Fall 12-2010

DETECTION OF SOYBEAN SEED PROTEIN QTLS USING SELECTIVE GENOTYPING

Piyaporn Phansak
University of Nebraska-Lincoln

DETECTION OF SOYBEAN SEED PROTEIN QTLS USING SELECTIVE GENOTYPING

By

Piyaporn Phansak

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Agronomy (Plant Breeding and Genetics)

Under the Supervision of Professor James E. Specht

Lincoln, Nebraska

December, 2010

DETECTION OF SOYBEAN SEED PROTEIN QTLS USING SELECTIVE GENOTYPING

Piyaporn Phansak, Ph. D.

University of Nebraska, 2010

Advisor: James E. Specht

A quantitative trait locus (QTL) is a statically defined location of a gene governing that trait. QTL identification is the first step towards using marker-assisted selection (MAS) to introgress desirable QTL alleles into elite high-yield cultivars. Hundreds of high protein plant introductions (PIs) exist in the USDA germplasm collection and are a source of high protein alleles. Although 86 protein QTLs are currently listed in SoyBase, many are likely repeat discoveries of the same QTL(s), given the typical +/- 10 cM confidence intervals associated with QTL positions. Six germplasm accessions of maturity groups (MGs) II to IV that exhibited high seed protein (480 g kg^{-1} or more) were mated to a high-yielding cultivars of the same MG that exhibited normal seed protein (420 g kg^{-1} or less) to generate six F_2 populations. A total of 240 individual F_2 plants in each population produced $F_{2:3}$ seed progenies that were phenotyped for seed protein content. Selective genotyping, or phenotypic tail analysis, was used to genotype only those $F_{2:3}$ progenies occupying the lowest decile and the highest decile. A 1536-SNP

locus assay chip was used for the genotyping. In the six mapping populations, eight protein QTLs with LOD scores greater than 3.0 were detected and mapped on five linkage groups using R/qtl. Significant QTLs on LG-C2 (Chromosome 6), LG-O (10), LG-B2 (14), LG-E (15), and LG-I (20) were detected. A review of the currently listed QTLs in Soybase (2010) indicated that no seed protein QTLs had been previously reported on LG-O (10). The new seed protein QTL discovered in this study in populations 1076, 1121, and 1122 is located on LG-O (10) near the two adjacent markers S19004 and S15265, and has an additive effect of 9.6, 7.9, and 6.5 g kg^{-1} greater seed protein, respectively. For improving the seed protein content in high yielding soybean cultivars, the accessions PI 437112A (1076), PI 398672 (1121), and PI 360843 (1122), which posses the high protein allele at this new LG-O (10) protein QTL, may be useful to soybean breeders.

ACKNOWLEDGEMENTS

It is a pleasure to thank all of the people who made this dissertation possible. I would never have been able to finish this dissertation without the guidance of my advisor and committee members, help from my friends, and support from my family.

I am heartily thankful and wish to express my deepest gratitude to my advisor, Dr. James E. Specht, whose encouragement, supervision, and support throughout the entirety of my dissertation enabled me to develop an understanding of the subject. It has been an honor to work under his supervision. His support has been insurmountable in a number of ways: He has exhibited excellent guidance, care, and patience, and has provided an excellent atmosphere in which to do research. I am grateful for Dr. George L. Graef, who enabled me to experience additional research in the soybean breeding program and provided invaluable plant breeding knowledge and practical issues beyond the textbook. I also appreciate him always being supportive and cheerful. I would also like to express my thanks to Dr. Ismail Dweikat for guiding and advising my research, especially with respect to molecular techniques. He has assisted me in developing my background in molecular genetics and has given me an opportunity to work and learn new methods from his laboratory. I would like to thank Dr. P. Stephen Baenziger and Dr. Gautam Sarath. Each has always been supportive and cheerful and has given very useful knowledge in plant breeding, genetics, and biochemistry.

I am also indebted to many of my colleagues who have supported me. First, Mike Livingston was very helpful both in the laboratory and in the field. Paul Nabity, Aaron Hoagland, and Travis Wegner were always willing to help and give advice with any problem I experienced but could not resolve on my own. Many thanks go out to Dr. Watcharin Soonsuwan, Brendan Borer, Sarah Brownell, Kyla Ronhovde, Trenton Hinze, and the other members and student workers of the Specht team for helping with field work and collecting samples. My dissertation would not have been possible without them.

I owe my deepest gratitude to Dr. Perry B. Cregan and Dr. David L. Hyten for their help with DNA extraction and SNP genotyping at the Soybean Genomics and Improvement Lab, Beltsville Agricultural Research Center-West, USDA. Without their help my research would have not been complete.

I would like to thank Dr. Julian Chaky, who is a good friend and is always willing to help and give his best advice. It would have been a lonely lab without him. I am grateful to have such good friends like Kayse Onweller, Scott Dworak, Jonathan Chaky, and Kittichai Chaiseeda, who have always helped and supported me—always there cheering me up and standing by through the good times and bad. I also would like to show my appreciation to another good friend, Dr. Nicholas Crowley, who always gives me his best wishes, support, and always keeps me updated on science outside my research. I am also so grateful to have been able to work by the sides Kyle Kocak, Joseph Jedlicka, and Yu-Kai Sun. They have been very helpful, supportive, and overall a joy with whom to work. Many thanks go to Dr. Nongluk Teinseree, Dr. Benjawan Siriwetwiwat, Dr. José Aponte, Dr. Lekgari A. Lekgari, Dr. Desalegn D. Serba, Dr. Neway Mengistu, Dr. Anyamanee Auvuchanon, and all my friends who always encouraged and supported me.

I owe my deepest thanks to Dr. Hugo Volkaert, my former boss who always had faith, trusted in me, and pulled me up when I felt weak. He always told me, “Never give up, Bee! You can do it!” He taught me everything he knew and shared lots of good memories. Without his immense support and advice, I would have not been able to start my Ph.D. program here in the United States.

I would also like to thank my parents, younger brother, and sister. They have always given me the moral support that I required and were always there to support and encourage me with their best wishes throughout my Ph.D. study.

Finally, I would like to thank the Royal Thai Government, who provided me with a scholarship to study here. Without the support from my country, I would have not been able to come here.

TABLE OF CONTENTS

PAGE

LIST OF ABBREVIATIONS

a – additive effect

AFLP – amplified fragment length polymorphism

BAC – bacterial artificial chromosome

bp – base pair

BC– backcross

cDNA – complementary deoxyribonucleic acid

Chr–chromosome

C.I. – confidence interval

cM– centriMorgans

d– dominance effect

DH–doubled haploid

DNA– deoxyribonucleic acid

EM–expectation maximization algorithm

EST– expressed sequence tag

GRIN– germplasm resources information network

LOD– logarithm of odds

LG– linkage group

MAS– marker-assisted selection

MR– marker regression

NGRP– national genetic resources program

NIR– near-infrared reflectance

OD–optical density

List of Abbreviations, continued

PCR– polymerase chain reaction

PI– plant introduction

Pop– population

QTL or QTLs– quantitative trait locus or loci

RAPD– random amplified polymorphic DNA

RFLP– restriction fragment length polymorphism

RIL or RILs– recombinant inbred line or lines

SNP– single nucleotide polymorphism

SSR– simple sequence repeat

STS– sequence-tagged site

USLP– universal soy linkage panel

LIST OF TABLES

PAGE

List of Tables, continued

Appendix Tables:

LIST OF FIGURES

PAGE

List of Figures, continued

Appendix Figures:

INTRODUCTION

Soybean [*Glycine max* (L.) Merr.] is one of the world's major crops grown for seed protein and oil content. Soybean seed is usually processed to obtain oil for human cooking oil industry and protein for the animal feed industry. Soybean processors crush the raw soybeans and then extract the oil from the meal. Because soybeans are high in protein, the meal is a key element in most livestock rations (Waldroup, 2009). A small percentage of soy-derived protein is used for soy milk, soy flour, soy protein, tofu, and many other retail human food products (Martin et al., 2006). Many non-food industrial products are also derived from soybean extracts.

Soybean hulls (i.e., seed coats) are removed from the soybean seeds before they are processed to separate oil from the meal. Soybean meal is a relatively high-energy, medium-protein feed. Soybean hulls are sometimes added back to the meal provide fiber in the diets of cows and calves. The oil may be refined for cooking and other edible uses, or sold for biodiesel production or industrial uses (Martin et al., 2006).

Soybeans account for more than half (53%) of the world oilseed production, and 38 percent of 2010 global soybean production was produced in the United States (Soystats, 2010). In USA, soybean is annually grown on more than 31 million hectares and with a most recent yield of 2.91 metric tons per hectare, and the total production was a of 90 million metric tons (Soystats, 2010).

The soybean seed constituents of protein and oil are the most economically important components of this seed crop. On average, seed of current USA soybean cultivars contains approximately 41% protein and 21% oil on a zero per cent dry weight basis (Hartwig and Kilen, 1991). However, the precise composition of soybean seed is

dependent upon cultivar, planting date, soil, and seasonal weather factors, notably temperature. Simpson and Wilcox (1983) noted that the same cultivar grown in different years could vary significantly in seed composition. Helms et al. (1990) noted that the protein concentration of soybean seed increased as planting date was delayed. Aside from genetic factors, the application of N, P, K fertilizers and lime are among the most important soil nutrient factors affecting the seed composition of soybean (Dornbos and Mullen, 1992). Because of the inverse relationship between seed protein and oil content, environmental factors that enhance seed protein content tend to depress seed oil content.

The temperature during seed-fill has also been shown to influence total oil and seed protein content of soybeans (Wolf et al., 1982; Simpson and Wilcox, 1983). Maximum seed oil content occurred in seed that developed at ambient temperatures between 25°C and 28°C, but decreased linearly if temperature exceeded that range (Dornbos and Mullen, 1992; Piper and Boote, 1999). The seed protein content was constant or slightly decreased when temperature increased from 15°C to 25 to 28°C, but at temperatures greater than 28°C, the protein content increased linearly with increases in temperature (Wolf et al., 1982; Dornbos and Mullen, 1992; Pipolo et al., 2004). Environmental stress during soybean seed-fill can alter the chemical composition of the seed. The impacts of drought and high air temperature on soybean seed protein and oil have been reported. Drought stress and high air temperatures during late seed-fill are known to raise seed oil content (Howell and Cartter, 1958; Specht et al., 2001).

Considering the nutritional and economic importance of soybean protein and oil content, attempts have been made to increase the concentration of both constituents in soybean seed (Panthee et al., 2005). However, the high negative correlation between

protein and oil content has not permitted a simultaneous increase in both constituents in a given seed, simply because the increase in the content of one almost invariably comes at the expense of the other (Chung et al., 2003). High yielding, high-protein soybean cultivars are also difficult to develop because of the inverse relationship between seed yield and seed protein content (Burton, 1987; Wilcox and Cavins, 1995; Cober and Voldeng, 2000; Chung et al., 2003)

The sources of most high protein genes used in breeding programs are typically unadapted plant introductions that have a very high seed protein content (Cober and Voldeng, 2000). Soybean seed protein is known to be quantitatively inherited, and is highly heritable (Burton, 1987; Chung et al., 2003). There is some evidence that the inheritance is oligenic rather than polygenic (Wehrmann et al., 1987). The final soybean seed protein content depends on the joint action of various genetic loci and the interactions of these loci with the environment, which can make genetic analysis of this trait complex and difficult to interpret. Over the last several years, many studies have identified genomic segments known as quantitative trait loci (QTLs) that govern seed protein content in populations derived from the matings of high-protein soybean plant introductions (PIs) with high-yield cultivars of ordinary or conventional protein content. A significant amount of genetic variation for seed protein content is often observed in the resultant populations. The protein-controlling QTLs in several such populations have been identified and their approximate genomic map positions can be found in Soybase QTL lists (Soybase, 2010).

Relative to previous studies, research conducted by (Ritchie, 2003) led to the identification of *five* high protein germplasm accessions that, when mated to high-yield

elite cultivars, produced populations that did *not* segregate for the well known protein QTLs in soybean linkage groups (LG) I, E, H-top, and H-bottom (now respectively known as chromosome 20, 15, and 12), yet still segregated for a wide range in $F_{2:3}$ progeny protein content. She identified one additional population that segregated for a protein QTL, but *only* in LG-E. Because these six high protein accessions likely possessed high protein-causing alleles at possibly unknown QTLs that may map to heretofore unknown map positions in the 20 soybean linkage groups, the identification and location of these QTLs would be of a great interest to soybean breeders. Therefore research objective pursued in this dissertation was to discover and map the QTLs that accounted for the high protein-causing alleles present in these six high protein PIs. This research would thus either confirm or refute the Ritchie (2003) conclusion that five of these six PIs do not have a high protein allele at LG-I, -E, or -H. The experimental approach used was a selective genotyping method, also known as phenotypic tail analysis, in which only the extreme decile fractions of an $F_{2:3}$ seed protein distribution would be genotyped with 1536 SNP markers having map positions distributed over the soybean genome. The ultimate goal was, of course to identify and map the QTLs influencing seed protein content in these six PIs.

LITERATURE REVIEW

Breeding for Protein Improvement in Soybean

Among all vegetable sources, soybean protein is known to provide the most complete amino acid balance for human food and animal feed. Still, additional improvement in the total protein content of soybean seed would not only benefit the soybean food industry, but also the feed industry (Wilson, 2004; Cianzio, 2007). There exists, however, a negative correlation between seed protein and oil content (Brim and Burton, 1979), and also a negative correlation between seed protein and seed yield (Wehrmann et al., 1987). As a result, it has been difficult to implement improvement in soybean seed protein content without a significant depression in seed oil and more importantly, without a significant reduction in seed yield.

The heritabilities of soybean seed protein and oil are high, especially in populations with parents that differ substantively in seed protein (or oil) content (Brim, 1973; Burton, 1987; Chung et al., 2003). The high heritability observed in many such soybean populations indicated that simple selection would be a reasonably effective method for achieving genetic gain. The seed protein means of a population derived from parentally homozygous strains of high protein randomly mated with those of low protein equaled the midparent values, suggesting that seed protein is determined mainly by additive gene action with little or no dominance effect (Thorne and Fehr, 1970). Panthee et al. (2005) found that the heritability in a high x low protein soybean F_6-derived RIL population was 0.96 for protein and 0.95 for oil. Chung et al. (2003) reported that the heritability for seed protein and oil was 0.89 and 0.84, respectively, in F_5-derived RIL derived from a high x low protein mating. The range of heritability for seed protein and

oil in eight F_2 soybean populations was 0.56 to 0.92 and 0.70 to 0.81, respectively (Brummer et al., 1997).

Extensive research on (a) the degree of genetic variability in the progeny of mated low and high seed protein lines, (b) the expression of seed protein content in temperate and tropical locations, and (c) the transfer of high-protein content to high-yielding lines has been reported in numerous publications. Wehrmann et al. (1987) reported moderate to strong inverse relationships between seed oil and seed protein and also seed yield and seed protein. Breeding programs for increased protein content in soybean seeds have been tested and the results described in many reports (Hartwig and Hinson, 1972; Shannon et al., 1972; Brim and Burton, 1979; Sebern and Lambert, 1984; Wehrmann et al., 1987; Hartwig and Kilen, 1991; Wilcox and Guodong, 1997; Helms and Orf, 1998; Cober and Voldeng, 2000). Hartwig (1973) concluded that it was not possible to retain high seed oil along with seed protein content because of the high negative correlation between the two traits, but also noted that selection for high protein itself (i.e., disregarding seed oil content) appeared to be a realizable objective in a breeding program. Openshaw and Hadley (1984) concluded that breeding methods designed to increase *both* seed protein and oil were not likely to succeed.

In the past, the pedigree and backcrossing methods have been used with limited success to select soybean lines with high seed protein, but with no or little reduction in seed oil. For example, Shannon et al. (1972) evaluated seed protein and oil of F_2-derived lines in the F_4 generation of six populations derived from all possible intercrosses among four homozygous lines, two with high seed protein (C1430 and C1460) and two with high seed yield (Calland and C1461). These authors reported that the association between

seed protein and oil was negative and significant in all six populations. Thus, increasing both seed protein and oil in any of these crosses would be extremely difficult. A more effective procedure to increase seed protein content with a lesser impact on seed oil content might be to set a minimum level for seed oil in the choice of parents so that when selecting for high seed protein in the progeny, acceptable oil levels would be maintained.

The correlation between seed protein and yield has been negative and statistically significant in most reports (Caldwell et al., 1966; Hartwig and Hinson, 1972; Simpson and Wilcox, 1983; Pantalone et al., 1996). Despite the negative relationship between seed protein and yield, many breeders have considered the negative correlation to not be so strong as to restrict progress in the selection of high-yielding, moderately high-protein strains. Wehrmann et al. (1987) evaluated 95 BC_2 progenies in each of three populations, where the recurrent parents were high yielding lines of ordinary seed protein content and the donor parent was Pando, which averaged 480 g kg^{-1} seed protein. In each of these populations, no backcross-derived lines were recovered that combined exceptionally high seed protein with a yield equivalent to that of the recurrent parent. However, in each of two populations, high protein lines were derived that had above-average seed protein contents of 422 and 433 g kg^{-1} and these lines did not differ significantly in yield or seed oil from the recurrent parent. In the third population, a line with substantively high 462 g kg^{-1} seed protein content was obtained, but it was significantly lower in both yield and seed oil concentration than the recurrent parent. The results indicated that when a low yielding, high protein donor parent was utilized as a source of alleles for high seed protein, selection for high protein during two backcross generations would effectively increase seed protein in the backcross progeny to some degree greater than that of high

yielding recurrent parent, and still result in progeny having a seed yield not significantly different from the high-yield parent.

Molecular Markers in Plant Breeding

The genetic analysis of desirable traits in breeding programs has been made substantially easier with the advent of molecular markers and marker-based linkage maps (Paterson et al., 1991). Most of molecular marker systems now utilized routinely by plant breeders involve the use of the polymerase chain reaction (PCR). In soybean, restriction fragment length polymorphisms (RFLPs) were initially used to construct detailed genetic maps, and to preliminarily map a few genes controlling complex traits (Keim et al., 1990; Diers et al., 1992). However, PCR-based markers were rapidly adopted by breeders because of the ease of use, minimal lab work, speed, and higher polymorphism frequency. A wide array of PCR-based markers were developed quickly for use by geneticist and breeders. These included random amplified polymorphic DNAs (RAPDs), sequence-characterized amplified regions (SCARS), arbitrarily primed PCR (AP-PCR) markers. Amplified fragment length polymorphisms (AFLPs) are almost invariably dominant markers, but have the advantage of generating an extremely large number of polymorphic AFLP markers in a single PCR reaction that can be scored on a single gel (Vos et al., 1995). However, the application of AFLP technology requires much skill and expertise. Microsatellite markers, also known as simple sequence repeats (SSRs), represent one of the most breeder-useful marker systems. Because of the potential for multi-allelism at each given locus, SSR markers possess exceptional information capacity. In soybean, SSR markers have proven to be quite informative in many

populations. Single nucleotide polymorphisms (SNPs) recently have become available in some crop species, including soybean (Zhu et al., 2003; Choi et al., 2007). The informativeness of the foregoing molecular marker types is dependent on how single-locus polymorphisms are visualized. Markers such as SSRs, RFLPs, and SNPs can distinguish between homozygous and heterozygous state (i.e., the visualized "alleles" are co-dominant), whereas RAPDs, and AFLPs predominantly generate only dominant-recessive "alleles".

Simple Sequence Repeats (SSRs)

Microsatellite markers or simple sequence repeats (SSRs) are, at the present time, the most convenient, effective, and popular markers used by soybean researchers. SSRs consist of tandemly repeated short sequence motifs of one to five base-pair repeats, and are ubiquitous in eukaryotic genomes (Akkaya et al., 1992; Ellegren, 2004). Thousands of forward and reverse primer pairs for SSR loci have been developed to prepare genetic linkage maps in eukaryotes. SSRs have been used to greatly improve the primary linkage maps in soybean that were initially constructed using RFLP markers. SSR markers, because of the sequence specificity of the forward and reverse primers, are effectively equivalent to 'sequence tagged sites' (STS), in that each SSR is almost invariably monogenic. However, SSRs are also highly polymorphic (multi-allelic), so virtually any segregating population can be used as a reference population for a linkage study (Morgante and Olivieri, 1993).

SSR polymorphisms in soybean were first reported by Akkaya et al. (1992). Currently, more than 1000 SSR loci have now been mapped (Cregan et al., 1999; Song et

al., 2004). Until now, the abundance of SSRs in the soybean genome was dependent on the development and testing to create a database of locus-specific SSR markers with a high likelihood of polymorphism. Now, however, a soybean SSR database (BARCSOYSSR_1.0) with the genome positions and primer sequences for 33,065 "potential" SSRs can be found in Soybase (Song et al., 2010). To characterize each SSR and create a linkage map, each SSR is amplified by PCR with specific forward and reverse primers of 20-30 bases. The amplified fragments will be polymorphic if the parental lines differ in the number of repeats in the primer-flanked amplicon. Thus, every SSR constitutes a genetic locus that maybe highly variable, depending upon its multi-allelicity amongst the parents that might be chosen to create a segregating population. For a given SSR, every amplified segment of different repeat length represents a different allele of same locus (Narvel et al., 2000). SSRs have been extensively used to identify the locations of genomic segments containing genes governing the inheritance of soybean traits like disease resistance, and recently, seed protein content, oil, and yield.

Single Nucleotide Polymorphisms (SNPs)

The most recent advance in molecular marker technology is the single nucleotide polymorphism (SNP). SNPs represent single DNA base differences between homologous DNA helices, though small insertions and deletions (known as indels) of base sequence are also treated as if they too were SNPs. SNPs have been shown to be the most abundant source of DNA polymorphisms in humans. Developments in SNP genotyping technologies and methodologies recently reported in human genomics offer a vision of future possibilities for molecular plant breeding. In contrast to humans, less progress has

been made in the discovery of sequence diversity in plants (Xu and Crouch, 2008). In the initial stages of soybean SNP discovery, analysis of sequence variation was limited to specific genes or DNA fragments (Zhu et al., 1995). However, Coryell et al. (1999), Zhu et al. (2003), Van et al. (2004) and Hyten et al. (2006) subsequently reported that SNPs were at least modestly abundant in soybean genome. Zhu et al. (2003) reported a total of 280 SNPs detected among 25 diverse soybean genotypes in more than 76 kb of amplified PCR product from primers designed from GenBank genes, cDNAs, BAC subclones, and SSR-flanking regions.

As genome-wide SNP markers becomes more readily available in many plant species, it will be possible to construct SNP-based molecular marker linkage maps with a marker-to-marker density of less than 1 cM. Hyten et al. (2008) used the Illumina GoldenGate assay to demonstrate the multiplexing of as few as 96 to as many as 1,536 soybean SNPs in a single reaction over a 3-day period using genotypic DNA samples from three soybean RIL mapping populations. The high multiplex capacity of the Illumina GoldenGate assay allows the analysis of sufficient loci (e.g., 1536) to provide the density needed to be successful in one-step (i.e., 3-day) QTL discovery strategies, once the mapping population has been created and phenotyped. Most recently, Hyten et al. (2010) used the GoldenGate assay to map an additional 2,500 SNPs in the soybean genome. The authors then identified 1,536 SNPs that were distributed approximately uniformly over the 20 chromosomes in the genome, and also had an optimal minor allele frequency in both exotic and adapted soybean germplasm. New technologies for assaying genotypes for SNP allele type are expected to make SNP markers the replacement marker system for the currently used SSR marker systems, relative to future soybean genetics

and breeding studies (Hyten et al., 2008). A particularly important advantage of the Illumina-based SNP allele detection over the SSR marker allele detection is the elimination of the tedious gel-based marker allele visualization required for the latter.

Development of Soybean Genetic Linkage Maps

The use of highly reproducible and abundant genetic markers greatly facilitates the development of a genetic map. Several soybean genetic linkage maps based on RFLP markers were constructed in the early 1990s (Keim et al., 1990; Shoemaker et al., 1992; Keim et al., 1997). The first genetic linkage map for the soybean was reported by Keim et al. (1990). This map consisted of 26 genetic linkage groups containing a total of 150 RFLP markers. This map was based on an F_2 population derived from a *G. max* x *G. soja* mating. Then, in 1993, Shoemaker and Olson (1993) used an F_2 population derived from the same mating to construct a revised molecular genetic linkage map of soybean that consisted of 25 linkage groups with 365 RFLPs, 11 RAPDs, three classical markers, and four isozyme loci. Subsequently, in 1995, Shoemaker and Specht (1995) succeeded in integrating some of the various classical genetic markers into that RFLP linkage map.

