicates the 'C' deletion. X-axis indicates the genetic position of the markers and Y-axis indicates the level of significance of association (negative log P-values).

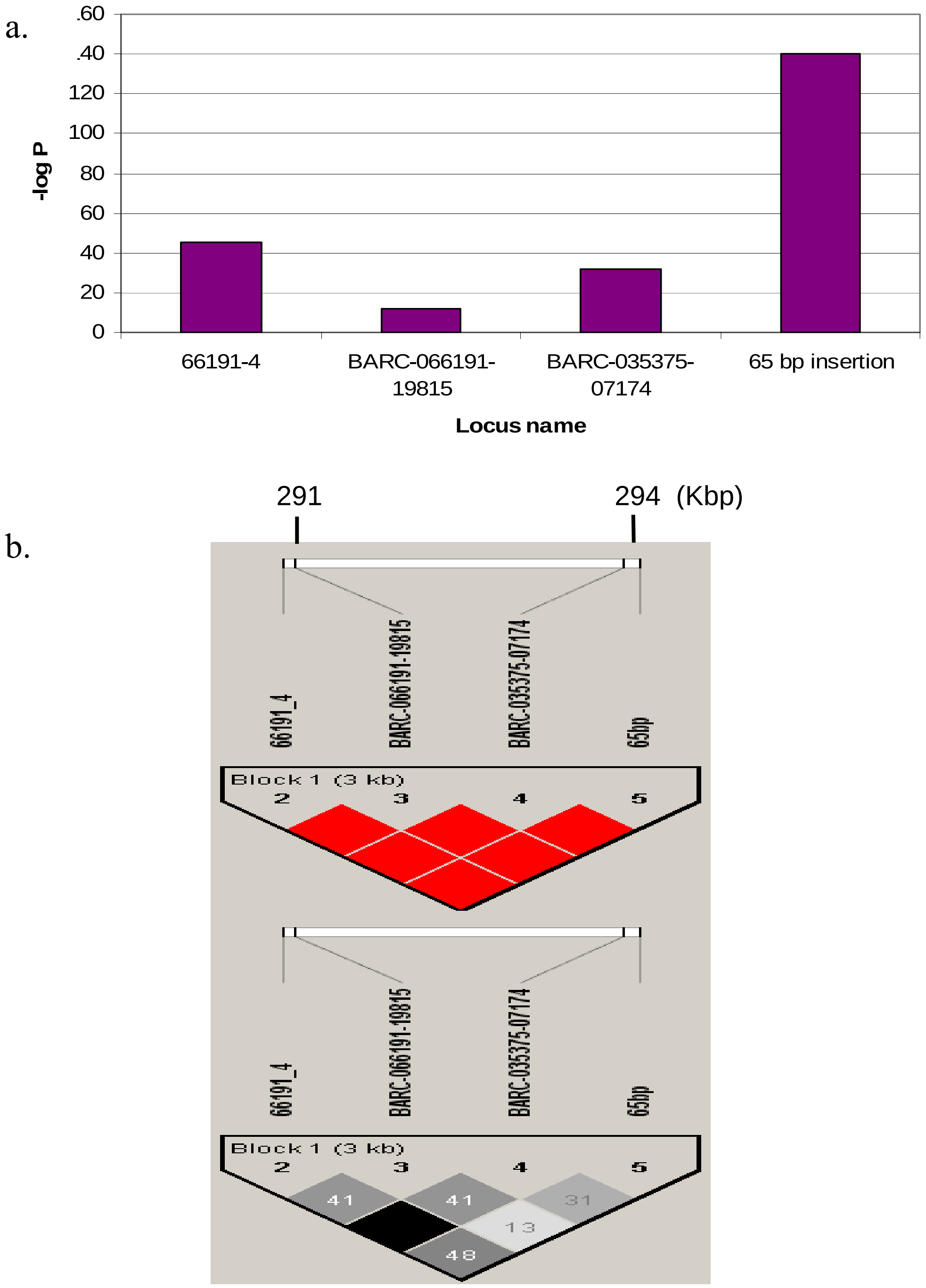


Figure 2-9. The degree of association of the SNP markers developed within the *GmF3'5'H* gene for pubescence color. a shows the level of significant ($P<0.001$) of the four markers developed from the *GmF3'5'H* gene. b shows the level of LD among the SNP markers in the red color plot (*D'*) and gray color plot (r^2). Values are expressed as $D'\times100$ and $r^2\times100$. The number on the top of the LD block indicates the order of the markers which have a minor allele frequency>0.10.

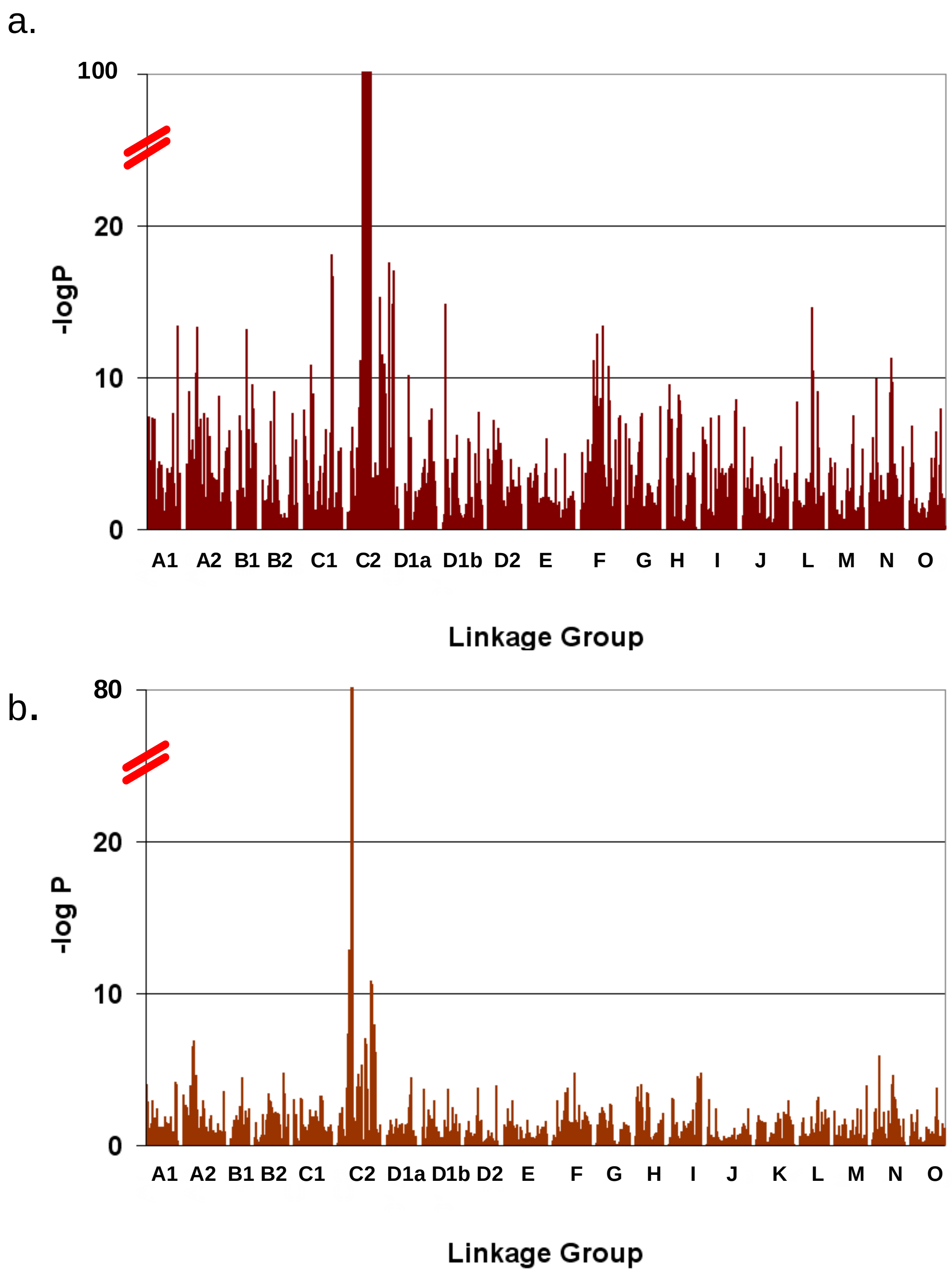


Figure 2-10. Genome-wide association analysis for gene controlling pubescence color. a is the result of naïve test (no adjustment for population structure) (Chi-square test) and b is the result after the structure adjustment using mixed-linear-model. The height of the peak indicates the degree of association. X-axis indicates all the markers on each linkage group and Y-axis indicates the negative log P-values.