Cregan et al. (1999) developed and mapped a set of 606 SSR markers together with 689 RFLPs, 79 RAPDs, 11 AFLPs, 10 isozyme, and 26 classical loci in one or more of three different populations, and aligned the linkage groups (LGs) derived from each of the three populations into a consensus map of 20 LGs that corresponded with the 20 homologous pairs of soybean chromosomes. The three populations included the USDA/Iowa state *G. max* x *G.soja* F_2 population, the University of Utah ‘Minsoy’ x ‘Noir 1’ recombinant inbred population, and the University of Nebraska ‘Clark’ x

'Harosoy' F_2 population. Song et al. (2004) subsequently reported a new integrated genetic linkage map of soybean using five widely used soybean mapping populations including the *G. max* x *G. soja* population USDA/Iowa State University 'A81-356-022' x PI 468916, plus the *G. max* x *G. max* populations of the University of Nebraska 'Clark' x 'Harosoy', and the three University of Utah populations of 'Minsoy' x 'Noir 1', 'Minsoy' x 'Archer', and 'Archer' x 'Noir 1'. Song et al. (2004) reported that a total of 420 new SSR loci had been developed to add to the 606 SSR loci that had been reported by previously Cregan et al. (1999). They also added a total of 66 new RFLPs into this linkage map. This integrated soybean genetic map with more precisely positioned markers served has been treated by the soybean research community as the main reference genetic map for soybean since 2004.

Most of the soybean genetic linkage maps to date were constructed with RFLPs, AFLPs, RAPDs, and SSRs. All of these markers are the actual or amplified fragments of genomic DNA. Recently, SNP markers have been incorporated into linkage maps. Choi et al. (2007) developed the first soybean transcript map by mapping 1141 SNP markers (derived from 1141 expressed gene sequences) onto the previous version of the soybean genetic map (Song et al., 2004), which included 1015 PCR-based markers (SSRs). On the basis of gene-based SNPs mapped, SNP markers were positioned in many of the 5 and 10 cM gaps that existed in the previous map. This map will be very useful for the case study of the diversity of gene function associated with these transcripts, as it will offer researchers an opportunity to identify potential candidate genes for >1,150 QTLs that have been reported to date.

Hyten et al. (2010) recently reported on the latest version of soybean integrated genetic linkage map (Consensus Map 4.0) by adding 2,651 new SNP markers into the previous genetic map developed by Choi et al. (2007). There are a total of 5,500 genetic markers in this new genetic linkage map. Hyten et al. (2010) selected a set of 1,536 SNPs to create a universal soybean linkage panel (USLP) for use in high-throughput soybean QTL mapping, using three mapping populations, including the University of Utah 'Minsoy' x 'Noir 1', 'Minsoy' x 'Archer', plus the University of Minnesota 'Evans' x 'Peking'. In the earlier soybean map, Choi et al. (2007) had reported that there were 40 gaps of 5 to 10 cM, and seven gaps of >10 cM. However, the consensus 4.0 map of Hyten et al. (2010) now has only one gap >10 cM and just 18 gaps of 5 to 10 cM in total length. Version 4.0 of soybean linkage map has become the consensus standard for future QTL mapping applications.

Quantitative Trait Loci (QTLs) Analysis

Many agriculturally important traits such as yield, seed quality, and some forms of disease resistance are controlled by several or many genes and display a quantitative form of inheritance. The genes governing a quantitative trait are known as quantitative trait loci (QTLs). Identification of a single QTL or multiple QTLs based solely on conventional phenotypic evaluation of just the parents and progeny lines is almost always impossible. A major breakthrough in the genetic characterization of quantitative traits at the QTL level was the development of DNA (molecular) markers in the 1980s (Collard et al., 2005). One of the main uses of molecular markers in agricultural research has been in

the construction of molecular linkage maps and the use of markers and maps for QTL discovery in diverse crop species such as soybean.

Many QTLs for soybean seed protein content have been identified to date. A ranking of the 86 seed protein QTL-marker associations listed in SOYBASE (2010) based on the magnitude of the QTL additive effect, indicated that the "strongest" protein QTLs were (in descending rank order) located on LGs I (which corresponds to Chromosome 20), E (Chr 15), H (Chr 12), M (Chr 7), A1 (Chr 5), C1 (Chr 4), F (Chr 13), and G (Chr 18), respectively (Diers et al., 1992; Lee et al., 1996; Orf et al., 1999; Sebolt et al., 2000; Specht et al., 2001). The protein QTL with the strongest allelic effect discovered to date is the one on LG-I (Chr 20), and it was first reported by Diers et al. (1992), who used genetic markers to study the genetic control of seed protein and oil concentration. Actually, these authors detected two major QTLs on both LG-I (Chr 20) and LG-E (Chr 15) controlling protein concentration in a population developed from a cross between PI 468916, an unadapted *G. soja* accession with high protein concentration, and A81-356022, an adapted maturity group (MG) III *G. max* breeding line. Lines homozygous for the *G. soja* allele for the most significant molecular marker linked to each QTL were associated with an increase in seed protein of 24 g kg^{-1} for the LG-I (Chr 20) QTL and 17 g kg^{-1} for the LG-E (Chr 15) QTL, when compared with lines homozygous the *G. max* allele. Diers et al. (1992) noted that introduction of these high protein QTL alleles into current cultivars could have a substantial effect of Northern USA soybean producers experiencing a lower than average seed protein content.

Mansur et al. (1993a) reported that an unlinked RFLP locus, L048, now known to map to LG-I (Chr 20), was associated with seed protein in an $F_{2:5}$ soybean population

from ‘Minsoy’ x ‘Noir1’. The study was later continued using 284 F_7 -derived RILs developed by single seed descent from ‘Minsoy’ x ‘Nior 1’ (Mansur et al., 1996). They mapped QTL for seed protein in this population, but could not confirm the original $F_{2:5}$ population detected association of L048 with protein in the F_7-RIL population. Instead, they reported three QTLs for seed protein on LG-U7 and U14, which correspond to LGs A1 (Chr 5) and L (Chr 19), respectively, of the linkage group alphanumeric naming system of Cregan et al. (1999).

Lee et al. (1996) mapped seed protein QTLs in the 120 F_4-derived lines population from a cross between ‘Young’ x ‘PI 416937’ and 111 F_2-derived lines population from a cross between ‘PI 97100’ x ‘Coker 237’, using RFLP markers. They reported seven seed protein QTLs on LGs E (Chr 15), C1 (Chr 4), J (Chr 16), N (Chr 3), P (LG-B2, Chr 14), and UNK1 in ‘Young’ x ‘PI 416937’ population. They also reported six seed protein QTLs on LG-E (Chr 15), H (Chr 12), K (Chr 9), and UNK2 in ‘PI 97100’ x ‘Coker 237’ population. Only the QTL on LG-E (Chr 15) was detected in both populations. The other seed protein QTLs were population-specific. Moreover, at each of the RFLP loci associated with QTLs, the allele associated with increased seed protein was associated with decreased seed oil, indicating that the negative correlation of protein and oil found at the genotypic level was also detectable at the QTL level.

Brummer et al. (1997) examined eight different populations of F_2-derived lines, and identified RFLP markers associated with what the authors called “environmentally stable” QTLs for soybean protein and/or oil content in nine linkage groups, LGs A2 (Chr 8), B2 (Chr 14), C1 (Chr 4), D1 (Chr 1), E (Chr 15), F (Chr 13), G (Chr 18), H (Chr 12), and I (Chr 20). In this study, the authors classified QTLs that were detected in at least

two of the three years and also in the 3-yr average value as “stable”. The authors also set the criterion for an environmentally stable QTL as a detection probability of $P \leq 0.05$ in two or more years and $P \leq 0.05$ in the average over the three years. All populations had at least one stable QTL for seed protein. The authors also noted that one population, derived from the mating of a breeding line (M82-806) with a high protein line (HHP; 25% *G. soja* by pedigree), possessed a very strong QTL for protein on LG-I (Chr 20) detected with RFLP markers A407-1 and A144. This QTL was likely the same QTL detected by Diers et al. (1992) based on a population constructed from a *G. max* x *G. soja* cross. Therefore, the QTL identified on LG-I (Chr 20) may specific to *G. soja* in the sense that *G. soja* accessions usually have high seed protein and may likely possess a high protein allele at this QTL that *G. max* cultivars do not. Additionally, Diers at al. (1992) identified a QTL for seed protein linked with RFLP marker A023; an environmentally sensitive QTL linked to this same marker reported by Brummer et al. (1997) is now known to be located on LG-L (Chr 19).

Qiu et al. (1999) identified two RFLP markers, B072 on LG-H (Chr 12), and B148 on LG-F (Chr 13), associated with seed protein QTL in a $F_{2:3}$ population derived from ‘Peking’ x ‘Essex’. Since the total phenotypic variation explained by the two QTLs was 30%, they assumed that there should be additional QTLs controlling this trait, but suggested that the QTLs could not detect these because of the background genetic effect of the population. The RFLP marker B072 on LG-H (Chr 12) is also associated with seed oil content. The authors reported that the B072 allele was allegedly associated with both increased seed protein and oil content. However, the protein-oil correlation in the population was quite negative. This is the only seed protein QTL detected so far that

increases seed protein and oil. If proven true, it may be useful to resolve the problem of negative correlation between protein and oil. However, this allele is suspect, and more research is needed to confirm its QTL effect.

Sebolt et al. (2000) continued the research of Diers et al. (1992) by showing that seed protein was increased by 20 g kg^{-1} when the high protein allele of the LG-I (Chr 20) QTL was backcross-introgressed into a normal protein cultivar to create a near-isogenic line (NIL) homozygous for this allele, and which exhibited high protein. Furthermore, the authors reported that NILs homozygous for the *G. soja* marker allele linked to the high protein allele had significantly greater plant height, and earlier maturity, but exhibited reduced yield, seed oil, and seed weight compared to the NILs homozygous for *G. max* alleles, suggesting that these latter effects maybe either pleiotropic effects of the protein QTL allele, or the effects of other alleles at QTLs that were phase-linked to the alleles at protein QTL.

Csanádi et al. (2001) mapped QTLs for seed protein content in an early maturing soybean population developed from the cross of cultivars ‘Maple Belle’ x ‘Proto’. They used 113 SSR, six RAPD, and one RFLP markers segregating in 82 individuals of an F_2 population. They found four QTLs for protein in LGs C2 (Chr 6), M (Chr 7), K (Chr 9), and D1a (Chr 1). They reported a close linkage of seed protein QTL and seed oil QTL on LG-C2 (Chr 6) with a negative correlation (i.e., increased seed protein but decreased seed oil content). The negative correlation between protein and oil content reported in this study at the QTL level was not; i.e., QTL alleles associated with high protein content inversely with low oil content and *vice versa* had been noted by others (Lark et al., 1994; Lee et al., 1996; Sebolt et al., 2000).

Chapman et al. (2003) reported soybean QTLs for agronomic and quality traits in an F_2 population, and in a population of 177 $F_{4:6}$ lines derived from the individual F_2 plants, for a cross of 'Essex' x 'Williams'. They identified two QTLs for seed protein concentration, one linked to Satt373 on LG- L (Chr 19) and one linked to Satt251 on LG-B1 (Chr 11). The protein QTL near Satt373 on LG-L (Chr 19) was not detected in the original F_2 population, and therefore the authors concluded that Satt373 may not be useful for early generation marker-assisted selection for seed protein based on F_2 DNA. The protein QTL near Satt251 on LG-B1 (Chr 11) was detected in the F_2 population, and may be the same protein QTL identified by Brummer et al. (1997) near RFLP marker A109_1 (approximately 10 cM upstream of Satt251).

Chung et al. (2003) documented the phenotypic effects of a high protein allele derived from PI 437088A, a *G. max* accession with a high protein concentration. The QTL mapped to the same region on LG-I (Chr 20) as the QTL reported by Sebolt et al. (2000). Chung et al. (2003) observed an 18 g kg^{-1} increase in protein concentration among lines homozygous for the allele from PI 437088A compared with lines homozygous for the allele from the elite parent. Similar to the QTL described by Sebolt et al. (2000), the high protein QTL detected by Chung et al. (2003) was associated with lower oil concentration, reduced yield, and earlier maturity.

Tajuddin et al. (2003) analyzed the soybean seed protein content trait in the RILs derived from a cross between *G. max* variety 'Misuzudaizu' and variety 'Moshidou Gong 503'. They reported ten QTLs for the seed protein content that were located on LGs I (Chr 20), E (Chr 15), D2 (Chr 17), A2 (Chr 8), C2 (Chr 6), K (Chr 9), L (Chr 19), N (Chr 3), and G (Chr 18), which are rank-listed here from the highest to the lowest additive

effect. The comparison of the mapping results for the seed protein in their study with the results reported previously by Dier et al. (1992), Brummer et al. (1997), and Sebolt et al. (2000) showed some agreement. Tajuddin et al. (2003) reported that an oil QTL was located in the same region as that of the protein QTL on LG-I (20), and like Chung et al. (2003) suggested that the inverse protein-oil phenotypic variation was controlled by the same QTL, or by two different but tightly linked QTLs -one for protein and one for oil.

Fasoula et al. (2004) followed up a prior report (Lee et al., 1996) on QTLs controlling seed protein content in two soybean populations to determine if those QTLs could be confirmed. The same two populations described by Lee et al. (1996), i.e., 'Young' x 'PI416937' and 'PI97100' x 'Coker237' were created anew by making new matings. They reported that only two protein QTLs on LGs E (Chr 15) and UNK2, of the four protein QTLs previously reported by Lee et al. (1996), were detectable in the new 176 $F_{2:4}$ population of 'PI97100' x 'Coker 237'. In their new 'Young' x 'PI416937' population, they could not confirm any of the previously reported seven protein QTLs, and noted that the unconfirmed QTLs were likely false positives (Type I errors) in the original population. Alternatively, they noted that the inability to confirm the QTLs could be caused by QTL environment sensitivity given that the latter phenomenon was observed in previous studies (Lee et al., 1996; Brummer et al., 1997; Xu, 2003). This phenomenon, known as selective bias or more commonly as the "Beavis effect", results in QTLs with true large effects to be more routinely detected than QTLs with true smaller effects, however, the additive effects are almost invariably over-estimated, particularly so for the small effect QTLs. The estimation of the total numbers of QTLs thus depends on the distribution of the magnitude of the true QTL effects (Xu, 2003). Fasoula et al. (2004)

concluded that soybean geneticists should continue efforts to detect, validate, and confirm QTLs that could be more successfully used in marker facilitated selection schemes by applied breeders.

Hyten et al. (2004) studied the seed protein content of a RIL population consisting of 131 F_6 -derived lines created from 'Essex' x 'Williams' using six different testing environments, and using 100 SSR markers spaced throughout the soybean genome. They reported only four protein QTLs in this population and all were previously reported in Soybase (1995) on LGs M (Chr 7), F (Chr 13), C2 (Chr 6), and K (Chr 9).

Kabelka et al. (2004) studied the putative alleles for increased seed protein content and increased seed yield in a population of 167 F_5-derived lines developed from a 'BSR 101' x 'LG82-8379' mating. They found 11 seed protein QTLs on LGs A2 (Chr 8), B2 (Chr 14), C1 (Chr 4), C2 (Chr 6), D1b (Chr 2), F (Chr 13), H (Chr 12), K (Chr 9), M (Chr 7), N (Chr 3), and O (Chr 10). Three of the 11 protein QTLs identified had alleles with positive pleiotropic effects on protein and yield. The LG82-8379 alleles of the QTLs on LGs B2 (Chr 14), M (Chr 7), and O (Chr 10), which increased protein concentration by 2 to 3 g kg^{-1}, also increased yield by 24 to 74 kg ha^{-1}. One LG82-8379 allele of particular interest segregated at the protein QTL linked to Satt358 on LG-O (Chr 10). It increased seed protein concentration 3 g kg^{-1} and also enhanced yield 47 kg ha^{-1}.

Zhang et al. (2004) reported on a soybean genetic linkage map they constructed using 184 RILs derived from the mating 'Kefeng No. 1' x 'Nannong 1138-2'. In this study, the markers included 189 RFLPs, 219 SSRs, 40 ESTs, three *R* gene loci, and one classical phenotype marker. The 452 markers were mapped into 21 linkage groups covering 3,595.9 cM of the soybean genome. All 20 linkage groups corresponded with

the 20 described by Cregan et al. (1999) except for linkage F (Chr 13) which they sub-divided into F1 and F2 due to a large interval between on the F1 and F2 clusters of markers. In their study, the seed protein difference between the parents was not significant. Still, they reported one protein QTL on LG-B2 (Chr 14).

Panthee et al. (2005) evaluated the seed protein QTLs in a population of 101 F_6-derived recombinant inbred lines (RILs) derived from 'N87-984-16' x 'TN93-99'. They genotyped the RILs with 94 SSR markers located on 19 molecular linkage groups. The marker coverage of the genome was only 1057.5 cM, with an average distance of 11.3 cM between markers. The order of most of the markers was in agreement with the public soybean molecular linkage map (Cregan et al., 1999). They found a novel major QTL located near Satt570 on LG-G (Chr 18) that was stable over environments for seed protein content, and another seed protein QTL located near marker Satt274 on LG-D1b (Chr 12) was environmentally sensitive. The QTLs on LGs D1b (Chr 12) and G (Chr 18) controlling seed protein concentration had map positions similar to the QTLs reported earlier by Brummer et al. (1997) and Hyten et al. (2004).

Nichols et al. (2006) fine mapped the seed protein QTL on LG-I (Chr 20). They used two sets of backcross populations developed from introgression of a high protein allele from *G. soja* into the genetic background of breeding line 'A81-356022'. The first set was comprised of three populations of BC_4 lines. The second set consisted of four populations of BC_5 lines. They reported that the LG-I (Chr 20) seed protein QTL was located in a 3-cM interval between SSR marker Satt239 and AFLP marker ACG9b.

Recently, Soares et al. (2008) reported using composite interval mapping to detect seed protein QTLs in a RIL population derived from mating BARC-8 (high protein) with

Garimpo (low protein). The authors identified six flanking pairs of markers, Satt422-Satt282, Satt384-Sat_112, OPAN09-OPAC02, Satt199-Satt594, OPAS07-OPP09, and Satt549-Satt084 on LGs-C2 (Chr 2), E (Chr 15), F (Chr 13), G (Chr 18), L (Chr 19), and N (Chr 3), respectively. Those six flanking markers were associated with the seed protein QTL, with the additive effect of the BARC-8 parental alleles of 11, 12, -8, -10, 9, and 8 g kg^{-1}, respectively.

Jun et al. (2008) detected seed protein QTL in an association mapping analysis of 48 high and 48 low protein germpalsm accessions. The authors reported that the markers Satt 431 on LG-J (Chr 16) and Satt551 on LG-M (Chr 7) were associated with seed protein QTLs.

The results from these foregoing studies show that many protein QTLs have been detected to date, though their high protein allele additive effects vary. Seed protein content improvement was achieved in some studies by marker-assisted backcross introgression of the high protein allele at either the LG-I (Chr 20) locus or the LG-E (Chr 15) locus into high yielding cultivars.

Selective Genotyping Strategy and Mapping Population

A population derived from two marker-polymorphic parental lines that have a quantitative trait contrast is useful because a linkage map can be created with molecular marker analysis, and then after phenotyping, a QTL analysis can be used to detect positions and genetic distances of markers among chromosomes that are associated with the quantitative trait. This association implies that a QTL has a map position near (i.e., linked to) the markers. Linkage map construction requires parental marker

polymorphism, so that marker linkage analysis to construct a marker linkage map (Podlich et al., 2004). Mapping populations in self-pollinating plant species can be created using backcross lines, F_2 progeny, recombinant inbred (RI) lines, or doubled haploid (DH) lines.

F_2 populations arise from an initial cross between two diverse parents that produces an F_1 hybrid. Allowing the F_1 hybrid to self-pollinate produce an F_2 mapping population. One advantage of the F_2 progeny for linkage mapping is the shorter population development time compared to the development of RILs lines; however, an F_2 population segregates in each subsequent generations and, unlike RI or immortal DH lines, F_2 plants cannot be replicated *per se* to conduct multi-year or multi-location field experiments (Collard et al., 2005).

Once a mapping population is finally obtained, DNA from each progeny or line is isolated and evaluated (i.e., genotyped) for the DNA marker polymorphisms that distinguish the parents. The genotyping process may pose a heavy burden of time and cost to the breeder, especially when dealing with thousands of individual plants. Therefore, alternative genotyping methods that could save time and money would be extremely useful, particularly if resources are limited. A short-cut method aimed for identifying markers linked to QTLs was the trait-based approach first described by Stuber et al. (1980; 1982). It was later described in more formal statistical terms by Lebowitz et al. (1987). This trait-based approach was also subsequently termed ‘selective genotyping’ by Lander and Botstein (1989) and is based on genotyping only those individuals or lines in any given population that exhibit the extreme phenotypes for the target trait. The association of markers with a given trait is inferred when one detects marker allele

frequency differences between extreme individuals in the population that have been grouped on the basis of contrasting phenotypes (Lebowitz et al., 1987; Lander and Botstein, 1989; Darvasi and Soller, 1992; Darvasi and Soller, 1994). In any QTL analysis, the most informative progeny are those residing at the extremes of phenotypic distribution (lower and upper tails) of the mapping population (Lander and Botstein, 1989; Bernardo, 2002; Sen et al., 2009). This is why phenotypic tail analysis is an alternative term sometimes used to describe the selective genotyping approach. Lander and Botstein (1989) showed that the best 17% and worst 17% of individuals of a given population (i.e., those progeny with phenotypes exceeding a plus or minus one standard deviation from the population mean) would account for 81% of the total linkage information. Additionally, these authors showed that for a given trait, the number of progeny needed to be genotyped decreases as the phenotypic difference between the two parents increases. Furthermore, in terms of maximal selective genotyping efficiency, it may not be useful to genotype more than the upper 25% and lower 25% of the population for a single trait studies (Darvasi and Soller, 1992; Darvasi, 1997). However, a selective genotyping strategy may allow sampling of only 5 or 10% of the individuals at each phenotypic extreme, as discussed by Ayoub and Mather (2002). In fact, the authors reported that selective genotyping of 5 or 10% of each tail of the phenotypic distribution (i.e., a respective 10 or 20% of the entire population) would have been satisfactory to detect all of the QTL regions that had been detected by interval mapping with the complete data set of the entire RIL population.

The selective genotyping approach is most appropriate for the cases where only one trait is being analyzed at a time (Darvasi, 1997). If two or more traits are of interest,

but are not highly or perfectly correlated, the number of progeny to be genotyped to detect QTLs in each trait greatly increases. This is due to the fact that the set of individuals selected for extreme phenotypic values of one trait will usually not be the same set of individuals selected for extreme phenotypic values for other traits (Tanksley, 1993; Ayoub and Mather, 2002). However, as mentioned above, this trait-based approach for QTL analysis has one advantage over a marker-based QTL analysis, which is that fewer progeny than those constituting the entire population still need to be genotyped, so long as all progeny are phenotyped.

Lebowitz et al. (1987) proposed some theoretical considerations for selective genotyping approach; however, there were some typographical and other errors were present in their original published paper, which made it difficult to interpret without some assistance (Rocha, 2003, personal communication). Hypothetically, the predicted difference marker allele frequencies between the lowest and the highest tails of an F_2 population (assuming decile tail fractions) can be calculated in the following manner:

$$\delta_M = \frac{(i_p)(2a)(m1)(m2)}{\sigma_p}$$

where,

i_p = 1.755, the standardized selection differential for <u>decile</u> selection in an F_2 population,

a = the additive effect of the parental alleles segregating at the QTL,

m1 = m2 = 0.5, the population frequencies of the two parental alleles at a given locus for the F_2 population case,

σ_p = the population's phenotypic standard deviation.

The above equation is an approximation that is only acceptable for primarily small QTL effects, but can be technically improved following the formulas in Falconer (1981) (Rocha, 2003, personal communication) by dividing the result of the above equation by the result of the following equation:

$$1 - [\frac{(i_p)(a)}{\sigma_p}]$$

The standard error ($SE_{\delta M}$) associated with an observed change in marker allele frequency between the two tails is:

$$SE_{\delta M} = \sqrt{\left\{\frac{[(2)(m1)(m2)]}{[(2)(n)]}\right\}}$$

where,

n = number of tail marker alleles (i.e., 2x the number individuals in each tail for a diploid organism),

m1 = m2 = 0.5, the F_2 population frequencies of the two parental alleles at a given locus.