One marker developed from a BAC-end sequence (BARC-057519-14781), positioned at 101 cM on LG C2, showed strong association with pubescence color. This marker, however, was not associated with any sequence scaffold of the Williams 82 genome. In addition, there were 13 SNPs developed from the exon and intron regions of the *F3'H* gene (AB191404) of which seven showed significant associations with pubescence color including 066187-6, 066187-4, 066187-1, 066189-16, BARC-066175_3-19800, BARC-012755-00393, and 066177_1 (Fig. 2-11). Of these seven markers developed from the *F3'H* gene, five markers, 066187-6, 066187-4, 066187-1, 066189-16, and BARC-066185_3-19800, were located in the intron and the other two markers, 066177-1 and BARC-012755-00393, were located in the 3' end of the exon. BARC-012755-00393 is a synonymous mutation (does not change amino acid sequence). Marker, 066177-1, is the C deletion that was reported as the determinant of pubescence color alteration from tawny to gray. This deletion produces a stop codon that prematurely terminates transcription of the *F3'H* gene (Toda et al., 2002). All of these markers were mapped to the 100-105 cM region on LG C2 (Fig. 2-8b) where the candidate gene region controlling pubescence color is located and all were associated with the sequence scaffold, super_82 of the Williams 82 whole-genome sequence. Figure 2-11 shows the magnitude of the significance of association of the 14 markers developed from the *F3'H* gene as well as the LD between these markers. Seven markers showing the highest associations with pubescence color were in high LD as indicated by r^2 but the other seven markers were not (Fig. 2-11). However, none of the markers, including the C deletion completely explained tawny *vs*. gray pubescence color in the entire set of 323 genotypes examined in this study. There

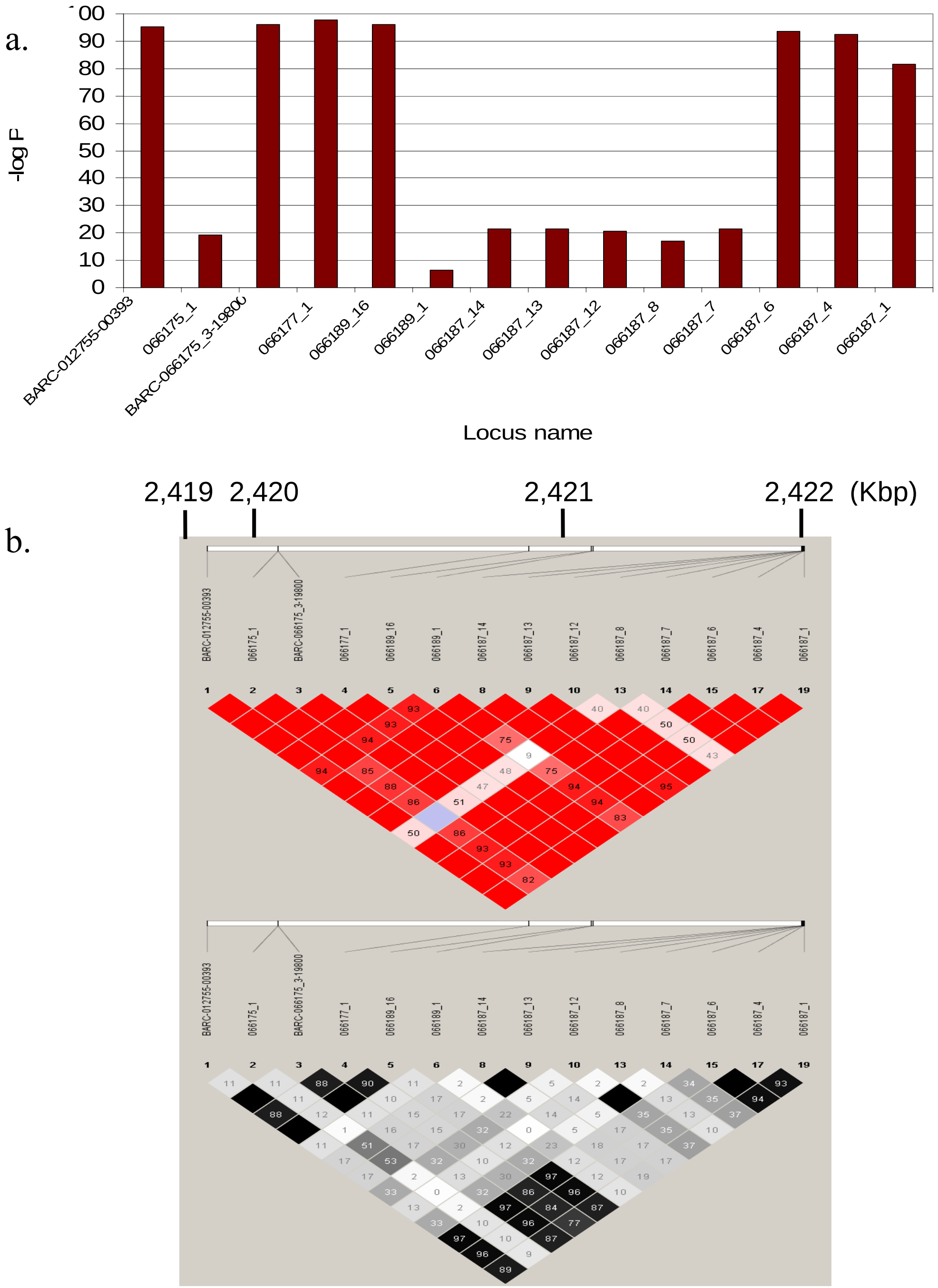


Figure 2-11. The degree of association of the SNP markers developed within the *F3'H* gene for pubescence color. a shows the level of significant (P<0.001) of the 14 markers developed from the *F3'H* gene. b shows the level of LD among the SNP markers in the red color plot (*D'*) and gray color plot (r^2). Values are expressed as *D'*×100 and r^2×100. The number on the top of the LD block indicates the order of the markers which have a minor allele frequency>0.10.

were 28 gray pubescence lines that did not contain the "C" deletion in the *F3'H* (AB191404) gene. The gray pubescence lines with the C deletion have one haplotype while three different haplotypes (MAF > 0.1) were present among the tawny pubescence lines. In addition, the 24 gray accessions in which the C deletion was absent had different haplotypes from the tawny pubescence lines (MAF > 0.10).

Whole-Genome Regression Analysis for the Genes Controlling a Quantitative Trait, Seed Protein Concentration

Seed protein concentration of 252 germplasm accessions with yellow seed coat color was examined over two locations in 2003. The genotypes were selected based upon data in the USDA Germplasm Resources Information Network (GRIN) database to represent two distinct groups of germplasm accessions – one group with normal (40 to 43%) seed protein concentration and the other with high (46 to 51%) seed protein concentration. However, seed protein concentration as measured in seeds grown in replicated hill-plot trials at Beltsville, MD and Linclon, NE showed continuous variation with a range from 40 to 47% (Fig. 2-12). The coefficient of variation (CV) of protein concentration was 2.18% in the experiment performed in MD and 2.30% in the NE experiment. The overall CV based on the analysis of the two locations was 2.21%. The results of analysis of variance showed that there were significant differences (P < 0.0001) in seed protein concentration among the accessions, but no significant difference between replications, and also no significant interaction between accessions and locations (Table 2-4). The correlation coefficient (r) of seed protein concentration between the MD and NE experiments was 0.863 (P < 0.0001). The correlation of the mean protein concentration measured in the NE

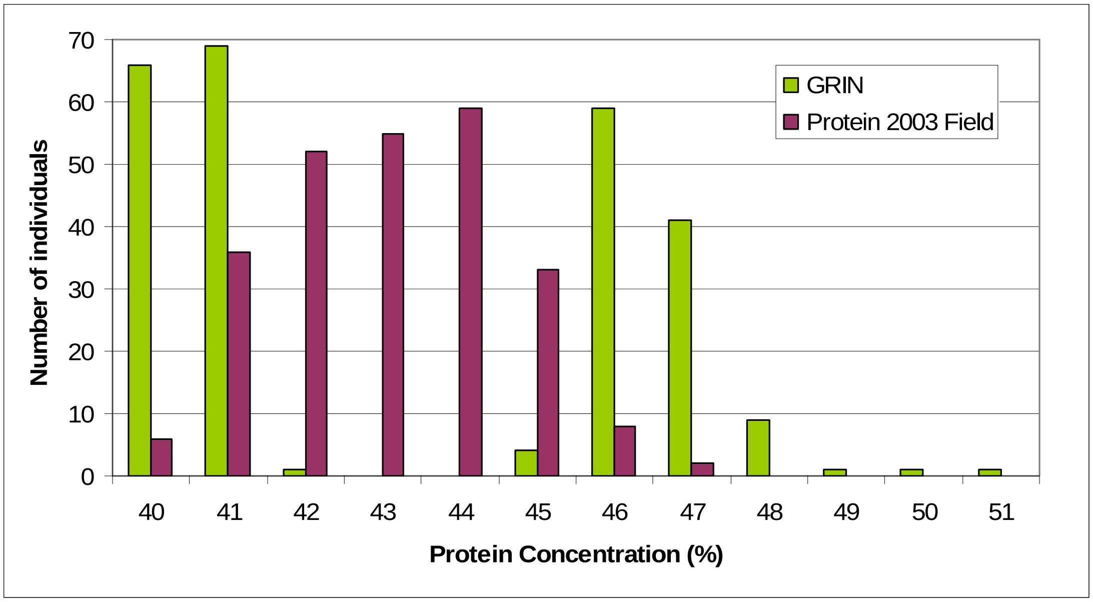


Figure 2-12. The comparison of seed protein concentration from the GRIN database *vs*. the 2003 field data in 252 accessions with yellow seed coat color. Green bars are seed protein concentration from the GRIN database and purple bars are seed protein concentration from the 2003 field test. X-axis indicates seed protein concentration. Y-axis indicates the number of individuals in a range of corresponding seed protein concentration.