In accordance with the above equations, statistical power of selective genotyping can be estimated assuming that one will genotype only the lowest and highest decile fractions of an F_2 population of a known phenotypic standard deviation, and that there is, a priori, a reasonably reliable estimate of the additive effect of the QTL to be detected by selectively genotyping that population. When using a selective genotyping approach to detect QTL in multiple segregating populations for a trait (i.e., seed protein), one should have available a suitable number of F_2 progeny to phenotype, and then a suitable fraction of the extreme progeny to genotype, in order to be able to detect, at a desired power, those QTLs of some specified additive effect. With regard to the latter criterion, the

known soybean seed protein QTLs of interest in the present case would be those on LGs I (Chr 20), E (Chr 15), and H (Chr 12). Of course, the detection of heretofore undiscovered QTLs with a similar-sized effect on protein are of even more interest.

In dissertation experiment, of interest was knowing what level of power would be available (given a chosen decile sample size of 22 progenies) to detect QTLs whose additive seed protein effect could be as large as, or larger than the 1.2 percentage points (i.e., 12 g protein/kg seed), which is additive effect that has been repeatedly reported for the large effect LG-I (Chr 20) protein QTL (Chung et al., 2003). To compute the power inherent in selective genotyping, consider an F_2 population of approximately 220 phenotyped individuals in which genotyping was limited to the contrasting decile tail samples (each tail consisting of 22 individuals). One can first compute the value of *beta* that associated with each 2a (2*additive effect) value given that power = 1-*beta*. The formula for the power calculation is:

$$Z_{beta} = [(\delta_{A(F2)})/ (SE\delta_{A(F2)})] - Z_{alpha}$$

where,

Z_{alpha} = the ordinate of a normal curve corresponding to the likelihood of *alpha* (Type I) error.

Z_{beta} = the ordinate of a normal curve corresponding to the likelihood of *beta* (Type II) error.

Even though there are no formal standards for choosing a power value, most researchers would plan experiments based on an imputed power of 0.80 which they would consider to be an adequate level of power. Accordingly, the goal in this study was to have a statistical power of 0.80 (i.e., the probability of <u>not</u> committing a type II error)

when searching for QTLs in any of our F_2 populations. To calculate the statistical power of selective genotyping, one must specify a phenotypic standard deviation and an additive effect of a QTL desired to be detected in the F_2 population. The parameter values selected for use in the above formula were some values derived in prior research conducted in the Dr. Specht lab. The standard deviation (σ_P) was set to 30 g protein per kg of seed (i.e., 3.0 seed protein percentage points). This estimate was obtained from a population of 557 $F_{2:3}$ seed progenies that Dr. Specht NIR-phenotyped and genotyped for SSR marker Satt496. The i_P parameter was set to 1.755 (for two-tail extreme decile genotyping). From this calculation, decile-based selective genotyping experiment (i.e., 22 high and 22 low protein extremes in a population of 220 F_2 individuals) would have a statistical power of 0.80 for detecting QTLs whose additive effect was 7.0 g protein per kg of seed (i.e., 0.7 seed protein percentage points or more). In fact, the power would be 100% for QTLs whose additive effect is 0.85 seed protein percentage points or more. This analysis suggested that there was sufficient power for detecting any QTL (new or known) with an additive effect similar to that of the LG-I (Chr 20) QTL.

QTL Detection Based on Selective Genotyping

In this study, the interval mapping method of Lander and Botstein (1989) and single marker regression method of Kearsey and Hyne (1994) were used to detect QTL in six $F_{2:3}$ high seed protein x low seed protein content populations. The interval mapping method employs pairs of neighboring markers to obtain maximum linkage information relative to the possible presence of a QTL within the enclosed segment of the chromosome (Ngwako, 2008; Broman and Sen, 2009), whereas the marker regression

approach investigates individual markers independently, without reference to their position or order. The regression method fits a QTL model to the marker allele means on a given chromosome simultaneously and obtains significance tests by simulation (Doerge, 2002; Ngwako, 2008).

Single marker analysis

Single marker analysis, the simplest method of associating markers with quantitative trait variation, tests for trait value differences between marker genotypes (i.e., AA, AB, BB) one marker at a time. Although simple, this analysis captures the basic simplicity in the idea of QTL mapping; however, there are several weaknesses associated with this simple method of QTL detection (Lander and Botstein, 1989). First, this method cannot tell whether markers are associated with one or more QTLs. Second, it does not estimate the likely positions of the QTLs. Third, the effects of QTLs are likely to be underestimated because they are confounded with the recombination frequencies. Finally, because of confounding effects, this method is not very powerful, and many individuals are required for the test to acquire sufficient power (Kearsey and Farquhar, 1998; Broman and Speed, 2002; Doerge, 2002).

Interval mapping analysis

The use of flanking marker interval mapping methods has proven to be a powerful tool for detecting QTL in segregating populations. Methods to analyze these data, based on maximum-likelihood, have been developed and provide good estimates of QTL effects in some situations (Muranty and Goffinet, 1997). Maximum-likelihood methods are,

however, relatively complex and can be computationally slow. The interval mapping method uses an estimated genetic map as a framework for the location of a QTL. Intervals defined by ordered pairs of markers are searched, and statistical methods are used to test whether a QTL is more likely than not to be present within the interval (Kearsey, 1998; Doerge, 2002; Ngwako, 2008). The approach of interval mapping considers one QTL at a time, and this can bias identification and estimation of QTL when multiple QTL are located in the same chromosome (Zeng, 1994). In this study, the simple or standard maximum likelihood interval mapping approach [via the Expectation-Maximization (EM) algorithm] was considered appropriate for selective genotyping because it uses a maximum-likelihood estimation method. By the property of the maximum-likelihood analysis, the estimate of locations and effects of QTL are asymptotically unbiased if the implicit assumption that there is at most one QTL per chromosome is true (Muranty and Goffinet, 1997; Broman and Speed, 2002).

.

RESEARCH OBJECTIVES

Regarding the QTLs previously studied by Ritchie (2003), the rationale of this research dissertation was to discover and map segregating protein QTL(s) in six segregating F_2 populations for which there was substantive variation in seed protein content, but for which Ritchie (2003) found no evidence of QTLs on LGs I (20), E (15), or H (12). In this dissertation research project, the six high-protein germplasm accessions, representing maturity groups II, III, and IV, were mated to a high-yielding public cultivar of an equivalent maturity group to create six F_2 populations segregating for seed protein content.

The research objectives were:

1. To locate and map QTL(s) governing the high seed protein in these six high protein germplasm accessions, of which five were hypothesized by Ritchie not to possess the high protein allele at LG-I (Chr 20). –E (Chr 15), or –H (Chr 12).
2. To demonstrate that selective genotyping via whole genome 1536 SNP marker analysis and a larger F_2 population than Ritchie is a convenient means for detecting large-effect seed protein QTLs.

MATERIALS AND METHODS

Parental Germplasm

Plant materials were selected based on previous studies conducted by (Ritchie, 2003), who identified five of 41 high protein germplasm accessions [PI 437112A (UNL parent number 1076) , PI 398672 (1121), PI 360843 (1122), PI 407788A (1139), and PI 407823 (1146)] that, when mated to high-yield elite cultivars, did not segregate in F_2 generation for high protein QTLs known to be located in specific regions of the linkage groups (LGs) I (Chr 20), E (Chr 15), H (Chr 12)-top, and H (Chr 12)-bottom, though the F_2 populations still segregated for a wide phenotypic range in $F_{2:3}$ progeny seed protein content. Ritchie (2003) identified one additional population [PI 398704 (1143)] that, while not segregating of the LG-I (Chr 20) QTL, did segregate for a protein QTL in LG-E (Chr 15). It would have been of interest to screen these six populations with markers linked to known QTLs on LGs other than LG-I (Chr 20), -E (Chr 15), or –H (Chr 12), but Ritchie (2003), because of time constraints, concentrated her marker genotyping to just a few SSRs known to map in the QTL regions of those LGs.

The foregoing six high protein plant introductions (PIs) were mated as females to one of three high-yielding recently released public elite cultivars of a similar maturity group. The six PIs and three cultivars are listed in Table 1 with the male parents listed just below the respective female parents. The parent code is a Nebraska nursery number. All other information in Table 1 was excerpted from the Germplasm Resources Information Network (GRIN) website. Note that the six female parents have a GRIN-based seed protein concentration ranging from 482 to 507 g kg^{-1}, whereas the three cultivars have a seed protein ranging from 382 to 424 g kg^{-1} (both on the zero moisture

dry weight basis). The latter values are typical of most of the cultivars used for commercial production in the North Central United States. To minimize the segregation of major genes for flowering and maturity in the F_2 generation, parental matings were restricted to the same maturity group (MG). Because the six female parents ranged from MG II, III, and IV, three different male parents of those MGs had to be used.

Population Development

In October 2006, 40 seeds of each high- and normal-protein parent were planted into 28-cm diameter by 28-cm deep pots filled with a 1:1 mixture of steam-sterilized soil and Metro-Mix 360 soilless media in a mating greenhouse bay located on the East Campus of the University of Nebraska-Lincoln (UNL). As each female parent flowered (in maturity group order from early to late), it was mated to a male parent of the same maturity group (MG) that synchronously flowered. Multiple crosses (10-20 pollinations) were attempted for each of these six matings: 1076 x 1106M, 1121 x 1137M, 1122 x 1137M, 1139 x 1181M, 1143 x 1181M, and 1146 x 1181M (Table 1). Greenhouse-based pollinations for two of the six matings were successful in generating putative F_1 seeds. Four of the six mating failed to pollinate due to synchronization problems with the two early flowering varieties. This failure necessitated a repeat of the mating generation in the subsequent summer field season.

On 7 May 2007, 40 seeds of each high- and normal-protein parent were planted into a 2.5-m row of an on-campus crossing block nursery, located on UNL's East Campus. The parent rows were arranged numerically by parent number from lowest to highest (1076, 1106M, 1121, 1122, 1139, 1143, 1146, and 1181M), with 60 cm of

spacing between rows to allow working space to perform pollinations. The PI accession x agronomic cultivar crosses were made during late June to late July 2007. Approximately 10 to 20 pollinations were attempted for each mating. Pollinations for all of these six matings were successful in generating putative F_1 seeds. The putative F_1 seeds of each mating were individually hand-harvested and packaged by pod at the end of the 2007 summer season.

During the winter of 2007-2008, an on-campus winter greenhouse was used for the F_1 to F_2 generation advance. The putative F_1 seeds from each mating and the parental seeds of the six matings were planted in December 2007 into 28-cm diameter by 28-cm deep pots filled with a 1:1 mixture of steam-sterilized soil and Metro-Mix 360 soilless media. For each mating, a maximum of 20 F_1 seeds were planted into “five-pointed star” F_1 seed positions available in each of four pots. In each pot, five seeds were placed on top of the pre-watered soil mixture, each positioned approximately 5 cm from the pot’s rim. The F_1 seed from each mating was then inoculated with rhizobia and covered with 5 cm of fine, dry soil. Five seeds of each parent were also placed into a pot. Water was not applied again until needed to moisten the soil surface to aid in the emergence of the F_1 seedlings. Subsequently, an automated drip-irrigation system was used to supply 5-minute irrigation to each pot every other morning. To enhance seed set, fertilizer was applied immediately after flowering by adding 100 mL of a standard plant nutrient solution to each pot.

Parental contrasts for one or more of the classical genetic markers controlling flower, pubescence, pod, and hilum color were evident for some matings. For these matings, F_1 plant hybridity was confirmed for those matings in which the female parent

marker was recessive, and the male parent marker was dominant, for one (or more) of these four pigmentation traits. For instance, at flowering stage, a flower color marker was used to distinguish an authentic F_1 hybrid (purple flowers) from an inadvertent recessive white-flowered female parent self that was supposedly mated to a dominant purple-flowered male parent. A small number of putative F_1 plants that resulted from unintentional female self-pollinations were discarded. For those matings in which the male parent was recessive for all four of these loci, but in which the female parent contained the dominant allele for at least one locus, confirmation of the F_1 authentic hybrid would not be possible until the F_2 generation, when marker segregation would become evident. To ensure an earlier indication of the authenticity of putative F_1 plants, a parentally polymorphic molecular marker (i.e., SSR) was used to genotype the F_1 and thus either confirm or refute the F_1 hybridity in these (and actually all six) matings. Each confirmed F_1 plant from each mating was harvested individually to obtain its F_2 seed progeny.

The F_2 seeds harvested from the individual, greenhouse-grown F_1 plants of each mating were planted in a field nursery located on the East Campus on 9 May 2008, along with seed harvested in the 2006-2007 winter greenhouse activity from the male and female parents and putative F_1 seeds (if available) from each mating. A 30-m long row was used in the on-campus soybean nursery for the parents, F_1, and F_2 population of each of the six matings. The six north-south rows were spaced 60 cm apart, with two rows of the Nebraska cultivar, NE3001, bordering the east and west sides of those rows. In each 30-m row, the first 2.30 m was planted with 30 F_2 seeds harvested from a greenhouse-grown F_1 plant, each seed spaced precisely 7.5 cm apart. The next 0.31 m was planted

with four similarly spaced female parent seeds and then the next 0.31 m was planted with four similarly spaced male parent seeds. Sixty F_2 seeds were planted (again, at a 7.5 cm spacing) into the next 4.60 m of row space, followed by planting of four female parent seeds (similar spacing) into the next 0.31 m of row space, and by four male parent seeds in the next 0.31 m of row space. This planting pattern of 60 F_2 seeds followed by four female parent seeds and four male parent seeds was repeated three additional times. However, at the third instance of this planting pattern, four putative F_1 seeds (if available) were included between female and male parents (similar spacing) in each row. Thirty F_2 seeds were planted in the final 2.30 m of each row. Enough F_2 seeds were planted to reach a goal of 300 F_2 seeds per mating. In those mating with fewer than 300 total F_2 seeds, any remaining unused row length was space-planted with seed of female parent from those crosses. The beginning of each of the six rows was planted with a 60-cm border of female parent seeds, and end of each of the six rows was planted with a 60-cm border of male parent seeds. After planting, the nursery was drip irrigated as needed to minimize plant water stress during the low-rainfall 2008 season.

Regarding Richie (2003), she used only 120 F_2 plants in her study. In this study, the power for detecting QTL was to be increased by doubling the population size. A final total of 240 F_2 plants, representing each six PI x agronomic cultivar matings, was desired. All F_2 plants in each of six matings that survived to maturity were carefully gathered one-by-one from a given population row and individually threshed to obtain their $F_{2:3}$ seed progeny. The F_2 plants in a given row (i.e., mating) matured within a few days of each other, with only minor maturity differences arising apparent within any mating and likely due to within-row microclimate, or soil type, variability. Each $F_{2:3}$ seed progeny was

labeled (barcoded) with a population number identical to the respective female parent code number (Table 1) and assigned an $F_{2:3}$ packet number (i.e. 1, 2, 3…, etc.) based on harvested number of F_2 plants from the beginning to the end of the row. The in-row female and male parent plants were also individually threshed and labeled with a parent code number (Table 1). The 60-cm border row of female parent plants and the row of male parent plants that bordered the south and north side, respectively, of each population were bulk-threshed. Each individual $F_{2:3}$ seed packet was sieved to remove threshing debris and split or damaged seeds. Moldy seeds were also discarded, as were unripe green or immature seeds. Any $F_{2:3}$ seed packets with an insufficient seed number (less than 30 seeds) for phenotypic seed protein and oil analysis were discarded from each population.

Phenotypic Trait Evaluation, Measurement, and Analysis

During the fall of 2008, the $F_{2:3}$ seed progenies were evaluated for protein (and oil) content using a near-infrared reflectance (NIR) analyzer (Infratec model 1255 NIR Food and Feed Grain Analyzer, Ultra Tec Manufacturing Inc., Santa Ana, CA) located at the Stewart Seed Laboratory, located on the East Campus. The Infratec analyzer evaluates whole seed samples simultaneously for seed protein, oil, moisture, and fiber. The measurements were based on reflectance of electromagnetic radiation in the near infrared region of the spectrum. In a typical NIR measurement, whole seed from a packet is placed into a sub-sample cup, with seed above the circular cup rim gently brushed off, thus leaving approximately 15g of soybean seed existing in a cup, not much below and mostly level with the cup rim. The NIR analyzer will examine either one sub-sample, or up to five consecutive sub-samples of a given seed sample. In this study, a one-cup

sampling was chosen as the method of NIR analysis, mainly because the $F_{2:3}$ seed numbers it accommodated resulted in the measurement of a larger fraction of $F_{2:3}$ progenies than would be possible with a two-or multiple-cup sample. For each population of $F_{2:3}$ seed progeny, male and female parent seed samples were also evaluated for seed protein, oil, and moisture using NIR. The data generated by the analyzer were output on a zero moisture basis.

The NIR analysis of a given mating was accomplished in the following manner: The seed packets for each assay were arranged, ordered, and analyzed as follows: four check samples (i.e. soybean cultivars IA2021, Macon, plus high protein line 1, and high protein line 2 of known average seed protein content of 490 and 510 g kg^{-1}, respectively), followed by (up to 20) individual female parent progeny, then by (up to20) individual male parent progeny, then by some putative $F_{1.2}$ progeny (if available), then by all $F_{2:3}$ progenies of the mating (arranged in order of the packet number), and finally a repeat assay of the (up to 20) female and (up to 20) male parents. The four check samples were used at the beginning of each assay to ensure that the NIR analyzer operated within performance standards.

After one complete replicate of the NIR-measured protein data was obtained for all $F_{2:3}$ progeny of each of the six populations, the $F_{2:3}$ seed progeny in each mating were ranked from lowest to highest based on the NIR-determined seed protein value. However, for the purpose of deciding whether a complete or partial second-replicate run of phenotyping was necessary in this experiment (discussed later in the Results and Discussion section), *all* $F_{2:3}$ progenies in just one population (1143) were re-assayed with NIR to provide a *complete* second-replication of seed protein content data in this mating.

In contrast, for other five populations, only those $F_{2:3}$ progenies occupying the lowest and the highest protein quintiles (20% in each tail, but in total representing 40% of the entire population) were re-assayed, thereby generating a two-replicate data set, but only for those $F_{2:3}$ progenies occupying the lowest and highest quintile fractions of the $F_{2:3}$ protein distribution. This latter process is referred to hereafter as a partial two-replicate phenotyping. During the second replicate assay of these six populations, a few odd instrument-reported outliers or extremely unusual values from the first replicate were re-examined. Ordinarily, if the re-examined third value was in agreement with the first or second replicate value, the re-examined value was substituted for the outlier value. However, such substitutions were rare.

The $F_{2:3}$ progenies within each quintile were then ranked from low to high seed protein content using the quintile-specific two-replicate protein mean values. Within the low quintile group, those $F_{2:3}$ progenies in the lowest half of the group were selected to represent the lowest decile fraction of the entire $F_{2:3}$ population. In contrast, within the high quintile group, those $F_{2:3}$ progenies in the highest half of the group were selected to represent the highest decile fraction of the entire $F_{2:3}$ population. Only these decile fractions of progenies with the lowest and highest seed protein contents were selected for the selective SNP genotyping assays. Figure 1 illustrates the generation advance and phenotyping protocols used for each population.

Leaf Collection and DNA Extraction Procedures

Parents:

Leaf collection and DNA extraction with respect to the parental lines were completed during the summer of 2007. Approximately 5 g of young leaflets material was collected from each parental plant and put in a plastic bag. Leaf tissue was placed on ice during sampling and transferred to the laboratory where it was stored at -20ºC until the tissue could be de-moisturized in a lyophilizer and then ground. The DNA extraction was performed by the mini-extraction CTAB method (Saghai-Maroof et al., 1984). Half of a 1.5 mL eppendorf tube was filled with lyophilized, ground leaf tissue and 800 µL of CTAB buffer containing 1% β-mercaptoethanol. Incubation at 65°C for 1 hour was followed by two chloroform-octanol (24:1) extractions, precipitation with cold isopropanol, three alcohol rinses [76% ethanol-0.2*M* NaOAc, 76% ethanol-10m*M* NH4OAc and 70% ethanol], and resuspension in 1X TE pH 8.0 buffer. The DNA was then quantified with a spectrophotometric optical density (OD) measurement at a wavelength (λ) of 280 nm. Then diluted DNA stock sampler to 20ng μL^{-1} for use in subsequent marker analyses. A total of 10 µL of the DNA suspension was also run on a 1% (w/v) agarose gel to check for DNA quality and to verify concentration.

F_1 Progeny:

One young trifoliolate leaf per plant of each mating-tagged F_1 plant (grown in the 2008-2009 greenhouse) was collected and placed into a 96-deep well collection plate, with care taken to ensure that plate well number matched with the tagged F_1 plant. During leaf collection, the tissue was placed on ice, transferred to the laboratory, and later frozen

and stored at -20º C until subsequent DNA extraction. For DNA extraction, 96-deep well collection plates were taken out from -20ºC and immediately placed on ice. Then, three of 2.0mm Zirconia beads (BioSpec Products, Inc.) were added into each well for grinding purpose, and then 400 µL of CTAB (2X) with β-mercaptoethanol solution was added into each well. Leaf tissue was ground using a mini-bead shaker for approximately 3 min. The F_1 DNA was subsequently extracted using a DNA plate extraction protocol modified from that described by Saghai-Maroof et al. (1984). For the modified protocol, all solution volumes were reduced to half that indicated in the original protocol. Each sample of DNA was re-suspended with 100 µL of TE buffer. The DNA concentration of each sample was measured using a spectrophotometer at a wavelength (λ) of 260 and 280 nm. The stock DNA samples were diluted to 20ng μL^{-1} for use in subsequent marker analysis.

F_2 Progeny:

F_2 leaf collection was accomplished during the summer of 2008, but DNA extraction was deferred until after the $F_{2:3}$ seed progeny had been phenotyped for selection purposes. Individual F_2 plants were tagged with a labeled tag (specifying population number and F_2 plant number) after plants had produced at least two trifoliolate leaves. One very recently developed trifolioate leaf of about 2 cm in length was removed from the stem tip of each F_2 plant in each population. The trifoliolate leaf sample was rolled between the thumb and forefinger before being inserted into a labeled well of a clean 96-deep well collection plate (using a forceps), being careful to ensure that labeled plate well matched the tagged F_2 plant number. The 96-deep well plates were labeled starting at the bottom left (column 1, row “H”) with sample number 1, and continuing to

the top left (column 1, row “A”) with sample number 8, then starting again at the bottom left of next column with sample number 9, and continuing to the top left, and following same pattern until the last number of each plate (i.e., number 96 for plate 1, number 192 for plate 2, and number 286 for plate 3). The total of three 96-deep well plates per population were labeled with the population number and assigned a plate ID number of 1, 2, and 3. Plates containing leaf tissues were placed on ice during sampling and transported to the laboratory where they were stored at -20°C until further marker analysis (i.e. SNP analysis).

After the completion of phenotying and the subsequent identification of $F_{2:3}$ progenies in the lowest and highest deciles, the corresponding F_2 trifoliolate leaf samples were retrieved from the three 96-deep well collection plates by matching the labeled F_2 plant well number with the $F_{2:3}$ progeny number in the decile fractions. One-half of the collected leaf sample was transferred into a well of a new 96 deep-well plate. Both the source and destination 96-well plates were kept on ice while transferring leaf tissues. Owing to the limited space in the 96 deep-well plate, only 22 samples from each decile group (i.e., highest and lowest $F_{2:3}$ progeny of the 220 or more F_2 plants in each population) were chosen for DNA extraction. The destination plates therefore contained two populations per plate (i.e., DNA samples for 22 high and 22 low protein samples per population, plus DNA samples of the female and male parents, and one F_1 plant from each population). For this plate layout, the 96-deep well plate was divided into two sections for two populations. The top four rows (i.e., rows A to D) of a plate were used for one population, and the bottom four rows (i.e., rows E to H) of that plate were used for the another population. In each section, wells were loaded starting at the first well

moving left to right with samples from the low decile (sample numbers 1 to 22), followed by samples from the high decile (sample numbers 1 to 22), and then a female parent, male parent, and F_1, in that order. Plates were shipped on dry ice to a USDA laboratory in Beltsville, MD, for robotic DNA extraction. The extracted DNA was then subjected to the SNP marker analysis, also completed at USDA laboratory, Beltsville, MD.