Table 2-4. Analysis of variance of seed protein concentration of 252 accessions with yellow seed coat color based upon the 2003 field tests in Maryland and Nebraska.

Source	DF	SS	Mean Square	F Value	Pr > F	
Entry	251	2472.91	9.8522	14.09	<.0001	$\sigma^2+4.74\sigma^2_{(Entry)}+ r\sigma^2_{(Entry*location)}$
Location	1	0.29	0.2978	0.43	0.4543	$\sigma^2+rv\ \sigma^2_{(location)}+\ r\sigma^2_{(Entry*location)}$
Entry*Location	247	172.34	0.6989	0.75	0.9962	$\sigma^2+ r\sigma^2_{(Entry*location)}$
Rep(location)	4	3.37	0.8439	0.91	0.4570	$\sigma^2+ v\sigma^2_{(Rep(location))}$
Error	864	800.67	0.9267			σ^2

*r, and v are number of replication(location) and number of entries.
$\sigma^2_{(Entry)}$= (9.852232-.6989)/4.74=1.9310
Heritability(%)=1.9310/(1.9310+0.9267)=67.60

experiment with the GRIN protein concentration data was 0.653 (P < 0.0001). The GRIN data *vs*. the MD experimental data had correlation of r=0.657 (P < 0.0001). The mean of seed protein concentration of the NE and MD experiments *vs*. the GRIN data had a correlation of r=0.653 (P < 0.0001). The correlation of seed size and seed oil concentration *vs*. seed protein concentration was 0.1105 (P<0.0001) and 0.2947 (P<0.0001), respectively. The heritability of seed protein concentration was 67.6%. For the estimation of the heritability of seed protein concentration, both entries and locations were considered to be random effects. Table 2-3 lists the four candidate regions of seed protein QTLs on LGs A1, E, I and M. The results of whole-genome regression analysis for seed protein concentration QTL is shown in Fig. 2-13a and b. After the correction for population structure using the mixed-linear-model approach (Fig. 2-13b), many of the associations detected by the naïve test were eliminated or their significance reduced. There remained two genome regions with markers showing strong associations (P<0.001) with seed protein concentration. These were on LGs E, and K (Fig. 2-13b). The two most significant markers associated with seed protein concentration were BARC-016533-02084 (P<0.001) and BARC-042349-08247 (P<0.001) which mapped at 19 cM on LG E (Fig. 2-8c). Previously reported markers associated with seed protein concentration on LG E were located at approximately the 27-31 cM region in the 2008 Soybean Genetic Map by linkage analysis (Brummer et al., 1997; Fasoula et al., 2004; Lee et al., 1996; Tajuddin et al., 2003). The SSR marker BARC-Satt384 at 19.6 cM on LG E is reported to be associated with a seed protein QTL (Tajuddin et al., 2003). To define the exact

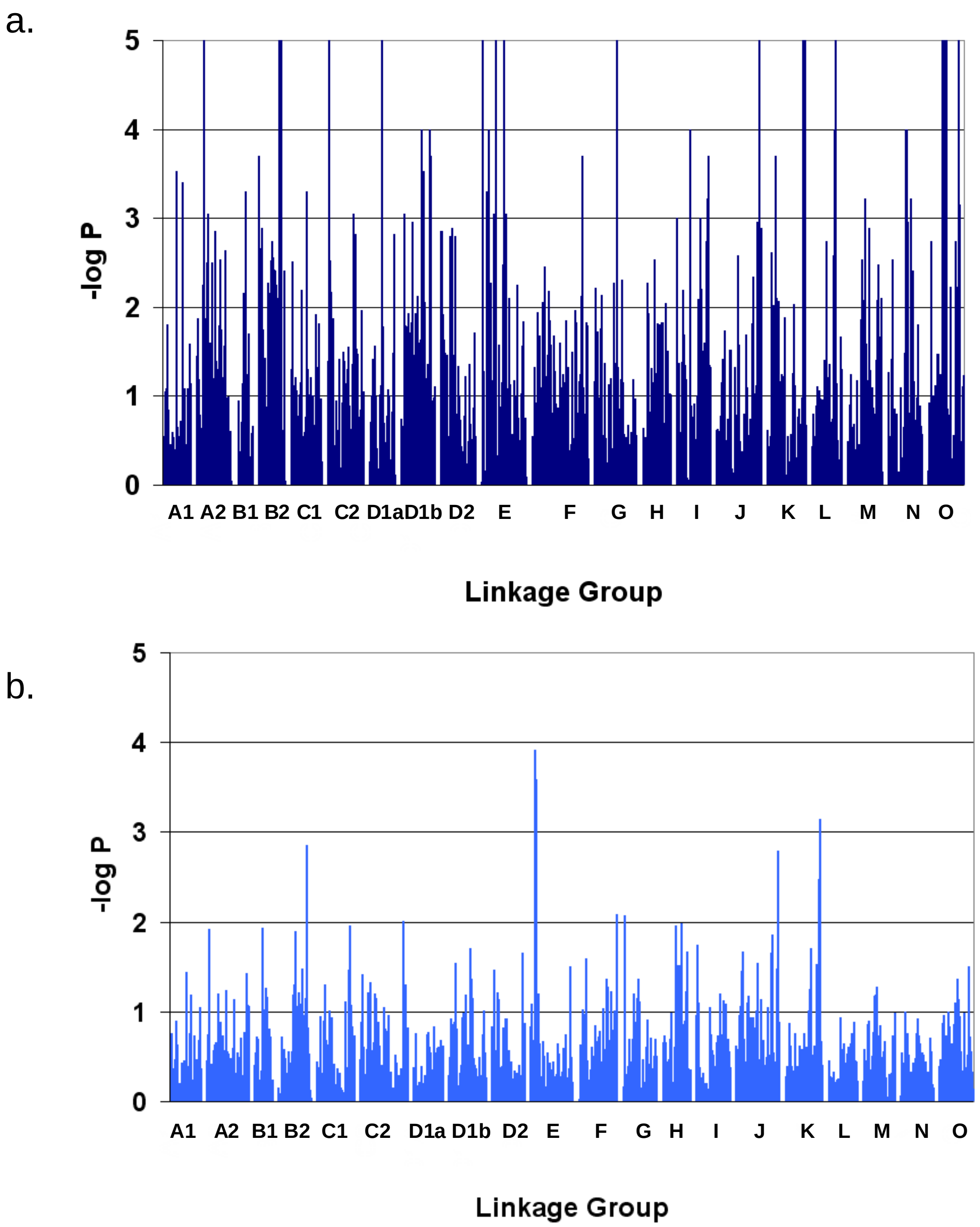


Figure 2-13. Genome-wide association analysis for QTLs controlling seed protein concentration. a is the result of naïve test (no adjustment for population structure) (logistic regression) and b is the result after the structure adjustment using mixed-linear-model. The height of the peak indicates the degree of association. X-axis indicates all the markers on each linkage group and Y-axis indicates the negative log P-values.

location of the seed protein QTL on LG E, the physical distance of markers which showed the most significant associations was determined using the sequence of the DoE, JGI *Glycine max*-scaffold_70. The sequence length of scaffold_70 was more than 4,000 Kbp. The 12 markers located in the scaffold_70 including the two markers most significantly associated with seed protein concentration, BARC-016533-02084 and BARC-042349-08247, were located at 3,900 and 3,959 Kbp, respectively, are shown in Fig. 2-14. Simple sequence repeat marker Satt384, reported by Tajuddin et al. (2003), to define the position of a seed protein QTL was also located in the same scaffold_70 at 3,889 Kbp. A comparison made with the 43 highest seed protein concentration lines (>45%) *vs*. the 43 lowest seed protein concentration lines (<41%) showed that there were two additional markers at 4,223.7 Kbp (BARC-030883-06960) and 4,224.1 Kbp (BARC-030453-06869) in the high seed protein lines which were not shown in Fig. 2-14 because they have a minor allele frequency less than 0.10 in the normal protein concentration accessions. A genome region spanning 323.9 Kbp where the two highly associated markers, BARC-016533-02084 and BARC-042349-08247 as well as BARC-030883-06960 and BARC-030453-06869 were located had a single haplotype in 95% of the normal seed protein lines while there were two haplotypes in the high seed protein lines with frequencies of 53.5 and 46.5%. These four markers, including the two most significant markers (BARC-016533-02084 and BARC-042349-08247) were also in high LD (*D'* and r^2) supporting the suggestion that the candidate gene most likely resides within a 323.9 Kbp region between BARC-16533-02084 at 3,900 Kbp and BARC-030453-06869 at 4,244.1 Kbp on LG E (Fig. 2-14).