SSR Marker Analysis

Screening for Parental Polymorphic Markers:

Molecular marker analysis of the parents was accomplished using the published base-pair (bp) sequences for PCR primer pairs of known SSR markers. A group of SSR markers (52 SSR markers) was selected on the basis of availability, probable parental polymorphism, and more importantly, closeness to the reported map position of the strong seed protein QTLs. The strongest protein QTLs are (in additive effect order from high to low) located on LGs I (Chr 20), E (Chr 15), H (Chr 12), M (Chr 7), A1 (Chr 5), C1 (Chr 4), and F (Chr 13), respectively (Appendix Table 1). Relative to availability and probable polymorphism, SSRs were initially selected on the basis of being polymorphic in all three published maps of the Utah map populations: Archer x Noir, Archer x Minsoy, and Noir x Minsoy (*G. max* x *G. max* populations) (Cregan et al., 1999). In the initial portion of this study, all parents were screened with those 52 SSR markers to identify parental polymorphisms. SSR primers were synthesized according to the sequences published on the SOYBASE website (SoyBase, 2009). Based on this screening, 13 markers were found to be parentally polymorphic for most of the six matings, and used in the next step.

F_1 Confirmation:

The initial group of 52 SSR markers served as a primary screen to test for polymorphism between the six PI (female) parents and the three cultivar (male) parents. The 13 polymorphic SSR markers from the parental screen that mapped close to protein QTL in LG-I (Chr 20), and –E (Chr 15) were used for the F_1 confirmation. Nevertheless, selecting SSRs polymorphic for all or at least most of the female parents that were mated to their specific male parent was the first goal. As a result, relative to those 13 polymorphic markers, three markers [Satt384 (LG-E), Satt496 (LG-I), and Satt651 (LG-E)] were ultimately chosen because they provide a marker genotyping means of F_1 confirmation in all six populations. Marker Satt384 was used in 1139 x 1181M and 1146 x 1181M populations; Satt496 was used in 1121 x 1137M, 1122 x 1137M, and 1143 x 1181M population; and Satt651 was used in the 1076 x 1106M population.

PCR Amplification and Gel Electrophoresis:

Amplifications were performed in 10-µL aliquots of reaction mix containing of 0.2 µM of each of the paired forward and reverse primers, 1.5 mM $MgCl_2$, 5X of reaction buffer (50 mM KCl, 10 mM Tris-HCl, 0.1% triton X-100), 0.7 units of DNA Taq polymerase, 0.15 mM of each of the four dNTPs, and 50 ng of genomic DNA. To prepare each PCR reaction, the PCR mix was loaded into a 96-well reaction plate. Additionally, a 50-ng sample of genomic DNA (one from each parent, F_1) was loaded into an individual well previously loaded with PCR reaction mix. A polypropylene-based film was used to seal the wells to minimize evaporation. The PCR reaction plate was then placed into a PTC-100 Programable PCR thermocycler (MJ Research, Watertown, MA). Initial steps

were performed at 94ºC for 25 s, followed by 32 cycles of denaturing of 94ºC for 25 s, annealing of 47°C for 25 s, extension at 68°C for 25 s, and followed by a final extension step at 72°C for 3 min as well as an incubation at 4°C.

A 2.5% (w/v) agarose gel (AMRESCO, Solon, OH, Super Fine resolution agarose) electrophoresis was used to separate the PCR product. To begin, 2 µL of loading buffer (6X) were added to the PCR products, and a 10 µL sample was loaded on the gel. A 0.5X TBE solution served as a running buffer. The gel was run at a constant 70 V for 5 h. Ethidium bromide (10 mg mL^{-1}) was prepared to visualize the bands under exposure to UV light. Gels were stained for 10 min in the ethidium bromide solution, and then de-stained for 20 min in 2 L of distilled, deionized water. Finally the banding pattern was visualized under UV light and the image captured on thermal-sensitive photography paper from an image analysis system (GelDoc2000, BioRad, Hercules, CA). Slight modifications in this procedure were necessary to optimize conditions for particular primers and equipment. All SSR amplicons fell within the 2642- to 50-bp size range of the markers in molecular weight standard XIII (Boehringer Mannheim). Bands were scored using ‘A’ to represent a P1 homozygote of the male cultivar type, ‘H’ to represent a heterozygote, and ‘B’ to represent a P2 homozygote of the female germplasm accession type, relative to the two alleles for each SSR locus.

SNP Marker Analysis

A high-throughput method was needed to determine the genotypes for a large set of SNPs in many individuals in an efficient and cost effective way. The Illumina genotyping platform enables these large scale genomic studies by combining several

technologies, which include a miniaturized array of individual arrays (Sentrix Array Matrix), a high resolution confocal scanner (BeadArray Reader) within which the arrays are read, and a highly multiplexed genotyping assay (GoldenGate assay) (Gunter et al., 2005). These combined technologies maximize ease of use, result in good quality genotype data, and have a low cost per data point. The GoldenGate assay is designed to easily multiplex many loci in a single reaction while maintaining the ability to target any particular SNP of interest. (Hyten et al., 2008) used the Illumina GoldenGate assay to demonstrate the multiplexing from 96 to 1,536 soybean SNPs in a single reaction over a 3-day period using genotypic DNA samples from three soybean RIL mapping populations.

In this study, the high multiplex capability of GoldenGate assay was the technique employed for the SNP genotyping of $F_{2:3}$ progenies in the low and high decile groups, the high and low protein parents, and one F_1 individual using the Illumina Genotyping Platform (Illumina Inc., San Diego, CA). The GoldenGate assay was developed for 1536 SNPs that were distributed throughout the 20 chromosomes of soybean genome (Hyten et al., 2010). A 1536-SNP marker assay would be expected to be parentally polymorphic for a sufficient number loci (i.e., ideally about 400 for the soybean genetic map) to provide a genotyping density needed to be successful in one-step QTL discovery strategies. The GoldenGate is conducted in assay 96-well plate over a 3-day period. The steps of assay are outlined as follows: the first day consists of (i) making activated DNA, (ii) adding DNA to oligonucleotides and hybridizing (iii) extension, ligation, and cleanup, and (iv) universal PCR cycle at 1,536-plex; the second day consists of (i) binding PCR product, eluting dye-labelled strands, and preparing for hybridization and (ii) hybridizing to the

Sentrix®Array Matrix or BeadChip; and finally, the third day consists of (i) washing and drying the Array Matrix or BeadChip and (ii) imaging Array Matrix or BeadChip (Illumina, 2009). All steps of SNP analysis were performed by personnel at the Soybean Genomics and Improvement Laboratory, USDA-ARS, BARC-West, Beltsville, MD.

The SNP loci were dispersed over the soybean genome (Fig. 2). Not all of the 1536 SNP loci were expected to be parentally polymorphic in each mating. Still, about 30-50% of the 1,536 SNPs (about 460-760 SNPs) were expected to be polymorphic in any of the given six matings which, depending upon the distribution of polymorphic SNPs across the genome and particularly within each LG, would be sufficient markers for QTL detection. The automatic allele calling for each locus was accomplished by using GenCall software (Illumina Inc., San Diego, CA). An “A”, “B”, and “H” genotype coding scheme was used for a homozygote of the normal-protein male parent allele type (A), a homozygote of the germplasm accession high-protein female parent allele (B) type, and a F_1 heterozygote of both alleles (H), respectively. A dash (-) was used for ambiguous or missing genotype calls.

Data Analysis

With respect to phenotypic data for the two main trait (protein and oil), distributional normality was analyzed using SAS (Statistical Analysis Systems, Cary, NC) version 9.1 software (SAS, 2002). Phenotypic correlations (i.e. seed protein content and oil) were calculated using the PROC CORR procedure.

Since, the individual F_2 plants (i.e., experimental units) cannot be replicated in this experiment, the genotype (G), environment (E), G x E interaction, and random

experimental error cannot be directly estimated from $F_{2:3}$ protein values using an expected mean squares computation based on an from analysis of variance. However, parents and F_1 genotypes can be, and were, field-replicated to obtain an estimate of the environmental and measurement variance estimated using this formula;

$$\sigma_e^2 = \frac{1}{3}(\sigma_{ph}^2 + \sigma_{pn}^2 + \sigma_{F1}^2)$$

where;

σ_{ph}^2 is the phenotypic variance of the self seed progenies obtained from replicated high protein parent plants

σ_{pn}^2 is the phenotypic variance of the self seed progenies obtained from replicated normal protein parent plants

σ_{F1}^2 is the phenotypic variance of the self seed progeny obtained from F_1 plants.

The genetic variance of the population of $F_{2:3}$ seed progenies derived from about 250 F_2 plants was estimated using the formula;

$$\sigma_g^2 = \sigma_p^2 - \sigma_e^2$$

where;

σ_p^2 is the phenotypic variance of the ~250 $F_{2:3}$ progenies in the given population.

Subsequently, a broad sense heritability (H^2) estimate was obtained by using the formula described by Kearsey and Pooni (1996) as follows;

$$H^2 = \frac{\sigma_g^2}{\sigma_p^2}$$

To conduct a preliminary test for any significant association between each SNP marker and a presumed nearby protein (and oil) QTL, the decile tail fractions of each F_2

mapping population were evaluated for an allele frequency difference at each parentally polymorphic SNP marker locus. The frequency of the "A" allele contributed by agronomic (ordinary protein) parent cultivar in the low decile and high decile fractions was evaluated using the *t*-test described by Bernardo (2002);

$$t = \frac{P_{A_{low}} - P_{A_{high}}}{\sqrt{\frac{p_{A_0}(1-p_{A_0})}{2N_L} + \frac{p_{A_0}(1-p_{A_0})}{2N_H}}}$$

where, $P_{A\ low}$ is the "A" allele frequency in the low decile $F_{2:3}$ progenies, $P_{A\ high}$ is the "A" allele frequency in the high decile $F_{2:3}$ progenies, N_L is the actual number of $F_{2:3}$ progenies in the low decile group, and N_H is the actual number of $F_{2:3}$ progenies in the high decile group.

Because this is an F_2 population, the expected frequency of the "A" allele is 0.5 in the above formula. The null hypothesis tested here is $P_{Alow} = P_{Ahigh}$, which if not rejected, signifies that the marker allele "A" at the given SNP locus was not linked to any nearby segregating seed protein QTL whose alleles could have contributed to the contrast between the two $F_{2:3}$ phenotypic tails. The alternative hypothesis is $P_{Alow} > P_{Ahigh}$, which if accepted, signifies that the marker allele "A" of the SNP locus had a significantly higher frequency in the low decile than in the high decile. The null hypothesis would ordinarily be a one-tailed test if one expects that the ordinary protein cultivar parent <u>not</u> to contribute a high protein allele at some QTL. However, two-tailed test is usually conducted because that expectation may not be a reliable expectation.

The SNP allele will be "coupled" (because of linkage) directionally with a change in the frequency of the QTL "A" allele being mostly in the low protein phenotypic tail, if

the cultivar “A” marker allele is linked to a low protein QTL allele. If the "high" parent is expected to be fixed for all high protein alleles, then using a one-tailed t-test is not an issue. However if the "low" parent carries alleles at some QTLs that condition high protein, then the use of a one-tailed test is problematic. Overall, we therefore cannot consider that the hypothesis tests in selective genotyping are intrinsically one-tailed unless we absolutely know beforehand that the parental origin of all of the favorable alleles linked to the markers being tested (e.g., as in a candidate gene approach) (Gallais et al., 2007; Sen et al., 2009).

Power Calculation in Selective genotyping

As discussed earlier, it is desirable to know if there is sufficient power to detect QTLs with additive effects that may range from effects as high as that of the LG-I (Chr 20) protein QTL (a QTL with known large additive effect of 12 g protein/kg seed) to effects as low as ½ or ¼ of that QTL’s effect. The detection of the QTL requires large sample sizes to achieve reasonable power (Soller et al., 1976). In this study the R/qtl software was used to help us calculate the power and determine an appropriate F_2 sample size to use in all six F_2 populations, for the detection QTLs with additive effect size of a minimum of 12 kg protein/kg seed (i.e., 1.2 percentage points) in seed protein content, assuming a desirable statistical power of 0.80.

Linkage Mapping and QTL Analysis

The program MAPMAKER/EXP VER 3.0 (Lander et al., 1987; Lincoln et al., 1993) was initially used for determining genetic linkage and distance between markers

because it uses a maximum-likelihood method for computation of recombination values between adjacent marker loci. The phenotypic data file contained several rows of phenotypic data for each of the $F_{2:3}$ populations. The first row of phenotypic data file were the $F_{2:3}$ progeny ID numbers, with the remaining rows containing phenotypic data. The genotype file contained the genotypes for only the 22 high and 22 low protein $F_{2:3}$ progenies identified in the "decile" fractions of the phenotypic tail analysis. All other progeny genotypes were treated as missing (-) values. Data files were imported into MAPMAKER/EXP as a text file. The parameters set for linkage mapping included using the Kosambi mapping function. The group command linkage criteria was set to a LOD score (logarithm to the base 10 of the likelihood odds ratio) of greater than 3.0. The group command maximum genetic distance was set to 37.2 centimorgan (cM). The error detection probability level was set at 1%. From files created in MAPMAKER/EXP, QTL mapping could be performed by using MAPMAKER/QTL VER 1.1 software (Lander et al., 1987; Lincoln et al., 1993). However, newly available software, R/qtl, was preferred for QTL mapping because it offers greater convenience, command and output, flexibility (Broman and Sen, 2009). The MAPMAKER/EXP population-specific raw and maps data files, which contain the marker genotypes/phenotypes and the saved linkage mapping results were imported into R/qtl using a command recommended by the author of the R/qtl software (Karl Broman, personal communication). The initial QTL analysis was performed using the single marker regression option, but the final QTL analysis was accomplished using the interval mapping option of the R/qtl software.

Only $F_{2:3}$ progenies within the extreme decile tails of a population's first replication seed protein distribution were genotyped. Technically, the decile tails are

actually the extreme half of the lower and upper quintiles—the only quintiles with two replicate values, and not truly the 1-replication seed protein tails. This was done to ensure two-replication means were the basis for selective genotyping. However, because QTL effects are biased in selectively genotyped populations, maximum likelihood-based QTL detection using interval analysis requires a phenotype for every individual in the population (including those without SNP marker genotypes) (Lander and Botstein, 1989). The interval analyses were first conducted on the first replication phenotypes of protein (and oil).

R/qtl was first used to check each population for marker segregation distortion using a Chi-square goodness of fit test to compare the observed segregation to expected 1:2:1 mendelian F_2 ratio. However, multiple testing (i.e. approximately 500 markers in each population) renders a single Chi-square test significance criterion of alpha=0.05 unsuitable, so a Bonferroni-adjusted significant criterion of alpha=0.05/500=0.0001 was used. Markers whose chi-square values exceeded the latter criterion were discarded. For the remaining SNP markers, the markers were ordered to be in the same order as the extensively rippled marker order that was recently published for each LG in soybean genetic map version 4.0 (Hyten et al., 2010). However, we also ran the ripple and switch marker order commands in R/qtl to compare the results using marker orders of the Hyten et al. (2010) genetic map. This data will be discussed in Results and Discussion section). Using R/qtl software, a single marker regression analysis (MR) was conducted using the scanone function as a preliminary QTL detection procedure. However, the final QTL detection involved the use of interval analysis using standard maximum likelihood interval mapping approach (via EM algorithm). For interval analysis, the sim.geno and

effectplot commands were used to estimate the A, B, and H genotypic values at each marker. In selective genotyped F_2 populations, a stratified permutation procedure is required to generate the appropriate a population-specific genome-wise LOD score significance criterion for use in determining the statistical significance of a QTL peak. Therefore, 1000 permutations were performed in each population for the traits of seed protein (and oil). Additive (a) and dominance (d) effects of each presumptive QTL were calculated from the genotypic values of the marker most tightly linked to that QTL. The percentage of variance explained by the QTL (i.e., R^2) was calculated using the formula $R^2 = 1\text{-}10^{(-2/n)LOD}$ where "n" represents number of phenotyped $F_{2:3}$ progenies in each population and LOD was the $\log_{10}$ likelihood ratio at the QTL peak.

Additional QTL analyses were also performed in $F_{2:3}$ population 1143. In this population, the seed protein and oil were measured with NIR when the seed was drier and then when the seed was moister than the original first replication NIR measurement. This experiment was performed to determine if the seed moisture that was greater and lesser than original influenced the NIR-based estimates of the $F_{2:3}$ progeny seed protein (and oil) values, and also the QTL analysis.

RESULTS AND DISCUSSION

Authentic F_1 Confirmation and F_2 Development

The F_1 plants of six $F_{2:3}$ populations were evaluated with SSR markers from the regions on LG-I (Chr 20) and LG-E (Chr 15), where the two most well-known seed protein QTLs are known to map. Based on the initial screen of 52 SSR markers, 13 markers were found to be parentally polymorphic for one or more of the six populations. Of these 13 markers, three markers [Satt384 (LG-E), Satt496 (LG-I), and Satt651 (LG-E)] were ultimately chosen because these three were sufficient for marker genotyping F_1 plants in all six populations. SSR marker Satt384 was used to verify the authenticity of F_1 plants in populations 1139 and 1146. SSR marker Satt496 was used in populations 1121, 1122 and 1143. SSR marker Satt651 was used in population 1076. Using these markers, the total number of F_1 plants confirmed as authentic in populations 1076, 1121, 1122, 1139, 1143, and 1146 were 11, 14, 18, 17, five, and five, respectively (Table 3). The F_1 plants of some populations were also additionally confirmable as authentic based on F_1 classical marker genotype (Table 2), although with less certainly due to phenotypic variation in pigment intensity. The total number of F_2 seeds obtained from the populations 1076, 1121, 1122, 1139, 1143, and 1146 were 703, 670, 1544, 1399, 415, and 2190 seeds, respectively (Table 3).

With respect to population development, the six female parents ranged from MG II, III, and IV, so three male parents of corresponding MG were used as mating partners. Although it would have desirable to use one male parent to generate six half-sib matings and thus directly compare the effect of six PI alleles at any given QTL against one

"standard" allele, maturity segregation in a high versus low protein segregating population could introduce confounding effects of maturity QTL x protein QTL interaction, which would likely have occurred because of seed-fill timing differences (relative to seasonal temperatures) between early and late maturity F_2 plants.

Phenotypic Data Analysis

Population 1076

The replicate one seed protein values for the 188 $F_{2:3}$ progenies of population 1076 exhibited continuous variation (Fig. 3). The progeny seed protein distribution was normally distributed, judging from the non-significant Shapiro-Wilk test value (hereafter this test will be referred to as the SW test), although the distribution was slight leptokurtic (sharper peakedness) (+0.32) and a slightly leftward skewed (-0.23) (Table 3). The parental means indicated that the high protein PI parent 1076 and the low protein agronomic parent 1106M differed by 43 g kg^{-1} in seed protein content (Fig. 3). The seed protein of the progeny ranged from 371 to 483 g kg^{-1}, with a mean protein for the population of 431 g kg^{-1} and a standard deviation of 18 g kg^{-1} (Table 4). The progeny seed protein mean (431 g kg^{-1}) was somewhat lower than the high protein PI parent 1076 mean (447 g kg^{-1}), but much higher than the low protein agronomic parent 1106M mean (404 g kg^{-1}). For traits that have an additive only type of inheritance (i.e., no dominance effects), one would expect the $F_{2:3}$ progeny mean to have exhibited a mid-parent value of about 442 g kg^{-1}. It is possible that genotypes with a very high protein potential may not have been able to express that potential proportionately as well as genotypes with a lower potential.

Although seed oil was not the focus of this selective genotyping project, phenotypic and genotypic variation in seed oil is almost invariably influenced in an inverse manner by variation in seed protein. For that reason, the progeny seed oil values are worthy of examination here.

The seed oil content of the 1076 $F_{2:3}$ population also showed continuous variation, and a histogram of the oil content is shown in Appendix Fig. 1. The high protein but low oil PI parent 1076 and the low protein but high oil agronomic parent 1106M differed by 23 g kg^{-1} in seed oil content (Table 5), The seed oil of the $F_{2:3}$ progeny ranged from 121 to 219 g kg^{-1}, though the mean seed oil content was 165 g kg^{-1} which was identical to that of the high oil parent, suggesting that dominance may play a role in the inheritance of seed oil in this population. Note that the seed moisture mean of $F_{2:3}$ population 1076 was 8.1% (Table 6). The $F_{2:3}$ distribution of seed moisture was not normal in this (or any other) population. The reason for this is not known nor readily explainable.

Population 1121

The progeny seed protein distribution exhibited a deviation from normality distribution, partly because of a slight platykurtic tendency (flatter peakedness) (-0.26) and a modest rightward skewness (+0.56) (Fig. 3 and Table 3). Nevertheless, this deviation from normality was moderated to be slight and likely not serious enough to affect QTL analysis. Means, standard deviations, ranges, and parental values for seed protein content measured in the $F_{2:3}$ population 1121 and parents are presented in Table 4. The high protein PI parent 1121 had a mean seed protein of 432 g kg^{-1}, the low protein agronomic parent 1137M averaged 401 g kg^{-1} seed protein. Seed protein in the $F_{2:3}$

progeny ranged from 377 to 472 g kg^{-1}. The mean seed protein content in the $F_{2:3}$ progeny was 419 g kg^{-1}, which was slightly lower than that of the high protein PI parent 1121, but substantively higher than that of the low protein agronomic parent 1137M. In this population, the $F_{2:3}$ progeny mean was closer to the mid-parent value, suggesting a mostly additive form of inheritance.

The seed oil content of the $F_{2:3}$ population showed continuous variation as is evident in the histogram of the oil content presented in Appendix Fig 1. The PI low oil parent 1121 and high oil agronomic parent 1137M differed by 17 g kg^{-1} in seed oil content. The seed oil of the progeny ranged from 137 to 229 g kg^{-1} (Table 5), and the mean oil content of the population was 187 g kg^{-1}, which was near the mid-parent value. The moisture mean of $F_{2:3}$ population 1121 was 8.9% (Table 6), which was not much different from that of the population 1076.

Population 1122

The $F_{2:3}$ population 1122 showed continuous variation for seed protein content (Fig. 3) and seed oil content (Appendix Fig. 1). The seed protein distribution of the $F_{2:3}$ population 1122 showed a normal distribution with a very slight leptokurtic (sharper peakedness) tendency (0.07) and a very slight leftward skewness (-0.02) (Table 4). The mean seed protein content was 423 g kg^{-1} for the high protein PI parent 1122 and 393 g kg^{-1} for the low protein agronomic parent 1137M. The mean seed protein of $F_{2:3}$ population 1122 was 419 g kg^{-1}, which was just slightly lower than high protein PI parent 1122's mean seed protein, but clearly higher than the low protein agronomic parent

1137M's mean seed protein, which would nominally suggest dominance of higher seed protein. The seed protein content of the progeny ranged from 373 to 459 g kg^{-1} (Table 4).

With respect to oil content (Table 5), the low oil PI parent 1122 averaged 190 g kg^{-1} seed oil content, whereas the high oil agronomic parent 1137M averaged 195 g kg^{-1}, a different of only 5 g kg^{-1}. However, the seed oil content of the $F_{2:3}$ progeny ranged from 148 to 224 g kg^{-1}, revealing considerable transgressive segregation for seed oil in this mating. The moisture mean of $F_{2:3}$ population 1122 was 8.2% (Table 6), again not differing much from the other populations.

Population 1139

The $F_{2:3}$ population 1139 showed continuous variation for seed protein content (Fig. 3) and seed oil content (Appendix Fig. 1). Both skewness and kurtosis of the seed protein content were less than 1.0 (0.06 and -0.62, respectively), and the SW test value suggested that the segregation of the seed protein content trait fit a normal distribution model (Table 4). The mean seed protein content was 447 g kg^{-1} for high protein PI parent 1139 and 411 g kg^{-1} for low protein agronomic parent 1181M. The mean seed protein of $F_{2:3}$ population 1139 was 434 g kg^{-1} , which was slightly lower than high protein PI parent 1139's mean seed protein but clearly higher than the low protein agronomic parent 1181M's mean seed protein. Though this would again nominally suggest that higher protein is dominant, the higher protein parent may possibly not be expressing its high protein potential. The seed protein content of the progeny ranged from 391 to 476 g kg^{-1}.

For seed oil content, phenotypic data analysis indicated that low oil PI parent 1139 and high oil agronomic parent 1181M differed by 37 g kg^{-1} (Table 5). The low oil

PI parent 1139 averaged 159 g kg^{-1} seed oil content and high oil agronomic parent 1181M averaged 186 g kg^{-1}. Seed oil content of the progeny ranged from 120 to 229 g kg^{-1}, but had a mean of 170 g kg^{-1} which was close to a mid-parent value. A histogram of the progeny seed oil distribution is shown in Appendix Fig. 1, and appears to be a normal one. The moisture mean of $F_{2:3}$ population 1139 was 8.4% (Table 6).