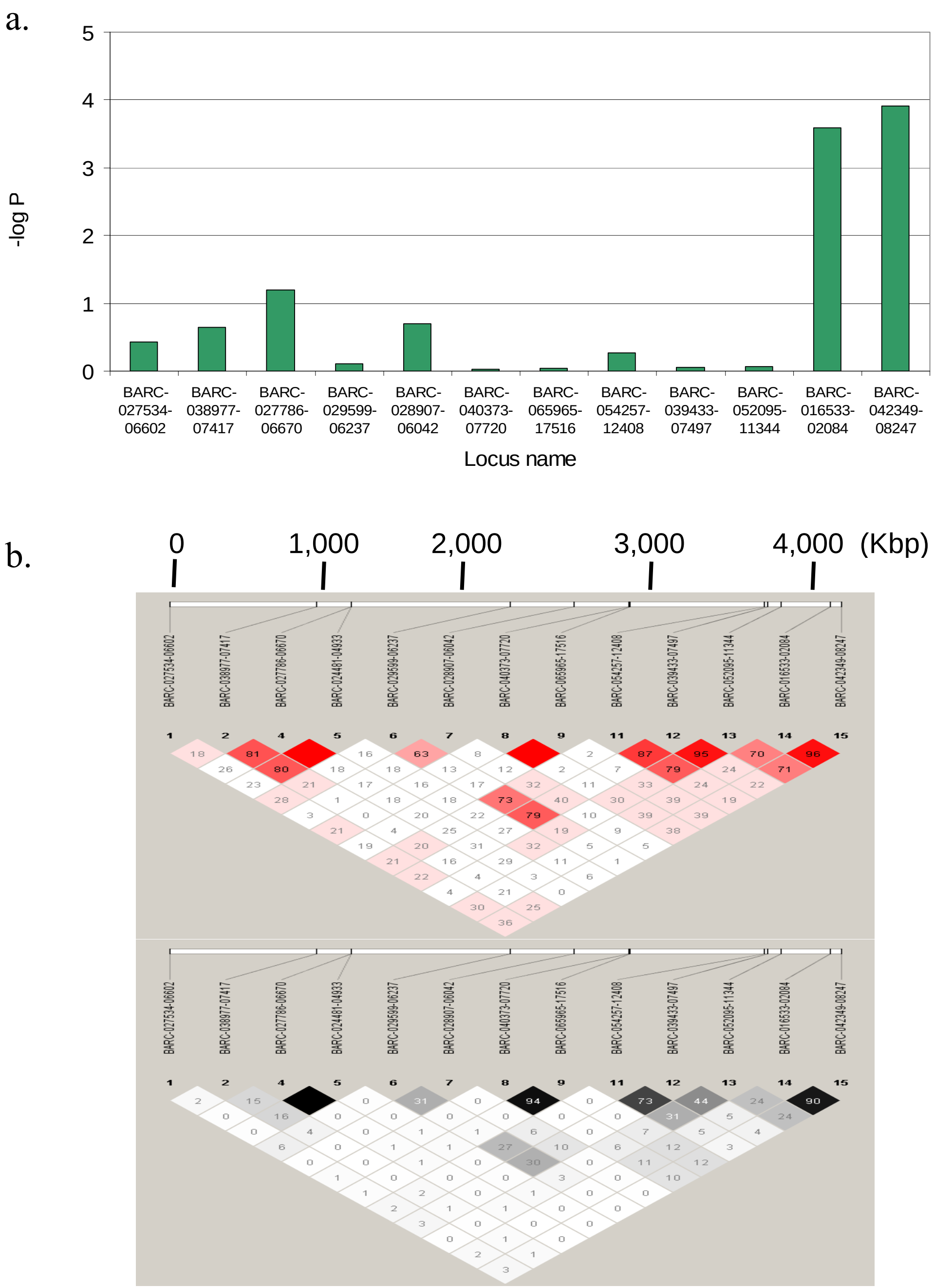


Figure 2-14. The degree of association of the SNP markers developed in close proximity to candidate region of seed protein QTL on LG E. The level of LD among the SNP markers are shown in the red color plot (*D'*) and gray color plot (r^2). Values are expressed as *D'*×100 and r^2×100. The number on the top of the LD block indicates the order of the markers which have a minor allele frequency>0.10.

The other markers showing strong associations with seed protein concentration were BARC-062731-18022 and BARC-042559-08304, were located at 91 cM on LG K. Therefore, using whole-genome association analysis a seed protein QTL on LG E was determined to reside in a 323.9 Kbp region and new seed protein QTLs on LG K was detected.

A Candidate Gene Approach to Detect the Seed Protein QTL on LG I

A major seed protein QTL has been reported in the 27 to 31 cM region on LG I in various studies (Brummer et al., 1997; Chung et al., 2003; Diers et al., 1992; Sebolt et al., 2000) with R^2 values varying from 24 to 44% (Table 2-3). To construct a fine map of this candidate region, a total of 66 SNP markers were developed in the vicinity of this candidate seed protein QTL, all of which mapped to the 29 to 30 cM region between SSR markers BARC-Satt239 and BARC-Satt496. Knowledge of the exact distance between markers would allow an estimate of LD among the markers and would assist in determining the location of a causal mutation controlling seed protein level. The physical order of 63 SNP markers was determined using the whole-genome DoE, JGI assembly_scaffold_1 spanning 18,635 Kbp (http://www.phytozome.net/soybean). There were 29 markers associated with the scaffold_1 with MAF > 0.10. There was a high level of LD among 11 of these 29 markers located in a genome region spanning 2,024 Kbp (Fig. 2-15). Of these 11 markers, six were developed from the subclones of BAC clones associated with the SSR marker, BARC-Satt496, which is reported to be closely linked with the seed protein QTL (Chung et al., 2003). The result of the mixed-linear-model association analysis indicated that none of the markers developed in the candidate region had a

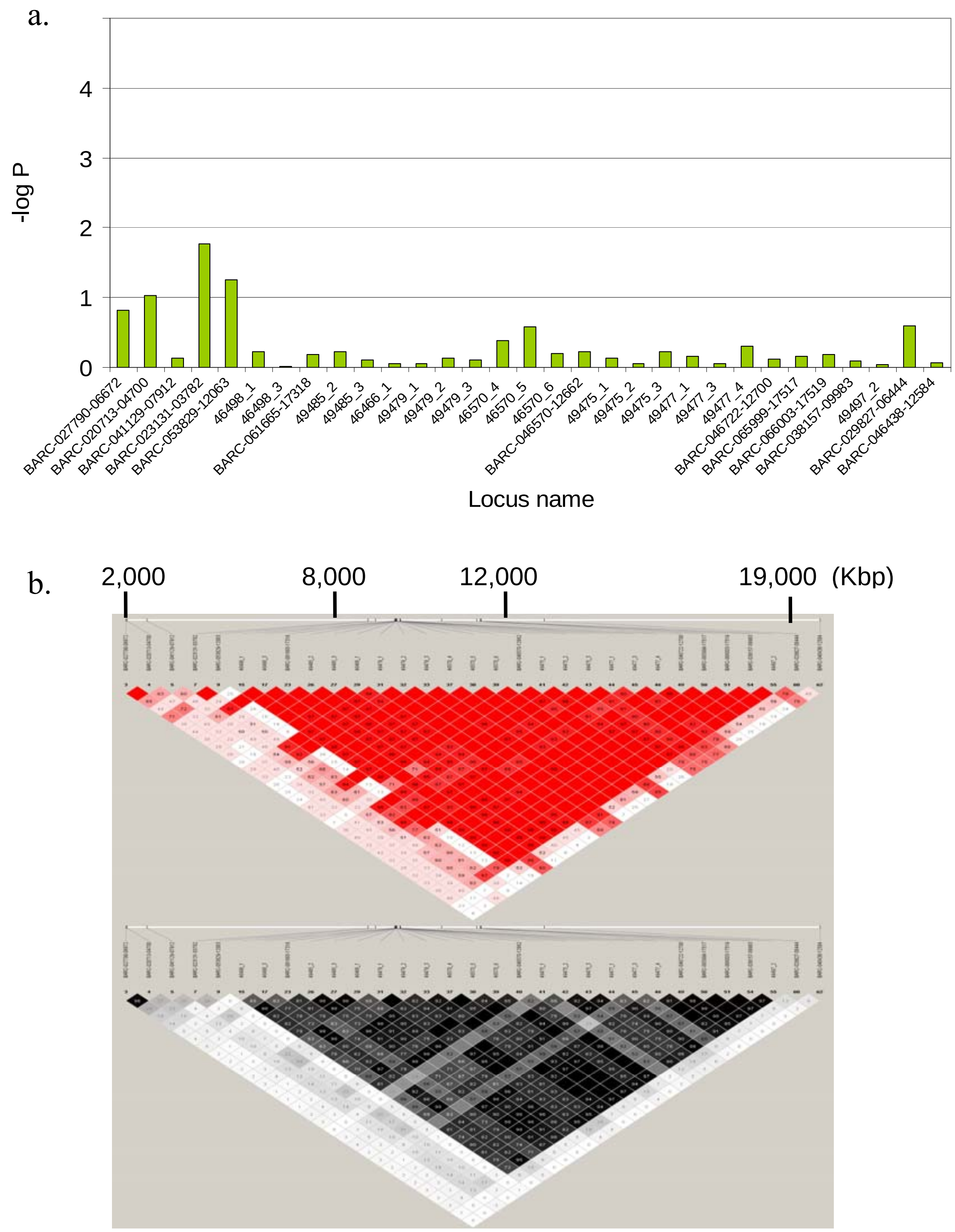


Figure 2-15. The degree of association of the SNP markers developed in close proximity to candidate region of seed protein QTL on LG I. a shows the level of significant of the markers developed from the candidate seed protein QTL region. The level of LD among the SNP markers are shown in the red color plot (*D'*) and gray color plot (r^2). Values are expressed as *D'*×100 and r^2×100. The number on the top of the LD block indicates the order of the markers which have a minor allele frequency>0.10.

significant association with seed protein concentration (Fig. 2-12b). In addition, these markers were in unusually extensive LD which extended 2,024 Kbp from the marker contrast to the average level of LD in the soybean germplasm accessions in this study which was less than 650 Kbp as indicated by $r^2 \geq 0.2$. In addition, there were two same main haplotypes in this high LD block in both high and normal seed protein groups.