Population 1143

The seed protein distribution of $F_{2:3}$ population 1143, when evaluated at an average seed moisture content of 8.5% (Table 6), exhibited a normal distribution with a modest platykurtic tendency (-0.43) and a slightly rightward skewness (+0.12) (Fig. 3). The mean phenotypic data for seed protein content revealed that the high protein PI parent 1143 averaged 437 g kg^{-1}, whereas the low protein agronomic parent 1181M averaged 411 g kg^{-1}, a different of 26 g kg^{-1} (Table 4). The mean seed protein of $F_{2:3}$ population 1143 was 431 g kg^{-1}, which was slightly lower than that of high protein PI parent 1143's mean seed protein but clearly higher than that of low protein agronomic parent 1181M. The progeny seed protein content ranged from 378 to 471 g kg^{-1}.

Regarding seed oil content (Table 5), the low oil PI parent 1143 averaged 172 g kg^{-1} in seed oil content whereas the high oil agronomic parent 1181M averaged 187 g $kg^{-1,}$ a difference of 15 g kg^{-1}. Seed oil content of the progeny ranged from 103 to 229 g kg^{-1}. A histogram of the oil phenotype is shown in Appendix Fig. 1.

For the purposes of determining if moister or drier than normal seed influenced the NIR estimates of seed protein and oil, population 1143 was re-evaluated when the $F_{2:3}$ seed progenies seeds were made moister, and then again after seed had been dried in the

dryer room for 24 hours. As noted above, the ordinary NIR analysis was conducted with seed of 8.5% moisture (standard deviation of 1.6%), then evaluated again with NIR at a 10.6% seed moisture content (standard deviation of 0.8%), and again at a 7.0% seed moisture content (standard deviation of 2.0%) (Table 6). The seed moisture increased by about 2.1% from the 15 November 2008 date of the first NIR evaluation to the 15 July 2009 date of the second NIR evaluation, even though the seeds were stored in the room temperature (25°C). In buildings, the interior ambient humidity is lower in the fall-winter than in the spring-summer, so the seeds simply gained moisture by equilibration.

Seed protein distribution of $F_{2:3}$ population 1143 in both moist and dry seed showed a normal distribution with a platykurtic tendency (-0.07 for moist and -0.09 for dry seed) and a rightward skewness (0.21 and 0.19 for moist and dry seed, respectively) (Table 4). The mean phenotypic data for seed protein content revealed that the high protein PI parent 1143 averaged 425 g kg^{-1} with moist seed and 440 g kg^{-1} with dry seed, whereas the low protein agronomic parent 1181M averaged 402 g kg^{-1} with moist seed, and 415 g kg^{-1} with dry seed. The differences in the parental means were 23 and 25 g kg^{-1}, respectively. The mean seed protein of $F_{2:3}$ population 1143 of moist and dry seed was 419 and 433 g kg^{-1}, respectively. The mean seed protein of $F_{2:3}$ moist seed was slightly lower than the high seed protein PI parent 1143's mean seed protein. But the mean seed protein of $F_{2:3}$ dry seed was higher than the high protein PI parent 1143's mean seed protein. However, both seed conditions were clearly higher than the low protein agronomic parent 1181M's mean seed protein. The $F_{2:3}$ progeny seed protein content ranged from 374 to 460 g kg^{-1} for moist seed and from 380 g kg^{-1} to 488 g kg^{-1} for dry seed.

Relative to seed oil content (Table 5), the PI parent 1143 averaged 170 g kg^{-1} seed oil content and agronomic parent 1181M averaged 186 g kg^{-1} for moist seed, but respectively averaged 169 g kg^{-1} and 181 g kg^{-1} for dry seed. Seed oil content of the progeny ranged from 108 to 219 g kg^{-1} for moist seed and 99 to 227 g kg^{-1} for dry seed. Histograms of the oil phenotype for both moist and dry seed are shown in Appendix Fig. 1.

In conclusion, based on the results of the NIR estimation of seed protein and oil differences among moist, normal, and dry seed, one can conclude that the greater the seed moisture, the lower the seed protein as well as seed oil content. While it appears that with higher seed moisture, NIR analysis detects lower seed protein values; however, the seed oil content for both moist and dry seed conditions were only slightly lower than initial seed moisture conditions.

Population 1146

The seed protein distribution of $F_{2:3}$ population 1146 showed a normal distribution (Fig. 3), with a modest platykurtic tendency (-0.40) and a slight rightward skewness (0.14) (Table 3). The mean phenotypic data for seed protein content revealed that the high protein PI parent 1146 averaged 438 g kg^{-1} and low protein agronomic parent 1181M averaged 413 g kg^{-1}, a different of 25 g kg^{-1}. The mean seed protein of $F_{2:3}$ population 1146 was 424 g kg^{-1}, which about at the mid-point between the high protein PI parent and low protein agronomic parent 1181M.

With respect to seed oil content (Table 5), the low oil PI parent 1146 averaged 164 g kg^{-1}, whereas the high oil agronomic parent 1181M averaged 180 g kg^{-1}. The seed

oil content of $F_{2:3}$ progeny ranged from 135 to 231 g kg^{-1}. A histogram of the oil phenotype is shown in Appendix Fig 1. The seed moisture means of $F_{2:3}$ population 1146 was 8.8% (Table 6), which was not differing much from the other population.

Other Phenotypic Considerations

Replicate one and replicate two seed protein content

Individual replicate one and replicate two seed protein values for each $F_{2:3}$ progeny occupying the lowest quintile of the replicate one distribution (red square symbols in Appendix Fig. 2 graphs) and in the highest quintile (blue circle symbols) $F_{2:3}$ were coordinately plotted on the abscissa and ordinate, respectively. The two-replicate mean seed protein contents of each progeny were also plotted as criss-cross symbols along the 1:1 line of Appendix Fig. 2. The graph of each population revealed that the lowest and the highest seed protein progenies remained well-separated into two distinct extreme quintile clusters based on the two-replicate mean value distributions. The seed protein standard deviation of the six $F_{2:3}$ progenies populations varied from 14 to 19 g kg^{-1} (Table 4). The replicate one and replicate two seed protein values were slightly different for some $F_{2:3}$ progenies (i.e., as evidenced by some scatter around the 1:1 line in the Appendix Fig. 2 graphs); however, these differences did not make much difference in the rank order of the progenies that occupied the extreme quintiles. The ranking of the two-replicate seed protein means served as the criteria for choosing $F_{2:3}$ progenies to genotype. Only the 22 lowest and 22 highest seed protein $F_{2:3}$ progenies in the corresponding lowest and highest quintile fractions were chosen for selective genotyping purposes. The 44 selected progenies in each population accounted for somewhat more

than 20% of the total $F_{2:3}$ progenies, as can be seen by the percentages listed in the last column of Table 3. This is because even though 220 or more F2 plants were harvested, many F_2 plants did not produce sufficient F3 seed for the NIR analysis.

Seed Protein and Oil Content of Parents

Seed protein content ranged from 393 to 415 g kg^{-1} for male low protein agronomic parents and 423 to 447 g kg^{-1} for the high protein PI female parents. There was only a slight difference between the male parent seed protein values measured in this study compared to those values published in the National Genetic Resources Program (NGRP) database, which ranged 382 to 424 g kg^{-1} (Table 1). On the other hand, all of the female parent seed protein contents observed in this study were substantively lower (approximately 5.9 to 6.0 g kg^{-1} lower) than those reported in the NGRP database. These differences were first noted by Ritchie (2003). The 2008 Nebraska production environment is possibly less optimum for high seed protein expression than the production environments in which seed was produced for the NIR protein analysis reported in the NGRP. Alternatively, perhaps use of one-cup NIR seed samples resulted in proportionately lower protein values that what might had been the true values.

The female parent seed oil content ranged from 159 to 190 g kg^{-1}. The seed oil of the male parents ranged from 181 to 196 g kg^{-1}. The female parent seed oil values measured in this study differed by about 6 to 8 g kg^{-1} from the values published in the NGRP database (Table 1). This was also true for the male parent seed oil values. However, male and female parent seed oil contents in this study were somewhat higher than the NGRP database by 1 and 6 to 8 k kg^{-1}, respectively. A similar difference was

also reported in Richie (2003). Again, the 2008 Nebraska environment possibly may have affected the seed oil content differently compared to the environment of experiments reported in the NGRP, or possibly the NIR calibration for seed oil content does not work as well with small versus large seed samples.

Heritability

The heritability for seed protein content computed for each of these six $F_{2:3}$ populations indicated that some or much (27 to 85%) of the phenotypic variation was genetic. Chung et al. (2003) reported the heritability for seed protein was 89% in their population. Brummer et al. (1997) reported a range of heritabilities for seed protein in eight soybean populations they studied, from 56 to 92%, depending on the population. The heritability observed in the dissertation project $F_{2:3}$ populations of 1076, 1121, 1122, 1139, 1143, and 1146 were 63, 85, 27, 69, 56, and 67%, respectively. The seed protein heritability of $F_{2:3}$ population 1122 was low, primarily because seed protein phenotypic variance observed in the male parent (1137M) was quite high. Additionally, the seed protein variance among the F_1 plants available in each mating was also high. In the literature, seed protein heritability has generally been found to be greater than 80% (Thorne and Fehr, 1970; Shannon et al., 1972; Helms and Orf, 1998; Cober and Voldeng, 2000; Chung et al., 2003). Heritability estimates are usually not reliably determined without replications in both space (i.e., locations) and time (i.e., years). In this study, the heritability estimates were mainly computed for comparative purposes relative to the six populations. Theoretically, the seed protein standard deviation of the high protein PI female parent is supposed to be lower than that of the corresponding $F_{2:3}$ progeny

population because the female parents are homozygous. Nevertheless, in this study, the seed protein standard deviation values for the female parents (Table 4.) were generally greater than seed protein standard deviation of male parents and the corresponding $F_{2:3}$ progeny populations. These results suggest the high seed protein female parents had a more variable seed protein expression compared the low seed protein cultivar male parents, possibly because high seed protein expression is affected proportionately more by microenvironment factors than is low protein expression, at least in the Nebraska environment. To deal with this problem, the environmental variance of the $F_{2:3}$ population was estimated using only the phenotypic variance of corresponding male parents. In a summary report provided by Brim and Burton (1979), seed protein heritability was reported as very low, ranging from 20 to 39%. However, this observation was attributable to the fact that their heritability estimates were based on progeny arising from the matings of parents that exhibited only modest difference in seed protein content.

Phenotypic Correlations

With respect to the relationship between seed protein and oil content in soybean, these two traits have been found to be negatively correlated, and frequently highly so, based on the literature reports summarized by Burton (1987). In the dissertation study, negative phenotypic correlations between seed protein and seed oil were observed in all six $F_{2:3}$ populations. Table 7 shows correlation values among traits in all six populations evaluated for statistical significant at an $\alpha = 0.05$ criterion. A highly negative correlation was observed between seed protein and seed oil content, ranging from r = -0.83 to -0.68 in each population. In this study, there were significant negative correlations (r = -0.25 to

-0.55) between seed protein and seed moisture contents (Table 7). The well-known negative association for seed protein and oil contents is in agreement with earlier studies in the literature. A strong negative phenotypic correlation between seed protein and seed oil in a mapping population ($r = -0.84$, $P < 0.001$) was also reported recently (Chung et al., 2003), and it was even stronger ($r = -0.98$, $P< 0.001$) in a prior report (Mansur et al., 1996). In mapping studies, the association between these two traits has been attributed to either two tightly linked QTLs for each trait with a repulsion phase allelic relationship, or simply to a single QTL that pleiotropicaly governs the inverse relationship between the two traits (Diers et al., 1992; Mansur et al., 1993b; Chung et al., 2003).

In contrast to the negative protein-oil correlations, highly significant positive correlations ($r = 0.63$ to 0.85) between seed oil content and seed moisture were observed in the six populations. On the other hand, the correlation between (progeny total) seed weight and seed moisture was significantly negative ($r = -0.39$ to -0.64). Note that seed moisture, protein, and oil were all measured simultaneously with the NIR instrument, whereas seed weight was measured separately. A negative correlation between seed oil content and seed weight was detected in each population, but was statistically significant in only three populations, notably 1139, 1143 and 1146. The correlation between seed protein content and seed weight was low, but sometimes negative and sometimes positive ($r = -0.06$ to 0.21), depending upon the population.

In summary, when the number of seeds in an $F_{2:3}$ progeny varied from low to high seed amount, the NIR instrument measurements of seed moisture also varied, but inversely so, from higher to lower values. The reason why smaller seed samples tended to have higher NIR-measured seed moistures is not known. In this regard, however, it is

worth nothing that the phenotypic correlations between seed protein and seed oil content, between seed protein content and moisture, between seed oil and moisture, and between seed weight and moisture in the NIR-generated values for moist and dry seed of the $F_{2:3}$ progenies of population 1143 were not much different in magnitude from the same phenotypic correlations observed in the mid-moisture seed of those same population 1143 progenies.

Genotypic Data Analysis

Construction of the genetic linkage map

In this study, 264 single nucleotide polymorphism (SNP) markers of the 1536 SNP markers were identified as not being parentally polymorphic in any of the six $F_{2:3}$ mapping populations. On the other hand, about 400 to 500 of the remaining 1272 SNP markers were parentally polymorphic in any given population. Interestingly, 400 SNPs were actually parentally polymorphic in all six populations. Based on these markers and their segregation in the $F_{2:3}$ population, MAPMAKER/EXP. VER. 3.0 (Lander and Botstein, 1989; Lincoln et al., 1993) was used to construct a genetic linkage map based on the marker order of the most recent version of the soybean integrated genetic linkage map (Consensus Map 4.0) that was published by Hyten et al. (2010). That version of the genetic map spans 2296.4 cM of Kosambi map distance, when summed over all 20 linkage groups (Fig. 2). Mean linkage group distance is 114.8 cM, and the genetic distance between any consecutive pair of mapped SNP markers averages about 0.6 cM (Hyten et al., 2010). Of the 1272 SNP markers, the total number of parentally polymorphic markers in each $F_{2:3}$ population was 497 (39%), 467 (37%), 425 (33%), 510

(40%), 472 (37%), and 497 (39%) markers for population 1076, 1121, 1122, 1139, 1143, and 1146, respectively. The high percentage of parentally polymorphic markers was expected given the mating type (i.e., landrace x modern cultivar). Though the marker numbers are large, parental polymorphism was sometimes not present in some parts of some linkage groups in each $F_{2:3}$ population, which necessitated the division of the chromosomal linkage maps into two or three sub-chromosomes (e.g., chromosome 1a, 1b, 1c), whenever non-polymorphic marker gap exceeded a 37.2 cM Kosambi map distance. The chromosomal marker linkage distributions are shown in the graphs of Fig. 4.

QTL Detection Based on Selective Genotyping

QTL mapping is conducted by searching for statistically significant associations between the quantitative trait phenotype and the two marker alleles segregating in the population (Zhi-Hong et al., 2005; Wang et al., 2007). A number of statistical approaches can be used to identify association between the trait and particular markers. When the phenotypic trait values are computed for the A and B genotypes of a marker locus, and those values differ substantially, this is usually an indication that a QTL is probably linked to the marker.

Single marker analysis

In this study, a two-sample two-tailed t-test (Bernardo, 2002) was initially used to identify markers whose “A” allele frequency was not identical in the selectively genotyped low and high seed protein fractions of each $F_{2:3}$ population. Table 8 shows the results of this t-test as applied to the SNP markers in each population. Those SNP

markers whose “A” allele frequency differed significantly (using an experiment-wise alpha of 0.0001) between the low and high decile seed protein fractions of $F_{2:3}$ progenies are likely to be linked to a protein QTL segregating in the population whose “A” allele frequency was also different and thus accounted for the large seed protein content difference in those fractions. A total of 43 markers were found to have significant t-test values in these six populations. These markers are located on LG-D1b (Chr 2), C2 (Chr 6), O (Chr 10), E (Chr 15), G (Chr 18), and I (Chr 20). Most of the 43 significant markers were present in linked clusters at a given map position, which would be consistent with a protein QTL being present at or near the cluster marker map positions. Because mean map positions for a QTL usually have plus or minus standard errors of 10 cM, the marker clusters discovered by the t-test were probably accounting for one nearly QTL per cluster. There were essentially eight clusters, with map locations on LG-D1b (two markers), C2 (10 markers), O (seven markers), E (four markers), G (four markers), and I [cluster 1 (seven markers), cluster 2 (six markers), cluster 3 (three markers)]. The highest t-test value in the single marker analysis was identified at the marker on LG-I (Chr 20) of population 1139. The second highest t-test value was identified on LG-O (Chr 10) of population 1076. This is t-test the simplest method one can use to ascertain potential QTLs, but the drawback of this method is that the additive and dominance effect on the seed protein traits cannot be estimated because of confounding of those effects with map distance between the marker and the QTL.

Subsequently, single-marker regression was used in this selective genotyping study. This method is generally used as a preliminary scan for detecting a QTL. A summary of the seed protein QTLs identified with this method is presented in Table 9,

with the genome-wide LOD score scans presented in Fig. 5. Using marker regression, 17 SNP markers were identified as having a LOD score ≥ 3.0 association with seed protein content in six $F_{2:3}$ populations. These 17 SNPs had map positions on LG-D1b (Chr 2), C2 (Chr 6), O (Chr 10), B1 (Chr 11), H (Chr 12), B2 (Chr 14), E (Chr 15), G (Chr 18), and I (Chr 20). Of these 17 markers, only nine (S17861, S30557, S19004, S15265, S30937, S20164, S27739, S27666, and S17070) were judged as having a significant association (determined by using the 95^{th} percentile of genome-wide maximum LOD scores obtained with 1000 permutations) with seed protein content across all six $F_{2:3}$ populations. The highest LOD score obtained in the single marker regression results was a LOD of 8.05 detected for seed protein QTL on LG-O or chromosome 10 (marker S19004 at its map position of 96.44 cM) in population 1076. This QTL also had the highest additive effect of 9.6 g kg^{-1}. The R^2 value (i.e., heritability) for this QTL was 16%. In Fig. 5, the LOD curve for each chromosome had a spiked appearance, because in a single-marker QTL analysis, the LOD score is estimated only at the marker positions (not in between). As is standard procedure when using the method of marker regression, a progeny with a phenotype, but without a given marker genotype, are omitted in that marker's regression analysis.

A summary of seed oil content QTLs detected with marker regression in the six $F_{2:3}$ populations is presented in Appendix Table 2. Ten SNP markers had LOD score values of ≥ 3.0 on LG-D1a (Chr 1), C2 (Chr 6), A2 (Chr 8), O (Chr 10), B1 (Chr 11), B2 (Chr 14), E (Chr 15), G (Chr 18), and I (Chr 20). However, only five markers located on LGs-C2 (Chr 6, marker S17861), A2 (Chr 8, S12625), O (Chr 10, S19004), G (18, S12541), and I (20, S13577) proved to be statistically significant in one or more of the six

populations using the 95th percentile of genome-wide maximum LOD score generated with 1000 permutations.

Interval mapping analysis

A summary of seed protein QTLs detected with the simple interval mapping (via the EM method) is presented in Table 9 (and in Figs. 6 and 7). Note that 14 markers were detected in the six $F_{2:3}$ populations that had a LOD score ≥ 3.0. These QTLs were located on D1b (Chr 2), C2 (Chr 6), O (Chr 10), B1 (Chr 11), H (Chr 12), B2 (Chr 14), E (Chr 15), G (Chr 18), and I (Chr 20). Of these 14 QTLs, eight were statistically significant (based on a LOD sore criterion generated with permutation tests) in one or more of the six $F_{2:3}$ populations (Table 10), with each QTL explaining about 10-19% of the variation. The highest LOD score in the simple interval mapping results was a 7.73 value, detected for a seed protein content QTL on LG-I or chromosome 20 near marker S17070, whose map position was 29.56 cM in population 1139. For this QTL, the additive effect of the high protein parent allele was estimated to be 11.4 g kg^{-1}, with this QTL having a (R^2) heritability of 19%. As previously reported in Soybase (2010) and the literature, the seed protein QTL present in this region of LG-I (Chr 20) (i.e. Satt239, Satt354, Satt439) has been detected many times by soybean researchers (Appendix Table 1). The second highest LOD score in the simple interval mapping results was 6.94, detected for a seed protein content QTL on LG-O (Chr 10) near marker S19004, whose map position was 96.44 cM in population 1076. For this QTL, the additive effect of the high protein parent allele was estimated to be 9.6 g kg^{-1}, with this QTL having a (R^2) heritability of 16%.

As previously mentioned, there are no reports of a seed protein QTL present in this region of LG-O (Chr 10) in Soybase (2010) or in the literature, as this is the newly discovered QTL from this study.

A summary of seed oil content QTLs detected by simple interval mapping is presented in Appendix Table 2. The simple interval mapping method detected ten oil QTLs that had a peak LOD score ≥ 3.0 on D1a (Chr 1), N (Chr 3), A1 (Chr 5), C2 (Chr 6), A2 (Chr 8), O (Chr 10), B2 (Chr 14), E (Chr 15), and I (Chr 20). However, only four markers on LG-A2 (marker S12625), LG-O (S19004), LG-E (S20164), and LG-I (S17070) proved to be statistically significant based on permutation determined genome-wide LOD scores and these QTLs explained 10-21% of the variation.

In the six $F_{2:3}$ populations, seed protein and oil content were negatively correlated (Table 7). This negative phenotypic correlation between the two characters was also reflected in the QTL detection results. For instance, in population 1076, the phenotypic data of seed protein and seed oil content were negatively correlated (r = -0.76). In the case of the QTL near marker S19004 on LG-O (Chr 10), the S19004 SNP marker allele from the high protein low oil PI female parent 1076 (PI 437112A) was associated with both high seed protein and low seed oil, whereas the allele from the low protein high oil agronomic male parent (PI 597386) was associated with low seed protein and high seed oil content (Table 9). Inversely, relative to the QTL near the SNP marker S12725 on LG-C2 (Chr 6), the low protein high oil agronomic male parent (PI 597386) marker allele was associated with low oil and high seed protein, whereas the high protein low oil female parent 1076 (PI 437112A) marker allele coded for high oil and low seed protein. In each of these two cases, it is not known if there are two tightly linked QTLs tightly

linked to the SNP marker, one controlling only seed protein and the other controlling seed oil, or if there is one pleiotropic QTL that governs both protein and oil, but in an opposite direction. What is interesting is that high protein parent is homozygous for a LG-O (Chr 10) SNP marker allele that confers high protein and low oil, but is also homozygous for a LG-C2 (Chr 6) SNP marker allele that confers low protein and high oil. This kind of situation is likely the genic basis for transgressive segregation in the F_2 population of this parental mating.

In population 1076, a fairly strong additive effect (9.6 g kg^{-1}) of the PI 437112A allele was estimated for the LG-O (Chr 10) marker S19004, with the R^2 value of 16%. This marker was also associated with the same seed protein QTL, using single marker regression analysis. Additionally, in population 1121 and 1122, moderate additive effects (7.9 and 6.5 g kg^{-1}, respectively) were found for LG-O (Chr 10) marker S15265, with R^2 values of 12 and 13%, respectively. A review of SoyBase (2010) and the literature indicted that no seed protein QTLs of any notable additive effect have been reported to date in this LG-O (Chr 10) region (Appendix Table 1), at least to the extent of having additive effects on seed protein that approached the magnitude of the additive effect of the LG-I protein QTL. Hence, this is a new seed protein QTL discovered on LG-O (Chr 10) at near the two adjacent markers S19004 and S15265. The R^2 value is the proportion of the phenotypic variance explained by the QTL. Although selective bias (i.e., Beavis effect) can upwardly bias the calculated additive effect, the bias is smaller the lager the true additive effect, which may mean the estimated additive effect of the LG-O (Chr 10) QTL may be close to its true effect (Broman and Sen, 2009). Almost all significant QTLs in this dissertation study exhibited additive effects that were magnitudinally greater than

the dominant effects. This is not unexpected given that most self-pollinated crops typically have minimal dominance effects. A large additive effect is conducive to breeding methods that exploit additive genetic variance, such as the pedigree method, bulk method, and backcross method.

The negative correlation between seed protein and seed oil content that has seen observed with inbred line genotypic means, has no repeatedly been observed in molecular marker genotype means, which implies that those markers are linked to QTLs governing seed protein and/or oil content. For example, QTL alleles coding for high protein content were invariably associated with low seed oil content and *vice versa* in several studies (Lark et al., 1994; Lee et al., 1996; Sebolt et al., 2000; Chung et al., 2003). The only report of a positive correlation at the QTL level was a LG-H (Chr 12) QTL reported by Qui et al. (1999) but those results have been questioned (see Appendix Table 1). In this dissertation study, strong negative correlations were observed in all six populations. Because of the strong correlation at the molecular marker level, one could speculate as to why genes involved with seed protein or with seed oil levels seem to be clustered more often than not, which would, intrinsically suggest pleiotropy is more common than two-QTL linkage. A close linkage between a seed protein QTL and a seed oil QTL (or a pleiotropic QTL) was suggested in this study by the joint mapping of protein and oil QTLs to LG-O (Chr 10), E (Chr 15), and I (Chr 20). A much higher map density and a substantively larger plant population (to provide opportunity for recombination to destroy the pleiotropy hypothesis by the revering the repulsed linkage phase to a coupled one) would be necessary to resolve whether protein and oil are inversely controlled by the same pleiotropic gene or by different genes linked in repulsion phase. Identifying QTL

for one trait but not the other would certainly be useful to breeders desiring to increase the level of one trait while holding the other constant or to increase the levels of both traits simultaneously.

In total, 17 and 14 QTLs were detected by the single marker regression method and by the simple interval mapping (via EM) method, respectively, and these QTLs mapped to similar positions of the chromosomes. The QTLs detected by the marker regression and the interval mapping methods also showed similar magnitudes of estimated QTL additive effects. The sign of the additive effect indicated the direction (increase or decrease) of the parent contributing the allele of the marker that serves as a proxy for the allele of the nearby QTL. In all of the QTL (but two) that were detected by both methods, the direction of the additive effect was consistent with the high to low protein difference between the parents. The exceptions were the significant QTLs on LG-C2 (Chr 6, population 1076 and 1121), and on LG-B2 (Chr 14, population 1146), for which the high protein allele came from the low protein content. Based on this study, it can be concluded that the methods of interval mapping and marker regression produced very similar results in this dissertation study even though the two methods employ different techniques.

Comparison QTLs Detected with QTLs Previously Reported

Three of the seven statistically significant seed protein QTLs identified via the interval mapping method [one on LG-C2 (Chr 6), one on LG-E (Chr 15), and one on LG-I (Chr 20)] were localized to regions where other researchers have previously reported similar QTLs. For example, on LG-C2 (Chr 6), SNP marker S17861 (which has a USLP

1.0 map position of 97.81 cM) is linked to a seed protein QTL. Kabelka et al. (2004) found SSR marker Satt363 to be associated with seed protein content QTL. The Satt363 marker has a USLP 1.0 map position of 89.70 cM and thus is only 8.11 cM distant from S17861 marker.

On LG-E (Chr 15), SNP marker S29437, located at 17.01 cM on the USLP 1.0 map, was strongly associated with seed protein content. Tajuddin et al. (2003) found SSR marker Satt384 to be associated with seed protein content. Satt384 is located at 19.61 cM on the USLP 1.0 map position (Hyten et al., 2010), and thus is only 2.60 cM distant from S29437.

On LG-I (Chr 20), SNP marker S17070, located at 30.00 cM on the USLP 1.0 map was very strongly associated with both seed protein and seed oil content. Chung et al. (2003) found SSR marker Satt239 to be associated with seed protein and seed oil content. Satt239 is located at 29.61 cM on the USLP 1.0 map position (Hyten et al., 2010), and thus is only 0.39 cM distant from marker S17070.

In practice, the detected QTLs provide information that is useful for selecting PI parents with desired genotypes for producing progeny in which a breeder can perhaps stack all of the high protein causing alleles at these QTLs. This would allow the breeder to create a very high protein (or inversely very low protein) breeding line that could be used as a single donor parent when the breeder wants to deploy one or more of the high protein alleles into recipient high yielding cultivars.

Power of QTL Mapping Based on Selective Genotyping

Selecting genotyping can be used instead of entire population genotyping without loss of QTL detection power if the entire population and the tail population sizes are large enough and a high density of markers is used (Gallais et al., 2007; Navabi et al., 2009; Sun et al., 2010).

In the previous study by Richie (2003), the author reported that among her set of 41 populations, five (1076, 1121, 1122, 1139, and 1146) did not seem to segregate for QTLs near SSR markers known to be near previously reported QTLs on LG-I (Chr 20), E (Chr 15), H (Chr 12; top), and H (Chr 12; bottom). Moreover, Ritchie (2003) noted that one additional population (1143) segregated only for a seed protein QTL on LG-E (Chr 15). All of her other populations segregated for the well-known LG-I (Chr 20) QTL. The present dissertation study confirmed Ritchie's discovery of a significant QTL in population 1143 located on LG-E (Chr 15). Ritchie (2003) did not examine markers at other chromosomal locations in her study. As noted above, Richie (2003) used only SSR markers located in very specific regions of four known chromosomes; whereas in this dissertation study, about 500 SNP markers were used to evaluate all 20 soybean chromosomes for seed protein QTLs. In the present study, the population size was of about 220 $F_{2:3}$ progenies, whereas in the Ritchie (2003) study about 120 or less $F_{2:3}$ progenies were evaluated, so the power of detecting QTL was certainly higher in this dissertation research. Ooijen (1992), Darvasi et al. (1993), and Kearsey and Farquhar (1998) have observed that confidence limits, power and reliability of QTL studies can be improved by increasing family size and number of families. Kearsey and Pooni (1996) also stated that the precision of a QTL position depends more on the population size than

the number of markers, and that no notable increase in QTL position accuracy is obtained with more than five uniformly spaced markers on each chromosome. Therefore, it is important to use a mapping population of relatively large size to ascertain QTLs of large effect and reliably estimate the QTL effects.

One of the main interests of selective genotyping is that it allows breeders to detect markers associated with QTLs by using only the selected individuals in extreme two tail fractions. This may require more phenotyping prior to a lesser amount of genotyping, but the savings in marker genotyping expense may be well worthwhile.

CONCLUSIONS

By studying six $F_{2:3}$ mapping populations for which there was evidence that the LG-I (Chr 20) QTL was not likely to segregate, some new QTLs affecting both seed protein and seed oil contents were identified. In this study, fewer than ten QTLs were detected for both seed protein and oil content. Statistically significant seed protein QTLs were identified on LGs-C2 (Chr 6), O (Chr 10), B2 (Chr 14), E (Chr 15), and I (Chr 20). The highest LOD score (7.73) detected for a seed protein QTL was the well-known QTL located on LG-I (Chr 20) near SNP marker S17070 in population 1139, and it had an additive effect of 11.4 g kg^{-1}. This QTL on LG-I (Chr 20) region has been reported many times by other researchers. However, the seed protein QTL discovered on LG-O (Chr 10) near marker S19004 in a population 1076, and near marker S15265 in populations of 1121 and 1122, has not been reported before. Hence, this is a new seed protein QTL that resides between the two adjacent SNP markers S19004 and S15265. At marker S19004, the additive effect was 9.6 g kg^{-1}. At the marker S15265, the additive effect was 7.9 and 6.5 g kg^{-1} for population 1121 and 1122, respectively.

Seed oil QTLs were also discovered on LG-O (Chr 10) near marker S19004, and on LG-I (Chr 20) near marker S17070. This suggest that new LG-O (Chr 10) QTL, like the LG-I (Chr 20) QTL, is either a single pleiotropic QTL or repulsion-phase linked protein and oil QTLs.

In conclusion, the results of this dissertation research indicated that the new QTL on LG-O (Chr 10) may be a seed protein QTL worthy of use by breeders interested in developing high protein breeding lines. If so, the germplasm accessions with the high

protein allele for this QTL are PI 437112A (1076), PI 398672 (1121), and PI 360843 (1122).

Finally, in this study larger population sizes were used, resulting in higher power than that of Richie (2003) in her study. The higher power allowed for detection of the well-known seed protein QTL on LG-I (Chr 20) that was not detected by Ritchie (2003). In addition, seed protein QTL on LG-E (Chr 15) was detected in this study, which confirmed the result that Richie (2003) reported in population 1143. A new seed protein QTL was discovered on LG-O (Chr 10), which inversely impacts seed oil content.

Table 1. Parental germplasm descriptions.

Maturity group	Parental code†	Seed‡		Germplasm accession			Flower color	Hilum color	Pod color	Pubesence color	Pubesence form	Seed coat color	Seed coat luster
		Protein	Oil	Number	Name (if any)	Origin							
		------- g kg^{-1} --------											
II	1076	482	154	PI437112A	VIR 249	USSR (FarE)	W§	Y	Tn	G	E	Y	S
II	1106M	382	195	PI 597386	Dwight	IL, USA	P	Bl	Tn	T	E	Y	D
III	1121	494	177	PI 398672	KAERI-GNT 301-1	S.Korea	Dp	Rbr	Br	T	E	Rbr	S
III	1122	484	184	PI 360843	Oshimashirome	Japan	W	Y	Br	G	E	Y	I
III	1137M	411	194	PI 597387	Pana	IL, USA	P	Bf	Br	G	E	Y	D
IV	1139	507	151	PI 407788A	ORD 8113	S.Korea	P	Bf	Tn	G	E	Y	S
IV	1143	488	158	PI 398704	KAERI-GNT 330-9-1	S.Korea	P	Bf	Br	G	E	Y	I
IV	1146	493	159	PI 407823	-	S.Korea	P	Bf	Tn	G	E	Y	I
IV	1181M	424	180	PI 606748	Rend	S.Korea	W	Bf	Br	G	E	Y	D

† The suffix M identifies the cultivars used as male parents in these mating: 1076 x 1106M; (1121,1122) x 1137M; (1139, 1143, 1146) x 1181M.

‡ Seed protein and oil values are those published in the soybean germplasm database (NGRP, 2009).

§ Abbreviations: White, Purple, Dark purple, Yellow, Black, Reddish brown, Buff, Tan, Brown, Gray, Tawny, Erect on leaf surface, Shiny, Intermediate, and Dull.

Table 2. Phenotypes of the classical marker genes used to comfirm the authentic F_1 plants.

Mating		F_1			
Female	Male	Flower color	Pub color	Pod color	hilum color
1076	1106M	P*	T	-	G
1121	1137M	-	-	-	-
1122	1137M	P	-	-	-
1139	1181M	-	-	Br	-
1143	1181M	-	-	-	-
1146	1181M	-	-	-	-

* Abbreviations: Purple, Tawny, Brown, Grey.

Table 3. List of the F_1 plants, the F_2 seed and plant numbers, and the total number of $F_{2:3}$ progeny obtained in each mating. All $F_{2:3}$ progeny were evaluated for seed protein and oil content, but only those progenies occupying the upper and lower quintiles were subjected to a second replicate evaluation. Because only the 22 lowest and 22 highest seed protein $F_{2:3}$ progeny were SNP-genotyped, the percentage of the population genotyped is shown here.

Mating		Confirmed	F_2	Harvested	Evaluated $F_{2:3}$	$F_{2:3}$ progeny quintiles		Progeny genotyped
Female	Male	F_1 plants	seeds	F_2 plants	progeny	low	high	
		----------No. ----------						---------%---------
1076	1106M	11	703	217	188	41	40	23.4
1121	1137M	14	670	262	217	45	45	20.3
1122	1137M	18	1544	213	189	41	40	23.3
1139	1181M	17	1399	283	167	36	38	26.3
1143	1181M	5	415	279	195	42	40	22.6
1146	1181M	5	2190	254	171	36	38	25.7

Table 4. Seed protein means and other statistical parameters relative to the populations of $F_{2:3}$ progenies derived from each parental mating.

Female & Pop.	Male	Female Parent			Male Parent			$F_{2:3}$ Progeny						
no.	no.	Mean	Std. Dev.	Max-min	Mean	Std. Dev.	Max-Min	Mean	Std. Dev.	Max-min	Shapiro-Wilk	Pr	Kurtosis	Skewness
		---------- g kg^{-1} ----------												
1076	1106M	447	23	474-419	404	11	427-384	431	18	483-371	0.99	0.61	0.32	-0.23
1121	1137M	432	7	442-420	401	13	420-376	419	18	472-377	0.97	0.00	-0.26	0.56
1122	1137M	423	13	441-404	393	12	408-368	419	14	459-373	1.00	0.84	0.07	-0.02
1139	1181M	447	25	473-412	411	10	426-395	434	18	476-391	0.99	0.39	-0.62	0.06
1143	1181M	437	16	457-409	411	12	428-391	431	18	471-378	0.99	0.14	-0.43	0.12
1143 (moist seed)[†]	1181M	425	13	446-404	402	9	413-387	419	15	460-374	0.99	0.38	-0.07	0.21
1143 (dry seed)[‡]	1181M	440	15	465-413	415	11	430-397	433	19	488-380	1.00	0.83	-0.09	0.19
1146	1181M	438	15	453-418	413	8	424-400	424	14	467-397	0.99	0.10	-0.40	0.14

[†]= same progenies of population 1143 but NIR-measured when the seed moister than original seed.

[‡] = same progenies of population 1143 but NIR-measured when the seed dryer than original seed.

Table 5. Seed oil means and other statistical parameters relative to the populations of $F_{2:3}$ progenies derived from each parental mating.

Female & Pop.	Male	Female Parent			Male Parent			$F_{2:3}$ Progeny						
no.	no.	Mean	Std. Dev.	Max-min	Mean	Std. Dev.	Max-Min	Mean	Std. Dev.	Max-min	Shapiro-Wilk	Pr	Kurtosis	Skewness
		---------- g kg^{-1} ----------												
1076	1106M	165	38	208-125	188	12	214-172	165	21.0	219-121	0.97	0.001	-0.001	0.51
1121	1137M	179	11	200-165	196	13	224-180	187	18.0	229-137	0.99	0.03	-0.08	-0.37
1122	1137M	190	19	220-167	195	15	221-176	185	15.0	224-148	0.99	0.33	0.10	0.20
1139	1181M	159	39	201-112	186	11	203-168	170	23.0	229-120	0.99	0.17	-0.46	0.22
1143	1181M	172	23	214-147	187	17	221-166	170	23.0	229-103	0.99	0.05	0.14	0.27
1143 (moist seed)†	1181M	170	19	206-139	186	11	207-166	169	21.0	219-108	0.99	0.15	0.21	0.16
1143 (dry seed)‡	1181M	169	17	208-144	181	8	190-162	167	22.0	227-99	0.99	0.15	0.10	0.23
1146	1181M	164	26	201-140	180	10	193-166	177	20.0	231-135	0.96	0.00	-0.09	0.59

† = same progenies of population 1143 but NIR-measured when the seed moister than original seed.

‡ = same progenies of population 1143 but NIR-measured when the seed dryer than original seed.

Table 6. Seed moisture means and other statistical parameters relative to the populations of $F_{2:3}$ progenies derived from each parental mating.

Female & Pop.	Male	Female Parent			Male Parent			$F_{2:3}$ Progeny						
no.	no.	Mean	Std. Dev.	Max-min	Mean	Std. Dev.	Max-Min	Mean	Std. Dev.	Max-min	Shapiro-Wilk	Pr	Kurtosis	Skewness
		---------- g kg^{-1} ----------												
1076	1106M	9.1	2.6	11.9-6.1	7.8	0.9	10.7-6.9	8.1	1.4	12.6-5.8	0.87	0.00	1.33	1.34
1121	1137M	9.4	1.0	11.6-8.4	7.8	1.2	10.3-6.5	8.9	1.3	12.9-6.8	0.92	0.00	0.22	0.93
1122	1137M	9.7	1.8	12.6-7.6	7.5	1.4	11.3-6.6	8.2	1.2	12.6-6.4	0.84	0.00	1.92	1.68
1139	1181M	9.6	1.9	10.9-6.4	7.8	0.9	10.0-6.7	8.4	1.6	13.1-6.2	0.88	0.00	0.09	1.03
1143	1181M	8.9	1.6	11.6-6.9	8.7	1.6	12.6-7.3	8.5	1.6	12.8-6.5	0.87	0.00	-0.18	0.96
1143 (moist seed)[†]	1181M	10.7	0.6	12.0-9.6	10.4	0.4	11.0-9.8	10.6	0.8	12.6-9.0	0.93	0.00	-0.56	0.66
1143 (dry seed)[‡]	1181M	7.5	1.5	11.2-4.8	6.9	0.8	8.0-5.7	7.0	2.0	12.6-4.0	0.89	0.00	-0.20	1.04
1146	1181M	8.6	1.7	11.0-7.2	7.7	0.4	8.4-7.2	8.8	1.6	12.6-6.7	0.90	0.00	-0.47	0.80

[†] = same progenies of population 1143 but NIR-measured when the seed moister than original seed.

[‡] = same progenies of population 1143 but NIR-measured when the seed dryer than original seed.

Table 7. Pearson phenotypic correlation coefficients among seed protein, seed oil, seed weight, and seed moisture of the $F_{2:3}$ progenies in each population.

Pop.	Seed protein vs. seed oil	Seed protein vs. seed moisture	Seed protein vs. seed weight	Seed oil vs. seed moisture	Seed oil vs. seed weight	Seed weight vs. seed moisture
1076	-0.76**	-0.31**	-0.12	0.69**	-0.11	-0.43**
1121	-0.73**	-0.25**	-0.16*	0.63**	-0.11	-0.51**
1122	-0.68**	-0.28**	-0.06	0.66**	-0.13	-0.39**
1139	-0.83**	-0.39**	0.06	0.70**	-0.24*	-0.57**
1143	-0.82**	-0.54**	0.21*	0.82**	-0.40**	-0.64**
1146	-0.74**	-0.55**	0.20*	0.85**	-0.38**	-0.62**

*,** significant at alpha = 0.05 and 0.01, respectively.

Table 8. SNP markers significantly associated with seed protein QTL based on a t-test of the frequency of the agronomic parent cultivar (A) allele in the low decile versus high decile fractions of $F_{2:3}$ progenies. The comparison-wise alpha of 0.05 was adjusted to a experment-wise alpha of 0.0001 by dividing the comparison-wise alpha by number of comparisons (i.e., 500 SNP markers).

Pop. No.	Hi-Pro	Low-Pro	Chr.	LG	USLP 1.0 Pos.	SNP	Total	Low-pro decile		Hi-pro decile		T-test
	female parent	male parent			------cM-------	marker	$F_{2:3}$ no.	$F_{2:3}$ no.	A frequency	$F_{2:3}$ no.	A frequency	value[†]
1146	PI 407823	PI 606748	2	D1b	30.39	S12694	171	20	0.30	22	0.75	-4.12
1146	PI 407823	PI 606748	2	D1b	30.67	S12843	171	22	0.32	22	0.77	-4.26
1076	PI 437112A	PI 597386	6	C2	103.11	S24435	188	22	0.30	21	0.74	-4.10
1076	PI 437112A	PI 597386	6	C2	103.33	S17254	188	22	0.30	21	0.74	-4.10
1076	PI 437112A	PI 597386	6	C2	103.87	S12561	188	22	0.30	21	0.74	-4.10
1076	PI 437112A	PI 597386	6	C2	104.50	S23346	188	22	0.30	21	0.74	-4.10
1076	PI 437112A	PI 597386	6	C2	104.50	S30557	188	22	0.30	22	0.75	-4.26
1076	PI 437112A	PI 597386	6	C2	104.51	S12725	188	21	0.29	22	0.75	-4.30
1076	PI 437112A	PI 597386	6	C2	105.16	S25487	188	21	0.29	21	0.74	-4.15
1076	PI 437112A	PI 597386	6	C2	106.47	S17110	188	22	0.30	21	0.74	-4.10
1076	PI 437112A	PI 597386	6	C2	107.01	S17519	188	20	0.30	21	0.75	-4.18
1076	PI 437112A	PI 597386	6	C2	114.13	S13752	188	22	0.27	21	0.71	-4.09

[†] = a negative value indicates that the low protein parent marker allele "A" is likely linked to a low protein parent QTL allele "A" that conditions high seed protein. A positive value indicates the inverse.

Table 8. (Cont.)

Pop. No.	Hi-Pro	Low-Pro	Chr.	LG	USLP 1.0 Pos.	SNP	Total	Low-pro decile		Hi-pro decile		T-test
	female parent	male parent			------cM-------	marker	$F_{2:3}$ no.	$F_{2:3}$ no.	A frequency	$F_{2:3}$ no.	A frequency	value†
1076	PI 437112A	PI 597386	10	O	94.97	S12744	188	21	0.71	21	0.14	5.24
1121	PI 398672	PI 597387	10	O	94.97	S12744	217	19	0.68	19	0.21	4.13
1076	PI 437112A	PI 597386	10	O	96.44	S19004	188	22	0.71	22	0.14	5.33
1121	PI 398672	PI 597387	10	O	96.44	S19004	217	21	0.69	21	0.24	4.15
1076	PI 437112A	PI 597386	10	O	99.69	S15265	188	22	0.73	22	0.16	5.33
1121	PI 398672	PI 597387	10	O	99.69	S15265	217	21	0.67	22	0.21	4.28
1122	PI 360843	PI 597387	10	O	99.69	S15265	189	22	0.66	21	0.19	4.35
1143	PI 398704	PI 606748	15	E	17.01	S29437	195	21	0.64	21	0.19	4.15
1143	PI 398704	PI 606748	15	E	17.01	S30937	195	22	0.66	22	0.18	4.48
1143	PI 398704	PI 606748	15	E	19.80	S20164	195	22	0.68	20	0.23	4.18
1143	PI 398704	PI 606748	15	E	28.33	S18494	195	22	0.66	21	0.19	4.35
1121	PI 398672	PI 597387	18	G	0.11	S29989	217	21	0.19	22	0.66	-4.35
1121	PI 398672	PI 597387	18	G	0.92	S18434	217	22	0.21	22	0.66	-4.26
1121	PI 398672	PI 597387	18	G	1.64	S27739	217	19	0.16	21	0.67	-4.55
1121	PI 398672	PI 597387	18	G	9.13	S12541	217	21	0.17	21	0.67	-4.58

† = a negative value indicates that the low protein parent marker allele "A" is likely linked to a low protein parent QTL allele "A" that conditions high seed protein. A positive value indicates the inverse.

Table 8. (Cont.)

Pop. No.	Hi-Pro	Low-Pro	Chr.	LG	USLP 1.0 Pos.	SNP	Total	Low-pro decile		Hi-pro decile		T-test
	female parent	male parent			------cM-------	marker	$F_{2:3}$ no.	$F_{2:3}$ no.	A frequency	$F_{2:3}$ no.	A frequency	value[†]
1139	PI 407788A	PI 606748	20	I	11.35	S25004	167	19	0.71	22	0.23	4.36
1139	PI 407788A	PI 606748	20	I	17.03	S25744	167	21	0.69	20	0.18	4.67
1139	PI 407788A	PI 606748	20	I	17.03	S25746	167	20	0.73	20	0.18	4.92
1139	PI 407788A	PI 606748	20	I	17.14	S25758	167	22	0.71	22	0.25	4.26
1139	PI 407788A	PI 606748	20	I	17.14	S25775	167	20	0.70	21	0.19	4.61
1139	PI 407788A	PI 606748	20	I	17.69	S25739	167	21	0.71	22	0.21	4.73
1139	PI 407788A	PI 606748	20	I	18.51	S13338	167	18	0.72	22	0.21	4.61
1139	PI 407788A	PI 606748	20	I	29.56	S13608	167	22	0.75	21	0.17	5.41
1139	PI 407788A	PI 606748	20	I	29.56	S27666	167	22	0.77	22	0.18	5.54
1139	PI 407788A	PI 606748	20	I	30.00	S17070	167	18	0.72	21	0.17	4.89
1139	PI 407788A	PI 606748	20	I	33.21	S13844	167	21	0.76	22	0.18	5.38
1139	PI 407788A	PI 606748	20	I	38.61	S16725	167	22	0.71	22	0.21	4.69
1139	PI 407788A	PI 606748	20	I	38.63	S13489	167	21	0.74	22	0.21	4.95
1139	PI 407788A	PI 606748	20	I	41.90	S15256	167	22	0.68	22	0.18	4.69
1139	PI 407788A	PI 606748	20	I	43.99	S13577	167	22	0.73	22	0.18	5.12
1139	PI 407788A	PI 606748	20	I	50.85	S13604	167	21	0.69	22	0.23	4.29

[†] = a negative value indicates that the low protein parent marker allele "A" is likely linked to a low protein parent QTL allele "A" that conditions high seed protein. A positive value indicates the inverse.

Table 9. Summary of seed protein QTL peak LOD scores > 3.00, ordered by chromosome, then positions, that were identified by marker regression (MR) and standard interval mapping using EM algorithm. A stratified permutation test of 1000 replicates was conducted in each popualtion to provide a genome-wide 95th percentile LOD score to serve as a statistical significance criterion for evaluating any given QTL LOD score peak. The additive (a) and dominance (d) effects were calculated on the basis of the substitution of a female high protein parent allele for a male low protein parent allele at the indicated marker locus.