Discussion

Soybean germplasm accessions that include mostly landraces represent an excellent reservoir of genetic variation for the application of association analysis. It is assumed that these accessions contain unique alleles for numerous agricultural traits that were absent in the 17 ancestors that form the genetic base of elite U.S. cultivars (Gizlice et al., 1994). A recent study indicated that currently cultivated soybeans in the U.S contain about 72% of the sequence diversity of the Asian landraces from which they derive (Hyten et al., 2006). In the current study, the feasibility of whole-genome association analysis was evaluated in 319 soybean germplasm accessions to detect the location of the genes controlling flower color, pubescence color, and seed protein concentration. The level of LD in the soybean germplasm accessions declined rapidly to less than 0.8 within 50 Kbp as indicated by *D'* and decreased to less than 0.6 at 1,000 Kbp (Fig. 2-2). This was a surprisingly low level of LD given that soybean has selfing rates greater than 99%. The relatively low level of LD in *Glycine max* may be the result of the fact that *Glycine soja*, the weedy wild soybean from which soybean was domesticated, has a higher out-crossing rate than *Glycine max*. The average out-crossing rate of *Glycine soja* is reported to range from 2.3% (Kiang et al., 1992) to 19% (Fujita et al., 1997) which varies depending on the number of pollinators visiting the soybean flower (Fujita et al., 1997). Also, the distance that pollen could be carried by bees is reported to be as much as 60m. In general, the distance of seed dispersal by pod dehiscence is up to 3m (Li et al., 1997). Long distance seed dispersal could also occur by streams or rivers or by animals. Therefore, the low level of LD in *Glycine max* could be the result of recombination that accumulated in the wild soybean *Glycine soja* before domestication occurred. In addition, wild soybean is distributed over a broad area from the central and the northern part of Eastern Asia to eastern Russia. This wide geographical range has produced a highly diverse *Glycine soja* population which is adapted to local conditions and most particularly to photoperiod duration. Thus, it is this diversity from which the early domesticates arose and became the ancestors of cultivated soybean and would contribute to the surprisingly low level of LD in the soybean germplasm accessions analyzed in this study. These highly diverse and unique populations, from which the early domesticates derive, as well as the strong selection of soybean genotypes after domestication would likely result in the presence of the extensive population structure in germplam accessions (Fig. 2-5). Strong selection after domestication can also increase the level of LD surrounding regions of the targeted loci controlling traits under selection.

The level of LD as indicated by r^2 is an important factor related to the likely success of whole-genome association mapping and allows for an estimation of the density of genetic markers needed for association analysis. The extent of LD in these soybean germplasm accessions declined rapidly to 0.2 within 600 Kbp as indicated by r^2 (Fig. 2-2) which was similar to the report by Hyten *et al*. (2007). The extent of LD in maize, an out-crossing species, varied from 0.5 to 600 Kbp (r^2=0.2) depending on the genes being analyzed (Jung et al., 2004; Palaisa et al., 2004; Wilson et al., 2004). In the selfing species, *Arabidopsis thaliana,* LD declined to r^2=0.2 at 10 Kbp. The different levels of LD, as indicated by r^2, between different species and different genes within species, can be explained by a combination of population histories including mating system, mutation rate, founding effects, selection, etc. (Flint-Garcia et al., 2003; Gaut and Long, 2003). In human genetics "tag SNPs" that represent the allelic variation of a haplotype block are used to detect alleles/genes associated with disease susceptibility. Using tag SNPs also reduces the number of SNPs required for whole-genome scans (Johnson et al., 2001). In this study, a total of 1,747 markers were associated with 216 sequence scaffolds in the Williams 82 whole-genome-sequence spanning 609,469 Kbp to estimate the structure of haplotype blocks. The coverage of these markers was approximately 55.4% of the soybean genome which is the largest study performed to date in soybean. Of these genome regions, an average 32% of the haplotype variation could be explained by tag SNPs associated with haplotype blocks (Table 2-3). The magnitude of haplotype block size variation accounted for was dependent upon the method used to analyze haplotype blocks (Table 2-3). The results indicated that the number of markers used was insufficient in approximately two thirds of the genome regions analyzed and more markers are required to capture haplotype block diversity in the germplasm employed in this study. In addition, the low marker density made it difficult to estimate the number of markers needed to perform whole-genome association analysis in this highly diverse set of soybean germplasm accessions.

Despite the relatively low level of LD, high population structure, and low density of markers in the 319 soybean germplasm accessions, the known genes controlling the qualitative traits, flower color and pubescence color, were detected using whole-genome case-control analysis. However, the type I error rate present with each of the different statistical methods used to adjust for population structure indicated that the structure adjustment using a mixed-linear-model was not adequate to adjust for the complex population structure in soybean germplasm accessions collected from across Asia used in this study (Fig. 2-6). This is in contrast to the successful adjustment for population structure reported in maize using a mixed-linear-model (Yu and Buckler, 2006). As shown in the results of association analysis for flower color and pubescence color in Fig. 2-10, the true associations were clearly distinguished from the false associations even with the naïve test. The 65 bp insertion in the *GmF3'5'H* gene that was completely associated with flower color alteration was readily detected. This result corresponded with the report by Zabala and Vodkin (2007). In the case of pubescence color, seven markers developed from the *F3'H* gene (Fig. 2-8b), all of which were in high LD, showed significant associations (Fig. 2-13) with pubescence color. Nonetheless, none of the markers was completely associated with tawny *vs*. gray pubescence color. Although the C deletion (066177_1) at the 3' end of the *F3'H* gene was reported as a determinant of the pubescence color alteration by Toda *et al*. (2002), there were 24 accessions whose gray pubescence could not be explained by these seven markers including the C deletion. This result suggests that the 24 gray pubescence lines without the C deletion carry a different mutation in the *F3'H* gene or in another gene(s) in the pathway controlling pubescence color. The qualitative nature of the genetic effect of the two genes, *GmF3'5'H* and *F3'H,* made it relatively simple to detect the location of these genes using association analysis despite the presence of high population structure and low marker density. These results are similar to those reported in other species using association analysis to detect genes with major genetic effect such as in *Arabidopsis* (Olsen et al., 2004; Stinchcombe et al., 2004; Zhao et al., 2007) and in maize (Palaisa et al., 2003; Thornsberry et al., 2001; Wilson et al., 2004).

The biggest challenge for association analysis is the identification of genes associated with quantitative traits that include many human diseases and agriculturally important traits in plants. Unlike the analysis of the qualitative traits flower and pubescence color, in the analysis of seed protein QTL the results were decidedly less definitive. After the adjustment for population structure, the two most significantly associated genome regions were on LG K (Fig. 2-9b). Two markers showing the most significant associations with seed protein concentration did correspond to the previously reported candidate gene region on LG E (Brummer et al., 1997; Fasoula et al., 2004; Lee et al., 1996; Tajuddin et al., 2003). The analysis localized the causative mutation to a 323.9 Kbp region on LG E. Currently, 72 seed protein QTLs have been reported at numerous positions across 19 of the 20 soybean LGs (SOYBASE, http://soybase.org). However, none of these seed protein QTL has been reported at the 91 cM region on LG K. Assuming these QTLs are not false positives, it is rather surprising that the large number of previous studies using conventional QTL analysis have not discovered QTLs at these three positions.

One virtue attributed to association analysis *versus* linkage analysis is the ability to detect a range of genes controlling the phenotype under study rather than just those segregating in a specific bi-parental cross. Thus, it is surprising that the major seed protein QTL on LG I which has been reported in a number of studies using different populations (Brummer et al., 1997; Chung et al., 2003; Diers et al., 1992; Nichols et al., 2006; Sebolt et al., 2000) was not detected in this study. Based on the previous reports, the major seed protein QTL is located between SSR markers, BARC-Satt496 and BARC-Satt239 approximately at the 26 to 31 cM region in the 2008 Soybean Consensus Map on LG I with reported R^2 values ranging from 24 to 44% (Table 2-1). The fact that none of the 29 markers (MAF > 0.1) targeted to this candidate region showed significant association with seed protein may have a number of explanations. Firstly, 11 of 29 markers that were specifically targeted to this QTL region were in an extensive LD block that extended more than 2,000 Kbp (Fig. 2-15). The extensive LD suggested that this genome region may have undergone selection. Selection may have been for increased seed size which would have been a logical target of selection of ancient soybean farmers. The average seed protein concentration of *Glycine max* germplasm accessions is approximately 42.1% (Wilson, 2004) while *Glycine soja,* the wild progenitor of soybean, has a extremely small seed size and much higher seed protein concentration (> 50%) (Bao, 1989; Xu, 1985). It is reasonable to assume that during soybean domestication, selection would have been imposed for larger seed types which, given the negative correlation between seed protein concentration and seed size, would result in larger seeds with lower protein concentration. If a single founding event was responsible for the lower seed protein concentration, one would anticipate detecting one haplotype which would distinguish the selected lower seed protein lines from their higher seed protein progenitors. However, there were two main haplotypes in this extensive LD block in both high and normal protein groups. This would suggest the possibility that two different founding events, both responsible for larger seed size and decreased protein concentration occurred in different haplotypes of *Glycine soja*. A second explanation for the inability for our association analysis to detect the seed protein QTL on LG I with a large genetic effect may be the lack of markers in sufficiently close proximity to the QTL. As indicated in Fig. 2-15 there were a number of gaps in the genome region thought to contain the LG I seed protein QTL in which no SNP markers were present. Furthermore, using linkage analysis, the location of a causal mutation may only be resolved to a region spanning 10 to 20 cM. In this study, the markers targeted to the protein candidate region were developed within 2 cM of the putative position of the QTL. Thus, it is possible that the major gene controlling seed protein concentration is located outside of the putative candidate region shown in Fig. 2-15. The inability to detect the QTL on LG I, as well as the candidate QTLs on LGs A1 and M, may be due to the inadequacy of the analyses used to adjust for population structure and thus true associations were obscured. Soybean has unusually high population structure resulting in natural sub-populations, mainly dependent on geographical location and selfing habit (as in Arabidopsis) with selection over thousands of years of domestication by Asian farmers which would result in additional population structure. Similarly, another explanation for the inability to detect the previously reported QTLs is the lack of statistical power due to the relatively small sampling of 252 yellow seed coat germplasm accessions combined with relatively small genetic effects that may be associated with the seed protein QTLs on LGs A1 and M. Larger sample size will increase the power to detect alleles in weak LD with nearby SNPs. A final reason for the inability to detect these previously reported QTLs was the relatively low LD that is present in the soybean germplasm accessions. High-density SNP markers evenly distributed across the whole-genome will provide greater power to detect genes with moderate and small genetic effect. However, the relatively low number of markers used in this study made it difficult to estimate the exact number of markers required for whole-genome association analysis.

This was the first report of whole-genome association analysis to detect previously reported qualitative genes and known QTLs in soybean germplasm. Despite the relatively low level of LD, the low density marker coverage, and high population structure, the genome position of the *GmF3'5'H* gene controlling flower color on LG F and the *F3'H* gene controlling pubescence color on LG C2 were readily detected. The detection of four QTLs, previously associated with seed protein concentration, a quantitative trait, was not successful. While one known QTL was successfully mapped to a 323.9 Kbp region on LG E, none of the other three previously reported QTLs was detected in the set of germplasm accessions employed here. In addition, new seed protein QTL on LG K was detected. It is unclear why association analysis was not successful in the detection of the three previously reported seed protein QTLs. However, a number of reasons including incomplete adjustment for population structure, lack of statistical power, an inadequate number of genetic markers in light of the low level of LD, and the relatively modest genetic effects associated with certain of the seed protein QTL under study are suggested as possible reasons.

Bibliography

Aranzana, M.J., S. Kim, K. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. Tang, C. Toomajian, B. Traw, H. Zheng, J. Bergelson, C. Dean, P. Marjoram, and M. Nordborg. 2005. Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet. 1:e60.

Ardlie, K.G., L. Kruglyak, and M. Seielstad. 2002. Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. 3:299-309.

Bao, X.U. 1989. A decade study of Chinese wild soybean (*Glycine soja*). Jilin agriculture sciences 39:5.

Barrett, J.C., B. Fry, J. Maller, and M.J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263-265.

Brummer, E.C., G.L. Graef, J. Orf, J.R. Wilcox, and R.C. Shoemaker. 1997. Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci. 37:370-378.

Burton, P.R., D.G. Clayton, L.R. Cardon, N. Craddock, and P. DelouKas. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678.

Caldwell, K.S., J. Russell, P. Langridge, and W. Powell. 2006. Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. Genetics 172:557-567.

Cardon, L.R., and G.R. Abecasis. 2003. Using haplotype blocks to map human complex trait loci. Trends Genet. 19:135-140.

Chen, Y., and R.L. Nelson. 2004. Identification and characterization of a white-flowered wild soybean plant. Crop Sci. 44:339-342.

Choi, I.Y., D.L. Hyten, L.K. Matukumalli, Q. Song, J.M. Chaky, C.V. Quigley, K. Chase, K.G. Lark, R.S. Reiter, M.S. Yoon, E.Y. Hwang, S.I. Yi, N.D. Young, R.C. Shoemaker, C.P. van Tassell, J.E. Specht, and P.B. Cregan. 2007. A Soybean Transcript Map: Gene Distribution, Haplotype and Single-Nucleotide Polymorphism Analysis. Genetics 176:685-696.

Chung, J., H.L. Babka, G.L. Graef, P.E. Staswick, D.J. Lee, P.B. Cregan, R.C. Shoemaker, and J.E. Specht. 2003. The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci. 43:1053-1067.

Conrad, D.F., T.D. Andrews, N.P. Carter, M.E. Hurles, and J.K. Pritchard. 2006. A high-resolution survey of deletion polymorphism in the human genome. Nat. Genet. 38:75-81.

Csanadi, G., J. Vollmann, G. Stift, and T. Lelley. 2001. Seed quality QTLs identified in a molecular map of early maturing soybean. Theor. Appl. Genet. 103:912-919.

Darvasi, A., and S. Shifman. 2005. The beauty of admixture. Nat. Genet. 37:118-119.

Devlin, B., and K. Roeder. 1999. Genomic control for association studies. Biometrics 55:997-1004.

Diers, B.W., P. Keim, W.R. Fehr, and R.C. Shoemaker. 1992. RFLP analysis of soybean seed protein and oil content. Theor. Appl. Genet. 83:608-612.

Diwan, N., and P.B. Cregan. 1997. Automated sizing of fluorescent-labeled simple sequence repeat (SSR) markers to assay genetic variation in soybean. Theor. Appl. Genet. 95:723-733.

Doebley, J., A. Stec, and L. Hubbard. 1997. The evolution of apical dominance in maize. Nature 386:485-488.

Eeles, R.A., Z. Kote-Jarai, G.G. Giles, A.A. Olama, M. Guy, S.K. Jugurnauth, S. Mulholland, D.A. Leongamornlert, S.M. Edwards, J. Morrison, H.I. Field, M.C. Southey, G. Severi, J.L. Donovan, F.C. Hamdy, D.P. Dearnaley, K.R. Muir, C. Smith, M. Bagnato, A.T. Ardern-Jones, A.L. Hall, L.T. O'Brien, B.N. Gehr-Swain, R.A. Wilkinson, A. Cox, S. Lewis, P.M. Brown, S.G. Jhavar, M. Tymrakiewicz, A. Lophatananon, S.L. Bryant, A. Horwich, R.A. Huddart, V.S. Khoo, C.C. Parker, C.J. Woodhouse, A. Thompson, T. Christmas, C. Ogden, C. Fisher, C. Jamieson, C.S. Cooper, D.R. English, J.L. Hopper, D.E. Neal, and D.F. Easton. 2008. Multiple newly identified loci associated with prostate cancer susceptibility. Nat. Genet. 40:316-321.

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567-1587.

Fan, J.B., K.L. Gunderson, M. Bibikova, J.M. Yeakley, J. Chen, E. Wickham Garcia, L.L. Lebruska, M. Laurent, R. Shen, and D. Barker. 2006. Illumina universal bead arrays. Methods Enzymol. 410:57-73.

Fan, J.B., A. Oliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bentley, J. Haas, P. Rigault,

L. Zhou, J. Stuelpnagel, and M.S. Chee. 2003. Highly parallel SNP genotyping. Cold Spring Harb. Symp. Quant. Biol. 68:69-78.

Fasoula, V.A., D.K. Harris, and H.R. Boerma. 2004. Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. Crop Sci. 44:1218-1225.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S.t. Buckler. 2003. Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54:357-374.

Frary, A., T.C. Nesbitt, S. Grandillo, E. Knaap, B. Cong, J. Liu, J. Meller, R. Elber, K.B. Alpert, and S.D. Tanksley. 2000. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289:85-88.