Pop.	Hi-Pro	Lo-Pro	Marker or	Chr.	LG	USLP 1.0	QTL peak LOD (if >3.0)		Permutation-based LOD score		QTL Effect		R^2
No.	Female Parent	Male Parent	Nearest Marker	No.	Name	Pos.	MR	EM	MR	EM	a‡	d‡	EM
						---cM---					-----g kg^{-1}-----		--%--
1146	PI 407823	PI 606748	S12843	2	D1b	30.67	3.17	3.40	-	-	-6.5	0.7	9
1121	PI 398672	PI 597387	S17861	6	C2	97.81	5.05	5.69	4.31	4.85	-7.7	-4.4	11
1076	PI 437112A	PI 597386	S30557	6	C2	104.50	4.89	-	4.15	-	-	-	11
1076	PI 437112A	PI 597386	S12725†	6	C2	104.51	-	4.9	-	3.77	-9.3	-1.1	11
1122	PI 360843	PI 597387	S15512	10a	O	66.05	3.03	-	-	-	-	-	7
1122	PI 360843	PI 597387	S12648†	10a	O	66.85	-	3.41	-	-	2.3	-3.5	8
1076	PI 437112A	PI 597386	S19004	10	O	96.44	8.05	6.94	4.15	3.77	9.6	-3.3	16
1121	PI 398672	PI 597387	S15265	10	O	99.69	5.83	5.96	4.31	4.85	7.9	-2.3	12
1122	PI 360843	PI 597387	S15265	10b	O	99.69	5.99	5.70	4.19	3.76	6.5	-1.2	13
1143	PI 398704	PI 606748	S30838†	10b	O	117.86	-	3.40	-	-	6.1	-3.2	8
1143 (moist)	PI 398704	PI 606748	S30838†	10b	O	117.86	-	3.52	-	-	5.1	-3.4	8
1143 (dry)	PI 398704	PI 606748	S30838†	10b	O	117.86	-	3.51	-	-	7.0	-2.9	8
1139	PI 407788A	PI 606748	S17851	11	B1	110.63	3.38	3.45	-	-	5.0	-5.5	9
1121	PI 398672	PI 597387	S24429	11b	B1	115.55	3.52	-	-	-	-	-	7
1146	PI 407823	PI 606748	S12786	12	H	56.14	-	3.18	-	-	-5.8	-0.8	8
1146	PI 407823	PI 606748	S28782	12	H	62.14	3.04	-	-	-	-	-	8

† nearest marker.

‡ If the effect is negative, the high protein parent marker allele depresses seed protein content.

Table 9. (Cont.)

Pop.	Hi-Pro	Lo-Pro	Marker or	Chr.	LG	USLP 1.0	QTL peak LOD (if >3.0)		Permutation-based LOD score		QTL Effect		R^2
No.	Female Parent	Male Parent	Nearest Marker	No.	Name	Pos.	MR	EM	MR	EM	a‡	d‡	EM
						---cM----					-----g kg^{-1}------		--%--
1146	PI 407823	PI 606748	S13540	14	B2	11.89	-	-	-	-	-	-	9
1146	PI 407823	PI 606748	S30533	14	B2	13.10	3.87	3.94	-	3.92	-6.4	-0.4	10
1143	PI 398704	PI 606748	S29437†	15	E	17.01	-	5.04	-	3.94	9.7	1.0	11
1143 (moist)	PI 398704	PI 606748	S29437†	15	E	17.01	-	5.91	-	4.05	7.7	0.6	13
1143	PI 398704	PI 606748	S30937	15	E	17.01	4.43	-	4.37	-	-	-	10
1143 (moist)	PI 398704	PI 606748	S30937	15	E	17.01	5.01	-	4.29	-	-	-	11
1143 (dry)	PI 398704	PI 606748	S20164	15	E	19.80	4.53	4.99	4.18	3.97	9.3	0.5	11
1143	PI 398704	PI 606748	S18494	15	E	28.33	-	-	-	-	-	-	9
1121	PI 398672	PI 597387	S23378	18a	G	0.00	-	3.78	-	-	-6.9	1.1	8
1121	PI 398672	PI 597387	S27739	18a	G	1.64	4.61	-	4.31	-	-	-	9
1143	PI 398704	PI 606748	S13933	19	L	55.88	3.18	-	-	-	-	-	7
1143 (dry)	PI 398704	PI 606748	S13479	20a	I	18.91	3.80	-	-	-	-	-	9
1139	PI 407788A	PI 606748	S27666	20	I	29.56	7.86	-	4.40	-	-	-	19
1139	PI 407788A	PI 606748	S17070†	20	I	30.00	-	7.73	-	4.04	11.4	0.1	19
1143	PI 398704	PI 606748	S17070	20a	I	30.00	3.90	3.16	-	-	7.4	2.2	7
1143 (moist)	PI 398704	PI 606748	S17070	20a	I	30.00	5.03	3.74	4.29	-	6.2	1.6	11

† nearest marker.

‡ If the effect is negative, the high protein parent marker allele depresses seed protein content.

Table 10. Relative to the data presented in Table 9, shown here are the nearest markers, positions and LOD scores for the 95% Bayes confidence interval (C.I.) calculated for each statistically significant seed protein QTL detected by standard interval mapping using EM algorithm (EM).

			Left boundary of the C.I.			Marker	USLP 1.0		Right boundary for the C.I.		
Pop. No.	Chr. No.	LG Name	nearest marker	USLP 1.0 map pos.	LOD	or Nearest Marker	map pos.	LOD	nearest marker	USLP 1.0 map pos.	LOD
				cM			cM			cM	
1121	6	C2	S21568	80.20	5.01	S17861	97.81	5.69‡	S17861	97.81	4.61
1076	6	C2	S30961	100.91	3.66	S12725†	104.51	4.86‡	S13752	114.13	3.53
1076	10	O	S12744	94.97	6.06	S19004	96.44	6.94‡	S15265	99.69	5.92
1121	10	O	S12744	94.97	5.12	S15265	99.69	5.96‡	S15265	99.69	5.44
1122	10b	O	S15265	99.69	5.28	S15265	99.69	5.70‡	S13214	114.60	4.96
1146	14	B2	S13540	11.89	2.99	S30533	13.10	3.94‡	S30533	13.10	2.43
1143 (moist)	15	E	S13233	11.99	5.10	S29437†	17.01	5.91‡	S18494	28.33	5.05
1143	15	E	S13233	11.99	4.22	S29437†	17.01	5.04‡	S18494	28.33	4.13
1143 (dry)	15	E	S13233	11.99	4.20	S20164	19.80	4.99‡	S18494	28.33	4.08
1139	20	I	S13608	29.56	6.50	S17070†	30.00	7.73‡	S13577	43.99	6.60

† nearest marker.

‡ statistically significant (determined by using the 95th percentile of genome-wide maximum LOD scores of 1000 permutations).

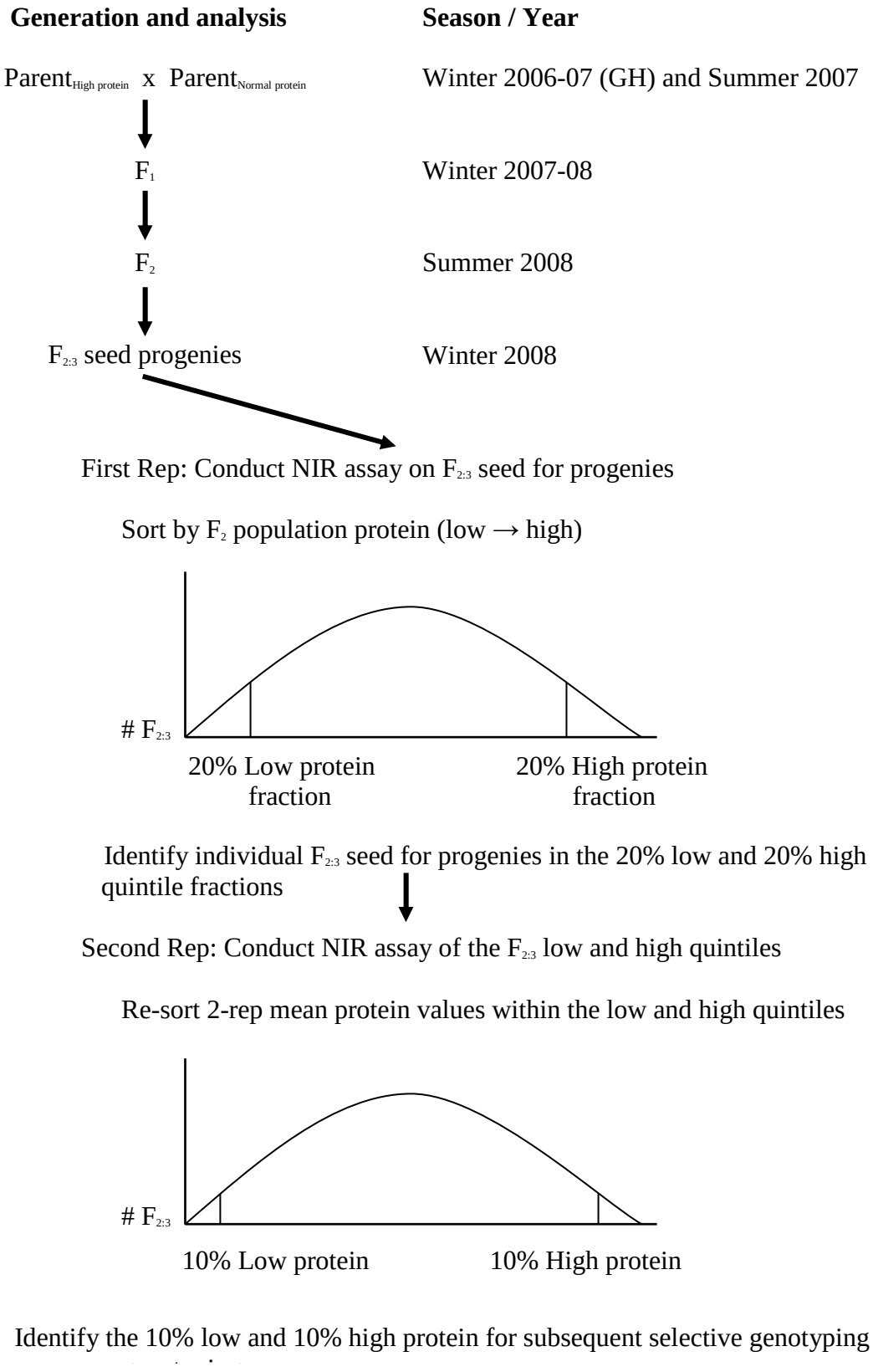


Fig. 1. Development of F_2 populations and the use of phenotyped $F_{2:3}$ seed progenies for selective genotyping with SNP markers

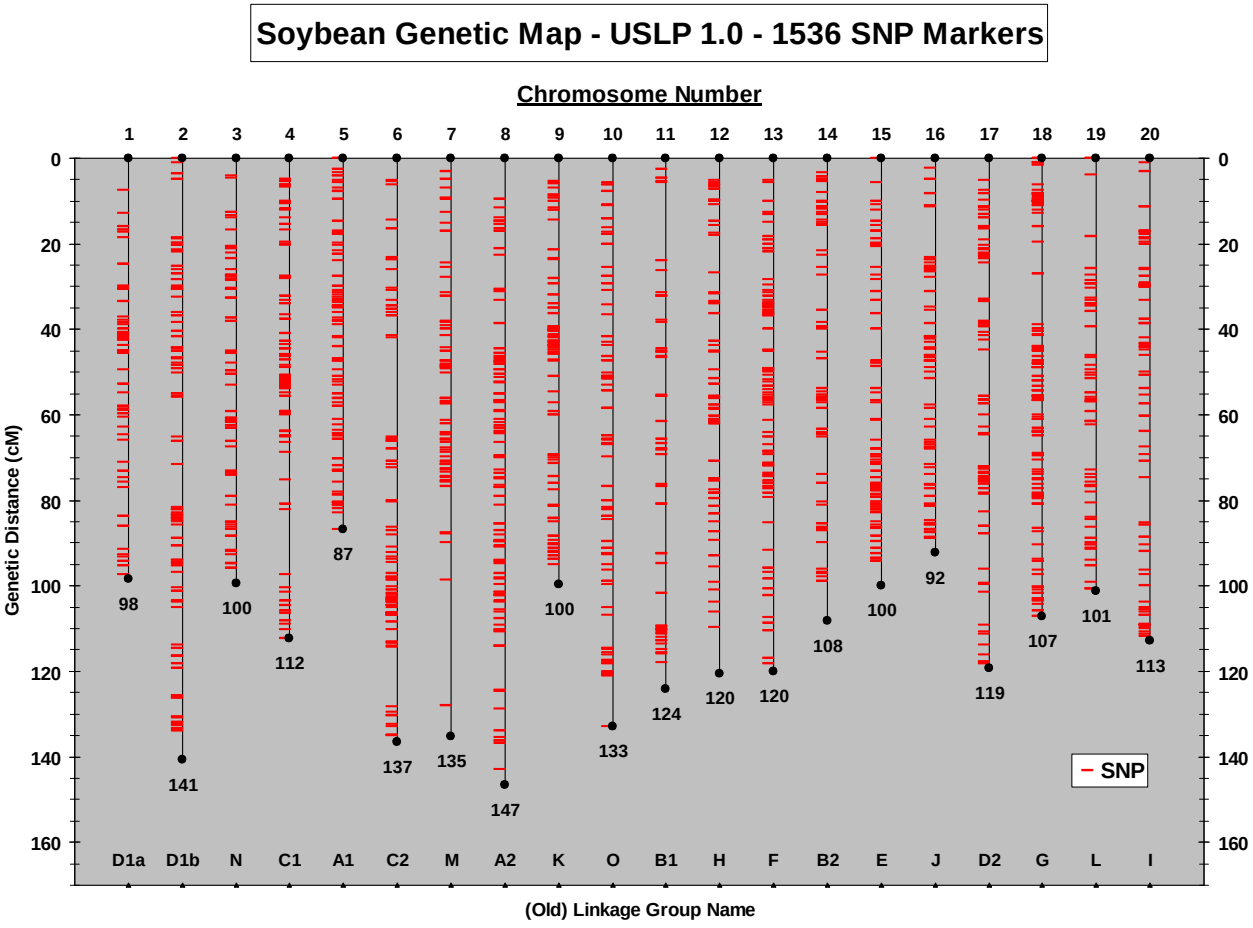


Fig. 2. The tickmarks on the vertical lines in this graph represent the map positions of the 1536 SNP markers within each of the 20 soybean linkage groups (bottom axis) and corresponding chromosomes (top axis). This set of SNP markers is called Universal Soy Linkage Panel 1.0 (Hyten et al., 2010). The vertical map distance is scaled in Kosambi centiMorgans.

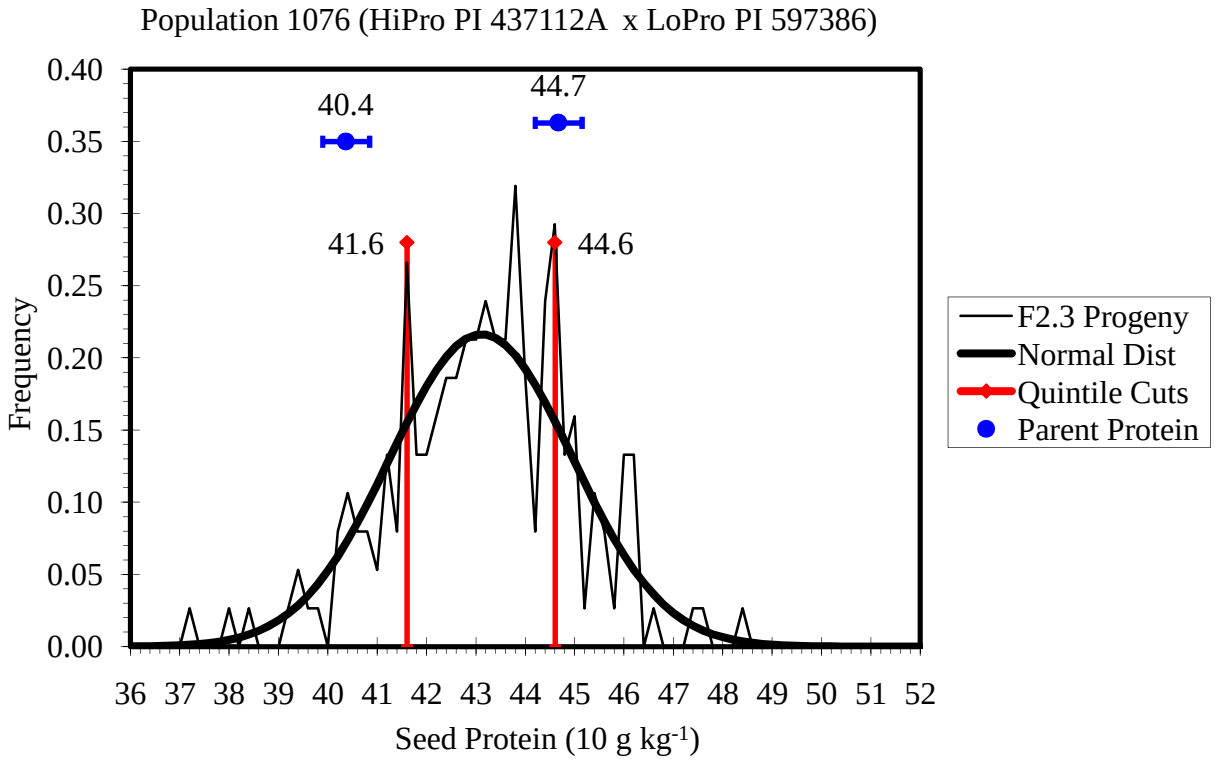
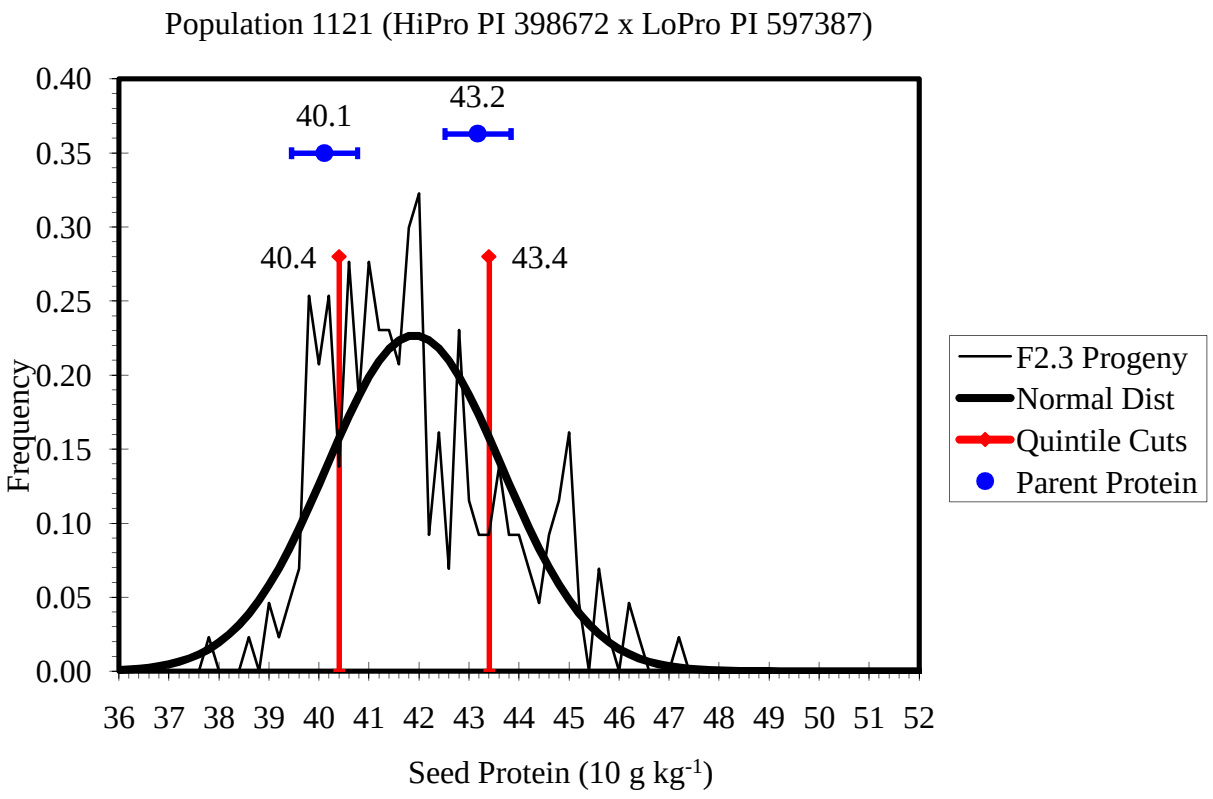


Fig. 3. Frequency distribution for seed protein content of $F_{2:3}$ progenies in the six soybean populations [1076, 1121, 1122, 1139, 1143, 1143 (moist), 1143 (dry), and 1146]. Also shown are mean seed protein values for the quintile (20%) low and high protein parents, as are the seed protein values defining the boundaries of the lowest and highest quintile fractions. See Table 3 for data on the tests for normality, skewness, and kurtosis of each distribution.


Population 1121 (HiPro PI 398672 x LoPro PI 597387)
40.1
43.2
40.4
43.4
Frequency
Seed Protein (10 g kg-1)
F2.3 Progeny
Normal Dist
Quintile Cuts
Parent Protein

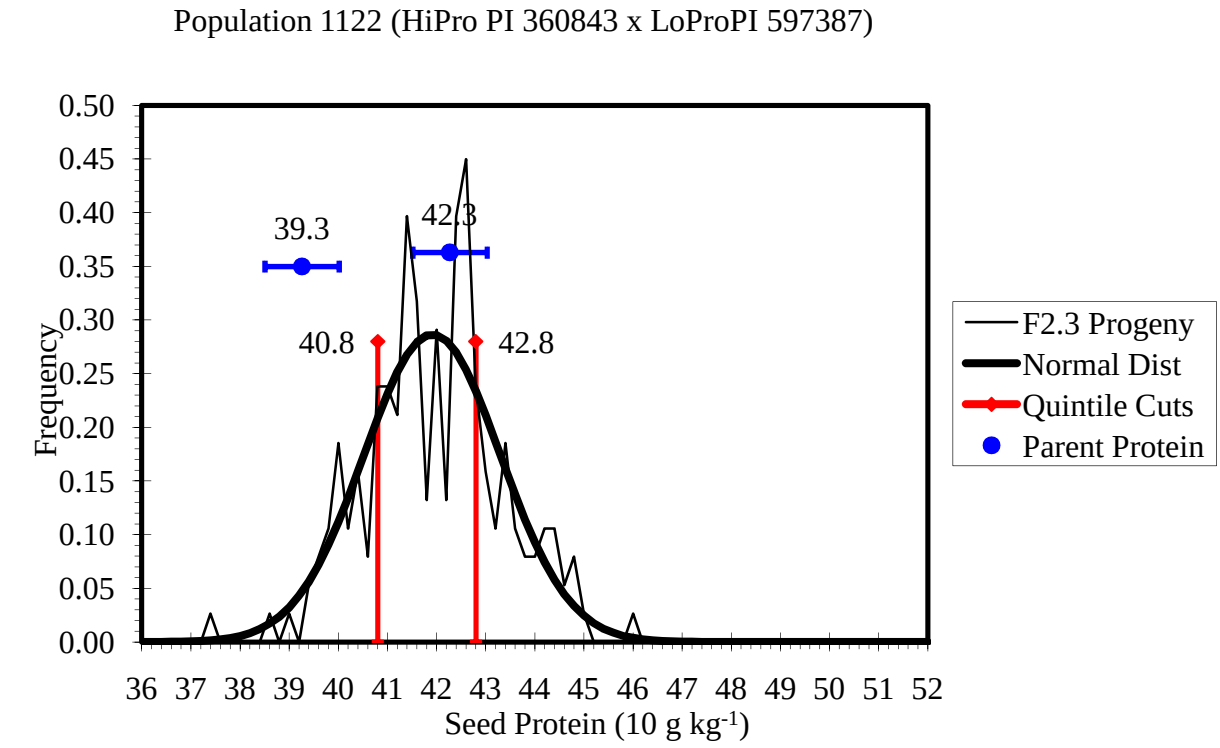

Population 1122 (HiPro PI 360843 x LoProPI 597387)
39.3
42.3
40.8
42.8
Frequency
Seed Protein (10 g kg-1)
F2.3 Progeny
Normal Dist
Quintile Cuts
Parent Protein

Fig. 3. (Cont.)