Frazer, K.A., D.G. Ballinger, D.R. Cox, D.A. Hinds, L.L. Stuve, R.A. Gibbs, J.W. Belmont, A. Boudreau, P. Hardenbol, S.M. Leal, S. Pasternak, D.A. Wheeler, T.D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S.B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R.C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M.M. Waye, S.K. Tsui, H. Xue, J.T. Wong, L.M. Galver, J.B. Fan, K. Gunderson, S.S. Murray, A.R. Oliphant, M.S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.F. Olivier, M.S. Phillips, S. Roumy, C. Sallee, A. Verner, T.J. Hudson, P.Y. Kwok, D. Cai, D.C. Koboldt, R.D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.C. Tsui, W. Mak, Y.Q. Song, P.K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, *et al.* 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-61.

Fujita, R., M. Ohara, K. Okazaki, and Y. Shimamoto. 1997. The extent of natural cross-pollination in wild soybean (*Glycine soja*). J. Hered. 88:124-128.

Gabriel, S.B., S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altshuler. 2002. The structure of haplotype blocks in the human genome. Science 296:2225-2229.

Gaut, B.S., and A.D. Long. 2003. The lowdown on linkage disequilibrium. Plant Cell 15:1502-1506.

Gizlice, Z., T.E. Carter, Jr., and J.W. Burton. 1994. Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci. 34:1143-1151.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8:195-202.

Gudmundsson, J., P. Sulem, A. Manolescu, L.T. Amundadottir, D. Gudbjartsson, A. Helgason, T. Rafnar, J.T. Bergthorsson, B.A. Agnarsson, A. Baker, A. Sigurdsson, K.R. Benediktsdottir, M. Jakobsdottir, J. Xu, T. Blondal, J. Kostic, J. Sun, S. Ghosh, S.N. Stacey, M. Mouy, J. Saemundsdottir, V.M. Backman, K. Kristjansson, A. Tres, A.W. Partin, M.T. Albers-Akkers, J. Godino-Ivan Marcos, P.C. Walsh, D.W. Swinkels, S. Navarrete, S.D. Isaacs, K.K. Aben, T. Graif, J. Cashy, M. Ruiz-Echarri, K.E. Wiley, B.K. Suarez, J.A. Witjes, M. Frigge, C. Ober, E. Jonsson, G.V. Einarsson, J.I. Mayordomo, L.A. Kiemeney, W.B. Isaacs, W.J. Catalona, R.B. Barkardottir, J.R. Gulcher, U. Thorsteinsdottir, A. Kong, and K. Stefansson. 2007. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat. Genet. 39:631-637.

Gupta, P.K., S. Rustgi, and P.L. Kulwal. 2005. Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol. Biol. 57:461-485.

Hardy, O.J., and X. Vekemans. 2002. spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular ecology notes 2:618-620.

Hegstad, J.M., J.A. Tarter, L.O. Vodkin, and C.D. Nickell. 2000. Positioning the *wp* flower color locus on the soybean genome map. Crop Sci. 40:534-537.

Hinds, D.A., L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, K.A. Frazer, and D.R. Cox. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072-1079.

Hurburgh, J., C.R., J.M. Brumm, J.M. Guinn, and R.A. Hartwig. 1990. Protein and oil patterms in U.S. and world soybean markets. J. Am. Oil Chem. Soc. 67:966-973.

Hymowitz, T. 1970. On the domestication of the soybean. Economic botany. 24:408-421.

Hymowitz, T. 2004. Speciation and cytogenetics. p. 97-136. *In* H.R. Boerma and J.E. Specht (ed.) Soybean: Improvement, production, and Uses. 3rd ed.

Hyten, D.L., V.R. Pantalone, C.E. Sams, A.M. Saxton, D. Landau-Ellis, T.R. Stefaniak, and M.E. Schmidt. 2004. Seed quality QTL in a prominent soybean population. Theor. Appl. Genet. 109:552-561.

Hyten, D.L., I.Y. Choi, Q. Song, R.C. Shoemaker, R.L. Nelson, J.M. Costa, J.E. Specht, and P.B. Cregan. 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175:1937-1944.

Hyten, D.L., Q. Song, Y. Zhu, I.Y. Choi, R.L. Nelson, J.M. Costa, J.E. Specht, R.C. Shoemaker, and P.B. Cregan. 2006. Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. U. S. A. 103:16666-16671.

Hyten, D.L., Q. Song, I. Choi, M. Yoon, J.E. Specht, L.K. Matukumalli, R.L. Nelson, R.C. Shoemaker, N.D. Young, and P.B. Cregan. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor. Appl. Genet. 116:945-952.

Iwashina, T., E.R. Benitez, and R. Takahashi. 2006. Analysis of flavonoids in pubescence of soybean near-isogenic lines for pubescence color loci. J Hered. 97:438-43.

Johnson, G.C., L. Esposito, B.J. Barratt, A.N. Smith, J. Heward, G. Di Genova, H. Ueda, H.J. Cordell, I.A. Eaves, F. Dudbridge, R.C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S.C. Gough, D.G. Clayton, and J.A. Todd. 2001. Haplotype tagging for the identification of common disease genes. Nat. Genet. 29:233-237.

Jung, M., A. Ching, D. Bhattramakki, M. Dolan, S. Tingey, M. Morgante, and A. Rafalski. 2004. Linkage disequilibrium and sequence diversity in a 500-kbp region around the adh1 locus in elite maize germplasm. Theor. Appl. Genet. 109:681-689.

Kabelka, E.A., B.W. Diers, W.R. Fehr, A.R. LeRoy, I.C. Baianu, T. You, D.J. Neece, and R.L. Nelson. 2004. Putative alleles for increased yield from soybean plant introductions. Crop Sci. 44:784-791.

Keim, P., T.C. Olson, and R.C. Shoemaker. 1988. A rapid protocol for isolating soybean DNA. Soybean Genet. Newsl. 15:150-152.

Khaja, R., J. Zhang, J.R. MacDonald, Y. He, A.M. Joseph-George, J. Wei, M.A. Rafiq, C. Qian, M. Shago, L. Pantano, H. Aburatani, K. Jones, R. Redon, M. Hurles, L. Armengol, X. Estivill, R.J. Mural, C. Lee, S.W. Scherer, and L. Feuk. 2006. Genome assembly comparison identifies structural variants in the human genome. Nat. Genet. 38:1413-1418.

Kiang, Y.T., Y.C. Chiang, and N. Kaizuma. 1992. Genetic diversity in natural populations of wild soybean in Iwate prefecture, Japan. J. Hered. 83:325-329.

Kim, S., V. Plagnol, T.T. Hu, C. Toomajian, R.M. Clark, S. Ossowski, J.R. Ecker, D. Weigel, and M. Nordborg. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. 39:1151-1155.

Kondrashov, A.S., and J.F. Crow. 1993. A molecular approach to estimating the human deleterious mutation rate. Hum. Mutat. 2:229-234.

Latch, E.K., G. Dharmarajan, J.C. Glaubitz, and O.E. Rhodes Jr. 2006. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conservation Genetics 7:295-302.

Lee, S.H., M.A. Bailey, M.A.R. Mian, T.E. Carter, Jr., E.R. Shipe, D.A. Ashley, W.A. Parrott, R.S. Hussey, and H.R. Boerma. 1996. RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor. Appl. Genet. 93:649-657.

Li, J., S.Z. Zheng, J. Qian, W.W. Ren, and P.H. Ye. 1997. Seed rain of *Glycine soja*. Chinese J. Appl. Ecol. 8:372-376.

Liu, J., J. Van Eck, B. Cong, and S.D. Tanksley. 2002. A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. Proc. Natl. Acad. Sci. U.S.A. 99:13302-13306.

Mansur, L.M., J.H. Orf, K. Chase, T. Jarvik, P.B. Cregan, and K.G. Lark. 1996. Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci. 36:1327-1336.

Marth, G.T., I. Korf, M.D. Yandell, R.T. Yeh, Z. Gu, H. Zakeri, N.O. Stitziel, L. Hillier, P.Y. Kwok, and W.R. Gish. 1999. A general approach to single-nucleotide polymorphism discovery. Nat. Genet. 23:452-456.

Matukumalli, L., J. Grefenstette, D.L. Hyten, I. Choi, P.B. Cregan, and C.P. Van Tassell. 2006. Application of machine learning in SNP discovey. BMC Bioinformatics 7:4.

Mitchell-Olds, T., J.H. Willis, and D.B. Goldstein. 2007. Which evolutionary processes influence natural genetic variation for phenotypic traits? Nat. Rev. Genet. 8:845-856.

National Agricultural Statistics Service. 2007. http://usda.mannlib.cornell.edu/usda/current.CropProd/CropProd-01-11-2008.pdf.

Nichols, D.M., K.D. Glover, S.R. Carlson, J.E. Specht, and B.W. Diers. 2006. Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. Crop Sci. 46:834-839.

Nowotan, A., H. Gambus, P. Liebhard, W. Praznik, R. Ziobro, W. Berski, and J. Krawontka. 2006. Characteristics of carbohydrate fraction of rye varieties. Acta Sci. Pol., Technol. Aliment. 5:87-96.

Olsen, K.M., S.S. Halldorsdottir, J.R. Stinchcombe, C. Weinig, J. Schmitt, and M.D. Purugganan. 2004. Linkage disequilibrium mapping of Arabidopsis CRY2 flowering time alleles. Genetics 167:1361-9.

Orf, J.H., K. Chase, F.R. Adler, L.M. Mansur, and K.G. Lark. 1999a. Genetics of soybean agronomic traits. II. Interactions between yield quantitative trait loci in soybean. Crop Sci. 39:1652-1657.

Orf, J.H., K. Chase, T. Jarvik, L.M. Mansur, P.B. Cregan, F.R. Adler, and K.G. Lark. 1999b. Genetics of soybean agronomic traits. I. Comparison of three related recombinant inbred populations. Crop Sci. 39:1642-1651.

Palaisa, K., M. Morgante, S. Tingey, and A. Rafalski. 2004. Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. Proc. Natl. Acad. Sci. U.S.A. 101:9885-9890.

Palaisa, K.A., M. Morgante, M. Williams, and A. Rafalski. 2003. Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 15:1795-806.

Palmer, R.G., T.W. Pfeiffer, G.R. Buss, and T.C. Kilen. 2004. Qualitative genetics. p. 137-214. In H.R. Boerma and J.E. Specht (ed.) Soybean: Improvement, production, and Uses. 3rd ed.

Panthee, D.R., V.R. Pantalone, D.R. West, A.M. Saxton, and C.E. Sams. 2005. Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. Crop Sci. 45:2015-2022.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000a. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Pritchard, J.K., M. Stephens, N.A. Rosenberg, and P. Donnelly. 2000b. Association mapping in structured populations. Am. J. Hum. Genet. 67:170-181.

Qiu, B.X., P.R. Arelli, and D.A. Sleper. 1999. RFLP markers associated with soybean cyst nematode resistance and seed composition in a 'Peking' x 'Essex' population. Theor. Appl. Genet. 98:356-364.

Redon, R., S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen, E.K. Cho, S. Dallaire, J.L. Freeman, J.R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J.R. MacDonald, C.R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M.J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D.F. Conrad, X. Estivill, C. Tyler-Smith, N.P. Carter, H. Aburatani, C. Lee, K.W. Jones, S.W. Scherer, and M.E. Hurles. 2006. Global variation in copy number in the human genome. Nature 444:444-54.

Reich, D.E., M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, T. Lavery, R. Kouyoumjian, S.F. Farhadian, R. Ward, and E.S. Lander. 2001. Linkage disequilibrium in the human genome. Nature 411:199-204.

Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler, IV. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. U.S.A. 98:11479-11484.

Rioux, J.D., R.J. Xavier, K.D. Taylor, M.S. Silverberg, P. Goyette, A. Huett, T. Green, P. Kuballa, M.M. Barmada, L.W. Datta, Y.Y. Shugart, A.M. Griffiths, S.R. Targan, A.F. Ippoliti, E.J. Bernard, L. Mei, D.L. Nicolae, M. Regueiro, L.P. Schumm, A.H. Steinhart, J.I. Rotter, R.H. Duerr, J.H. Cho, M.J. Daly, and S.R. Brant. 2007. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat. Genet. 39:596-604.

Rostoks, N., L. Ramsay, K. MacKenzie, L. Cardle, P.R. Bhat, M.L. Roose, J.T. Svensson, N. Stein, R.K. Varshney, D.F. Marshall, A. Graner, T.J. Close, and R. Waugh. 2006. Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc. Natl. Acad. Sci. U.S.A. 103:18656-18661.

Sachidanandam, R., D. Weissman, S.C. Schmidt, J.M. Kakol, L.D. Stein, G. Marth, S. Sherry, J.C. Mullikin, B.J. Mortimore, D.L. Willey, S.E. Hunt, C.G. Cole, P.C. Coggill, C.M. Rice, Z. Ning, J. Rogers, D.R. Bentley, P.Y. Kwok, E.R. Mardis, R.T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R.H. Waterston, J.D. McPherson, B. Gilman, S. Schaffner, W.J. Van Etten, D. Reich, J. Higgins, M.J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M.C. Zody, L. Linton, E.S. Lander, and D. Altshuler. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-933.

Schweder, T., and E. Spjotvoll. 1982. Plots of *P-value* to evaluate many tests simulaneously. Biometrika 69:493-502.

Sebolt, A.M., R.C. Shoemaker, and B.W. Diers. 2000. Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci. 40:1438-1444.

Simko, I., K.G. Haynes, and R.W. Jones. 2006. Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 173:2237-2245.

Sladek, R., G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, B. Balkau, B. Heude, G. Charpentier, T.J. Hudson, A. Montpetit, A.V. Pshezhetsky, M. Prentki, B.I. Posner, D.J. Balding, D. Meyre, C. Polychronakos, and P. Froguel. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881-885.

Specht, J.E., K. Chase, M. Macrander, G.L. Graef, J. Chung, J.P. Markwell, M. Germann, J.H. Orf, and K.G. Lark. 2001. Soybean response to water: a QTL analysis of drought tolerance. Crop Sci. 41:493-509.

Stinchcombe, J.R., C. Weinig, M. Ungerer, K.M. Olsen, C. Mays, S.S. Halldorsdottir, M.D. Purugganan, and J. Schmitt. 2004. A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA. Proc. Natl. Acad. Sci. U.S.A. 101:4712-4717.

Tajuddin, T., S. Watanabe, N. Yamanaka, and K. Harada. 2003. Analysis of quantitative trait loci for protein and lipid contents in soybean seeds using recombinant inbred lines. Breeding science 53:133-140.

Takahashi, R., S.M. Githiri, K. Hatayama, E.G. Dubouzet, N. Shimada, T. Aoki, S. Ayabe, T. Iwashina, K. Toda, and H. Matsumura. 2007. A single-base deletion in soybean flavonol synthase gene is associated with magenta flower color. Plant Mol. Biol. 63:125-135.

Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, and B.S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. mays L.). Proc. Natl. Acad. Sci. U.S.A. 98:9161-9166.

Tenaillon, M.I., M.C. Sawkins, L.K. Anderson, S.M. Stack, J. Doebley, and B.S. Gaut. 2002. Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. mays L.). Genetics 162:1401-1413.

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789-796.

Thornsberry, J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S. Buckler, IV. 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. 28:286-289.

Tishkoff, S.A., and B.C. Verrelli. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu. Rev. Genomics Hum. Genet. 4:293-340.

Toda, K., D. Yang, N. Yamanaka, S. Watanabe, K. Harada, and R. Takahashi. 2002. A single-base deletion in soybean flavonoid 3'-hydroxylase gene is associated with gray pubescence color. Plant Mol. Biol. 50:187-196.

Vigouroux, Y., J.S. Jaqueth, Y. Matsuoka, O.S. Smith, W.D. Beavis, J.S. Smith, and J. Doebley. 2002. Rate and pattern of mutation at mocrosatellite loci in maize. Mol. Biol. Evol. 19:1251-1260.

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. 71:1227-1234.

Wilcox, J.R., and Z. Guodong. 1997. Relationships between seed yield and seed protein in determinate and indeterminate soybean populations. Crop Sci. 37:361-364.

Wilson, L.M., S.R. Whitt, A.M. Ibanez, T.R. Rocheford, M.M. Goodman, and E.S. Buckler, IV. 2004. Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16:2719-2733.

Wilson, R.F. 2004. Seed composition. p. 621-677. In H.R. Boerma and J.E. Specht (ed.) Soybean: Improvement, production, and Uses. 3rd ed.

Wolfe, K.H., W.H. Li, and P.M. Sharp. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. U.S.A. 84:9054-9058.

Xu, B. 1985. The protein resources of *Glycine max* in China. Soybean Sci. 3:327-331.

Xu, D.H., J. Abe, M. Sakai, A. Kanazawa, and Y. Shimamoto. 2000. Sequence variation of non-coding regions of chloroplast DNA of soybean and related wild species and its implications for the evolution of different chloroplast haplotypes. Theor. Appl. Genet. 101:724-732.

Xu, M., and R.G. Palmer. 2005. Genetic analysis and molecular mapping of a pale flower allele at the W4 locus in soybean. Genome 48:334-340.

Yano, M., Y. Katayose, M. Ashikari, U. Yamanouchi, L. Monna, T. Fuse, T. Baba, K. Yamamoto, Y. Umehara, Y. Nagamura, and T. Sasaki. 2000. Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. Plant Cell 12:2473-2484.

Yu, J., and E.S. Buckler. 2006. Genetic association mapping and genome organization of maize. Curr. Opin. Biotechnol. 17:155-160.

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203-208.

Zabala, G., and L.O. Vodkin. 2005. The wp mutation of *Glycine max* carries a gene-fragment-rich transposon of the CACTA superfamily. Plant Cell 17:2619-32.

Zabala, G., and L.O. Vodkin. 2007. A rearrangement resulting in small tandom repeats in the F3'5'H gene of white flower genotypes in associated with the soybean W1 locus. The plant genome 47:S113-S124.

Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. 2007. An Arabidopsis Example of Association Mapping in Structured Samples. PLoS Genet. 3:e4.

Zhu, Y.L., Q.J. Song, D.L. Hyten, C.P. Van Tassell, L.K. Matukumalli, D.R. Grimm, S.M. Hyatt, E.W. Fickus, N.D. Young, and P.B. Cregan. 2003. Single-nucleotide polymorphisms in soybean. Genetics 163:1123-1134.