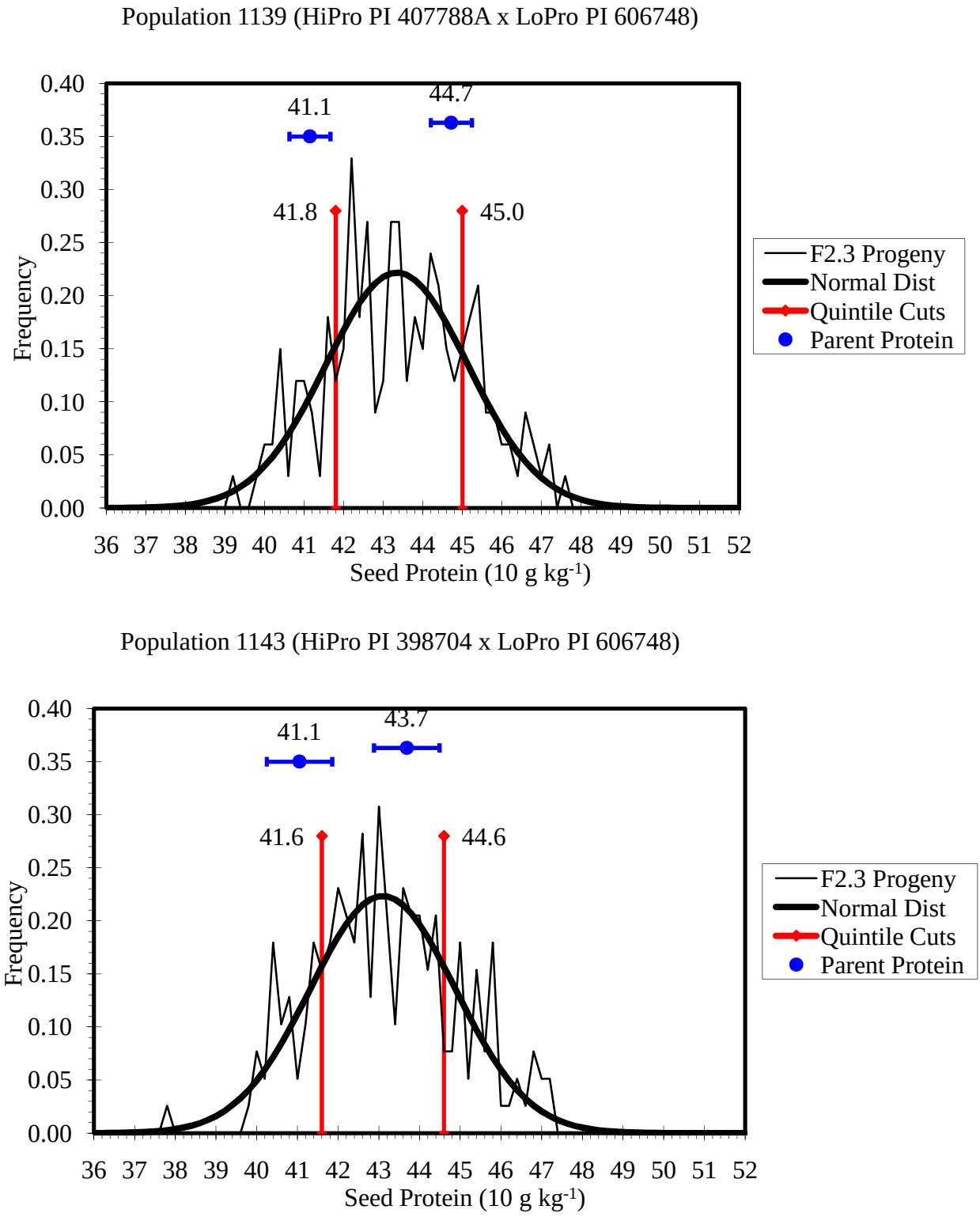

Population 1139 (HiPro PI 407788A x LoPro PI 606748)
41.1
44.7
41.8
45.0
Frequency
Seed Protein (10 g kg-1)
F2.3 Progeny
Normal Dist
Quintile Cuts
Parent Protein
Population 1143 (HiPro PI 398704 x LoPro PI 606748)
41.1
43.7
41.6
44.6
Frequency
Seed Protein (10 g kg-1)
F2.3 Progeny
Normal Dist
Quintile Cuts
Parent Protein

Fig. 3. (Cont.)

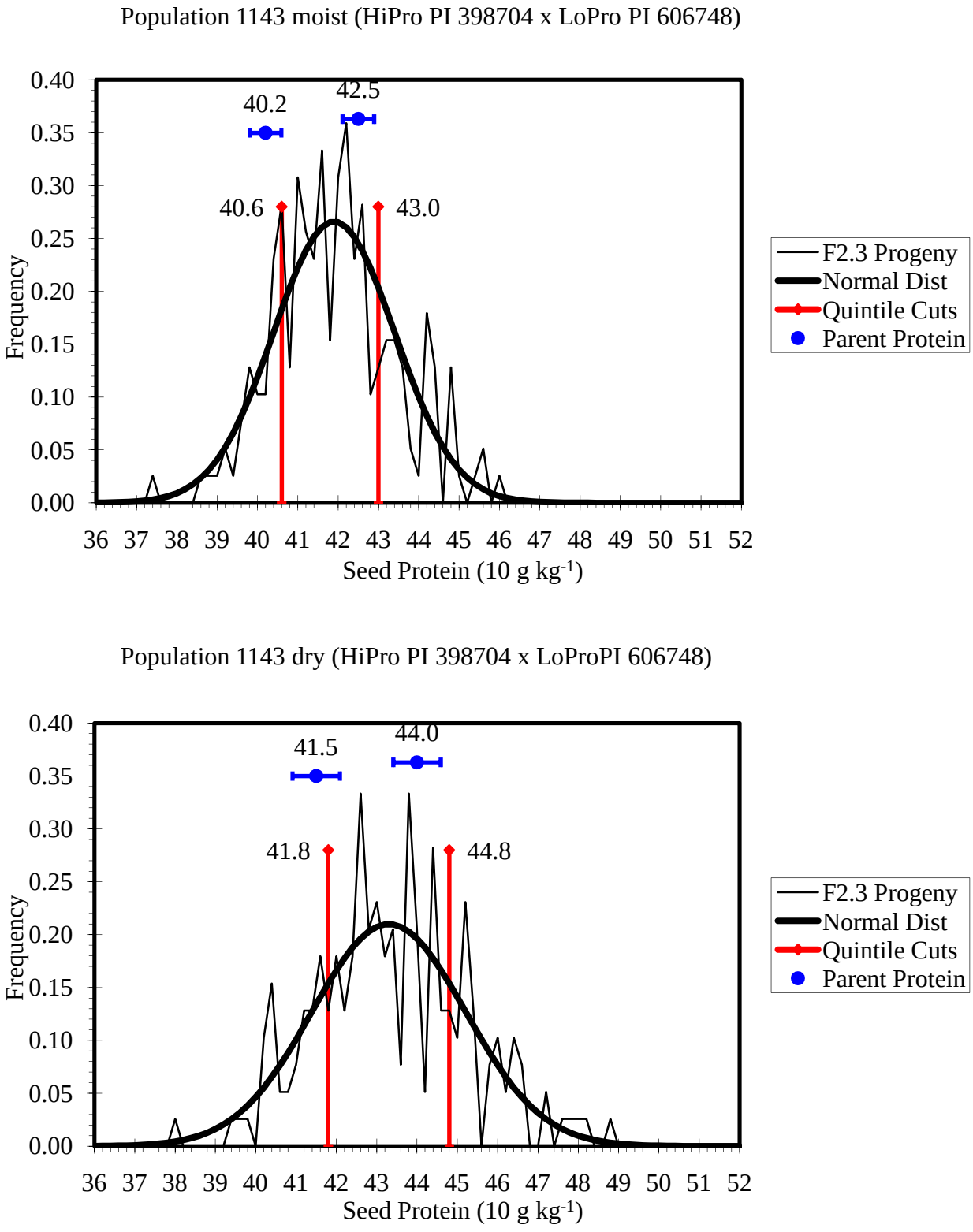

Population 1143 moist (HiPro PI 398704 x LoPro PI 606748)
Frequency
Seed Protein (10 g kg-1)
40.2
42.5
40.6
43.0
F2.3 Progeny
Normal Dist
Quintile Cuts
Parent Protein
Population 1143 dry (HiProPI 398704 x LoProPI 606748)
Frequency
Seed Protein (10 g kg-1)
41.5
44.0
41.8
44.8
F2.3 Progeny
Normal Dist
Quintile Cuts
Parent Protein

Fig. 3. (Cont.)

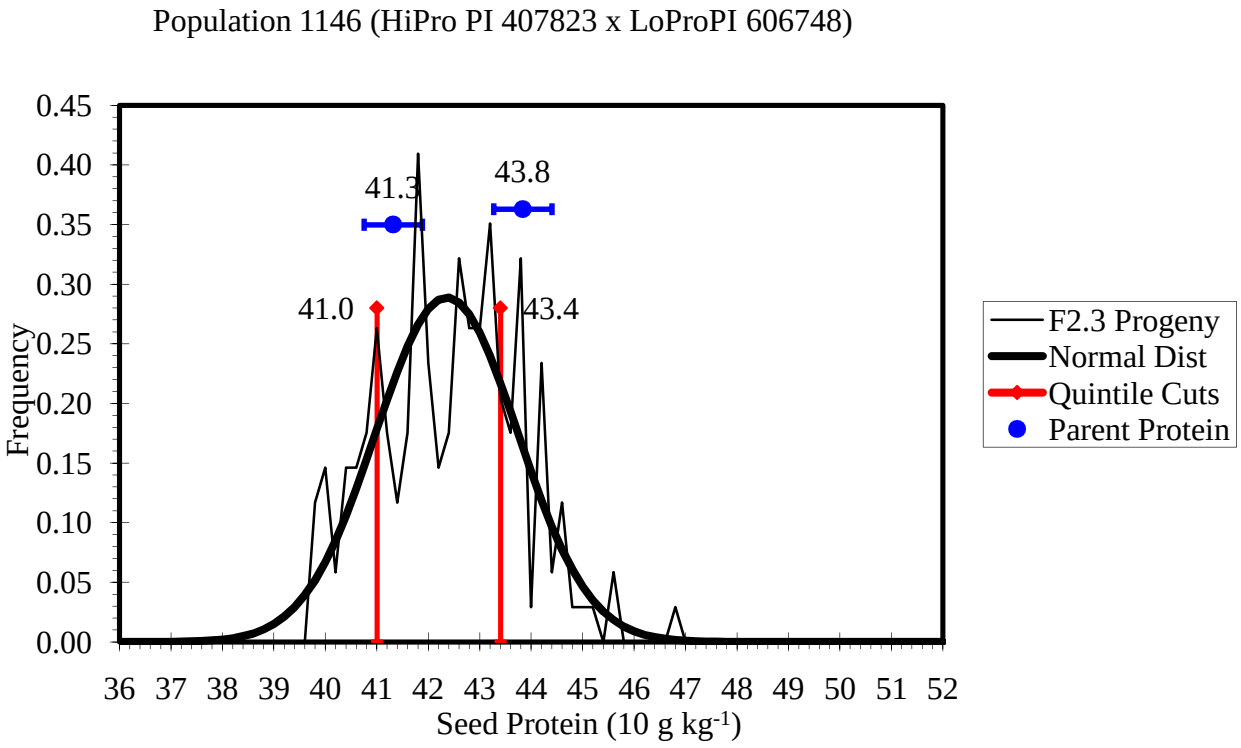

Population 1146 (HiPro PI 407823 x LoProPI 606748)
Frequency
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
41.3
43.8
41.0
43.4
36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
Seed Protein (10 g kg-1)
F2.3 Progeny
Normal Dist
Quintile Cuts
Parent Protein

Fig 3. (Cont.)

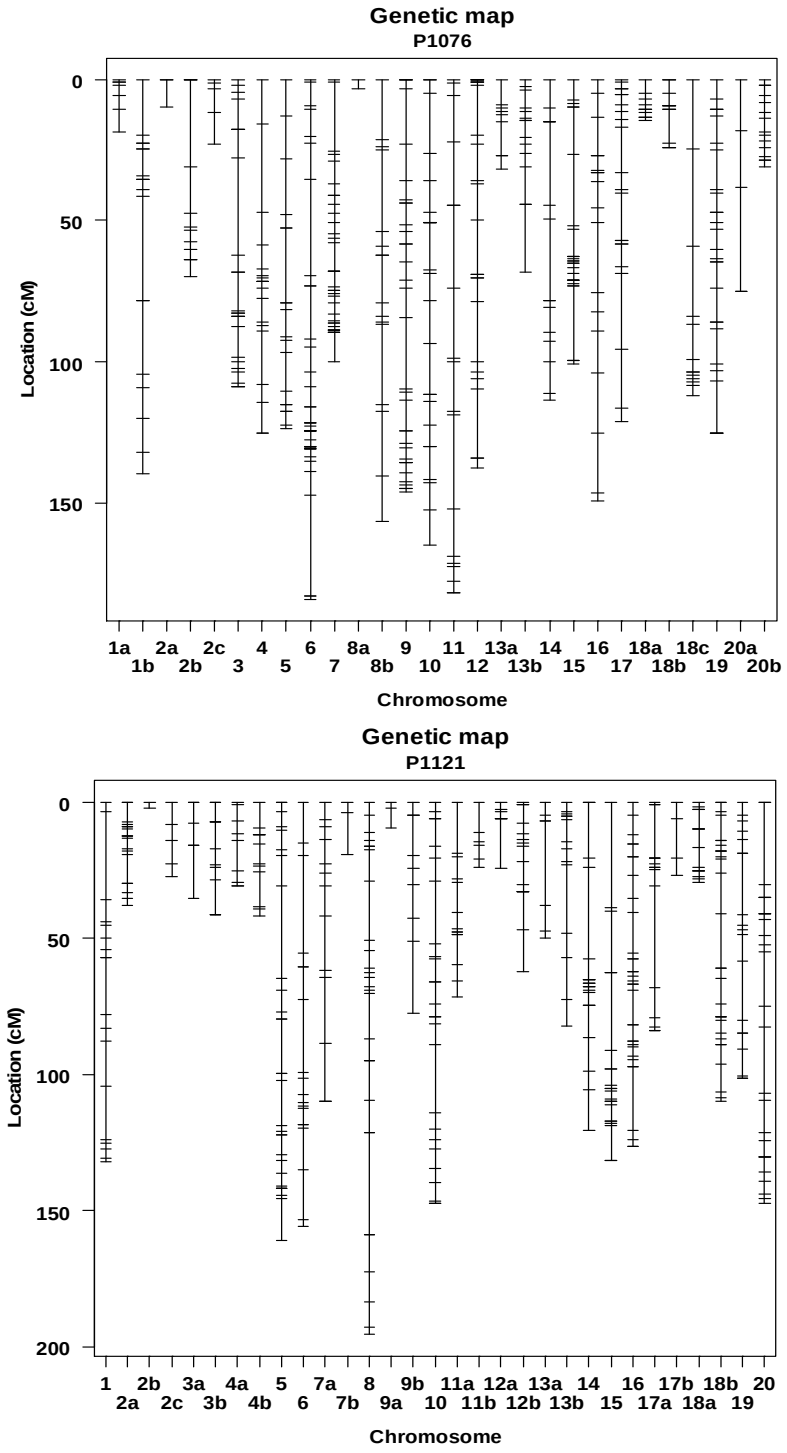


Fig. 4. The SNP marker genetic maps constructed for each of the six $F_{2:3}$ populations are presented here. About 400-500 SNP markers segregated in each population. In some instances, lack of marker polymorphism in some map positions required partitioning of a chromosome into two or sometimes three sub-chromosomes, which were labeled with a suffix of a, b, or c.

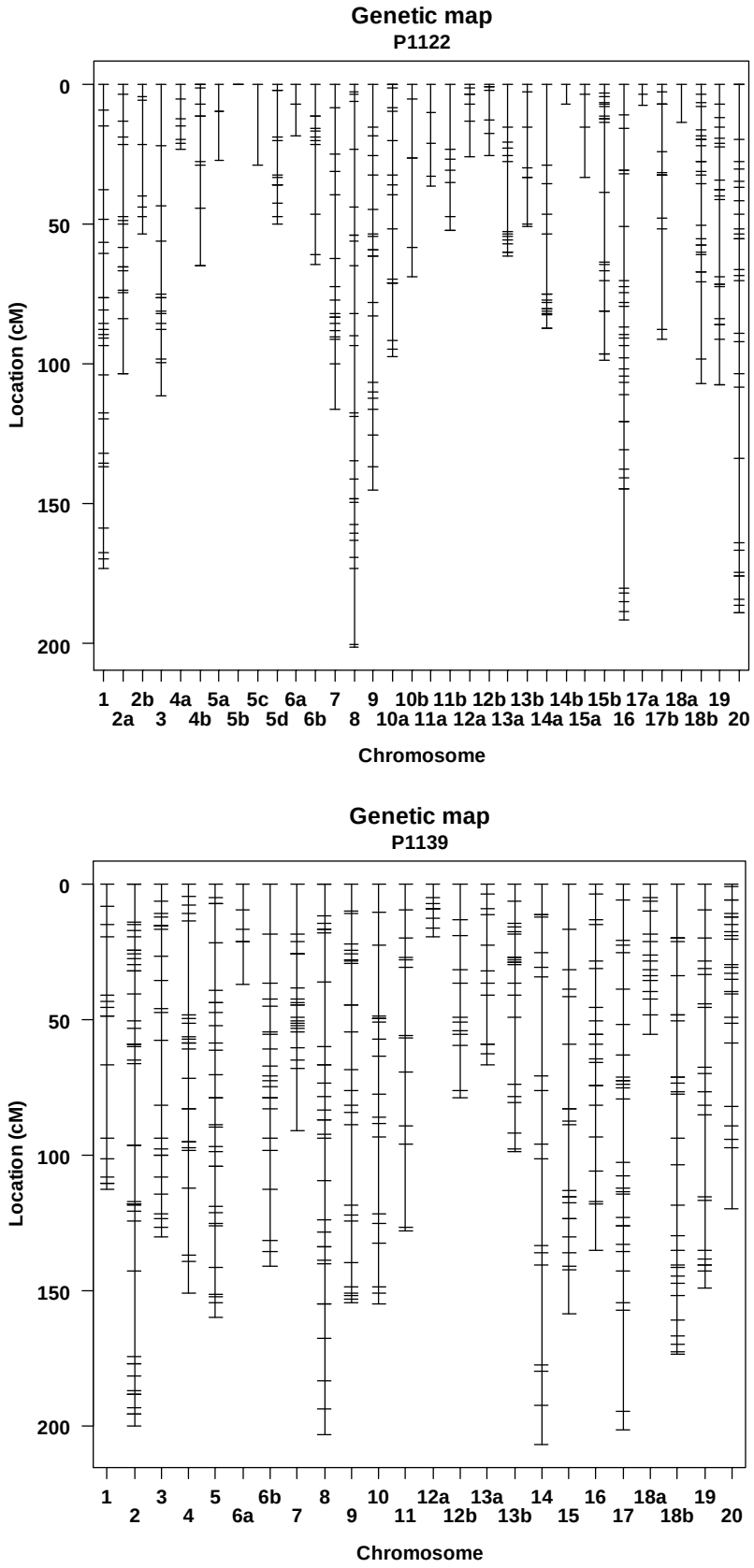
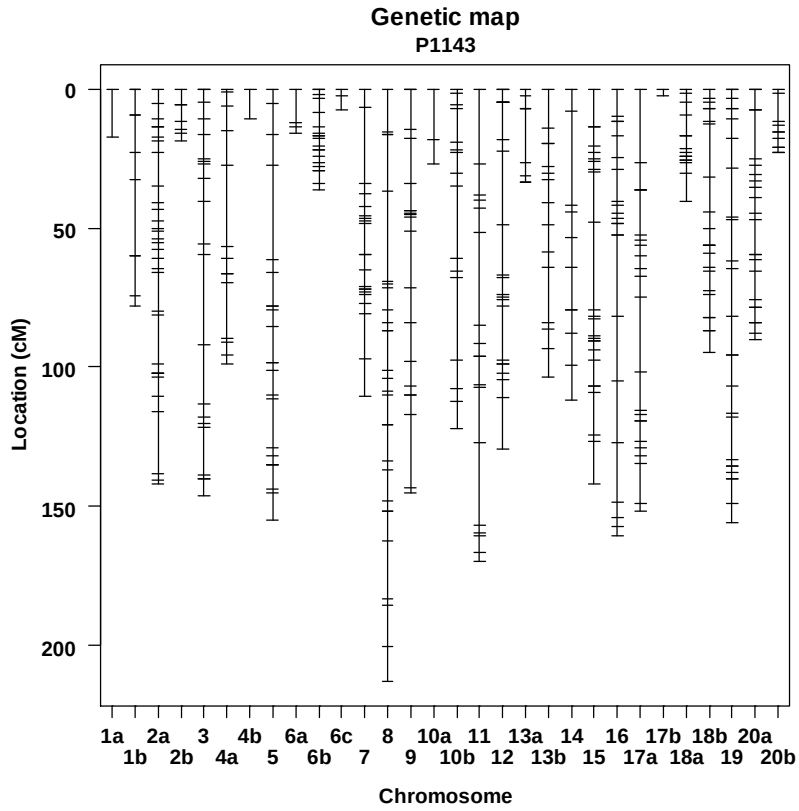

Genetic map
P1122
Location (cM)
0
50
100
150
200
1 2b 4a 5a 5c 6a 7 9 10b 11b 12b 13b 14b 15b 17a 18a 19
2a 3 4b 5b 5d 6b 8 10a 11a 12a 13a 14a 15a 16 17b 18b 20
Chromosome
Genetic map
P1139
Location (cM)
0
50
100
150
200
1 3 5 6b 8 10 12a 13a 14 16 18a 19
2 4 6a 7 9 11 12b 13b 15 17 18b 20
Chromosome

Fig. 4. (Cont.)


Genetic map
P1143
Location (cM)
0
50
100
150
200
1a 1b 2a 2b 3 4a 4b 5 6a 6b 6c 7 8 9 10a 10b 11 12 13a 13b 14 15 16 17a 17b 18a 18b 19 20a 20b
Chromosome

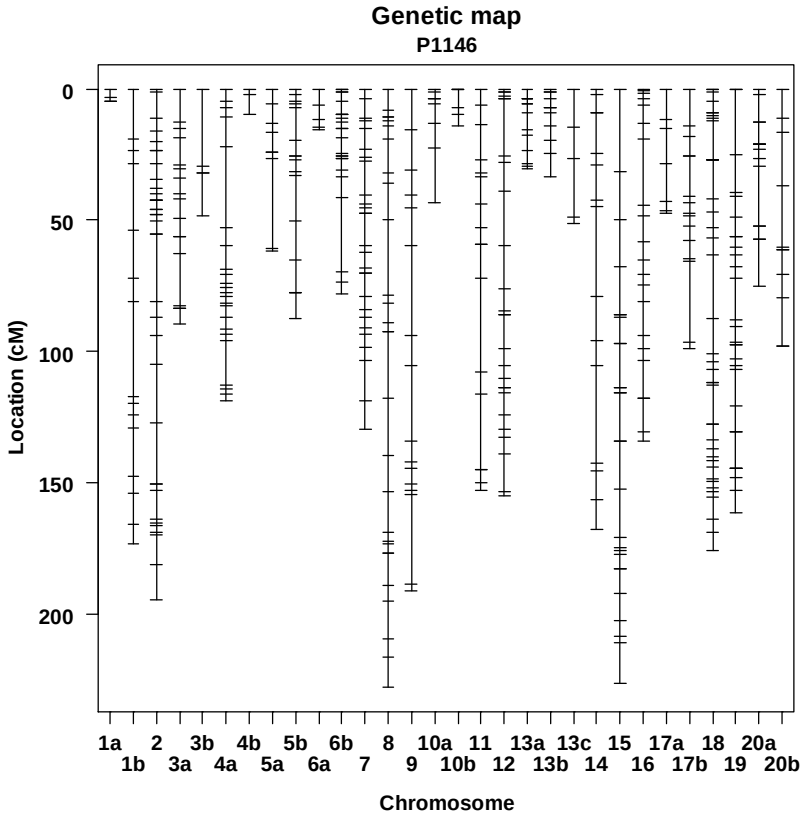

Genetic map
P1146
Location (cM)
0
50
100
150
200
1a 1b 2 3a 3b 4a 4b 5a 5b 6a 6b 7 8 9 10a 10b 11 12 13a 13b 13c 14 15 16 17a 17b 18 19 20a 20b
Chromosome

Fig. 4. (Cont.)

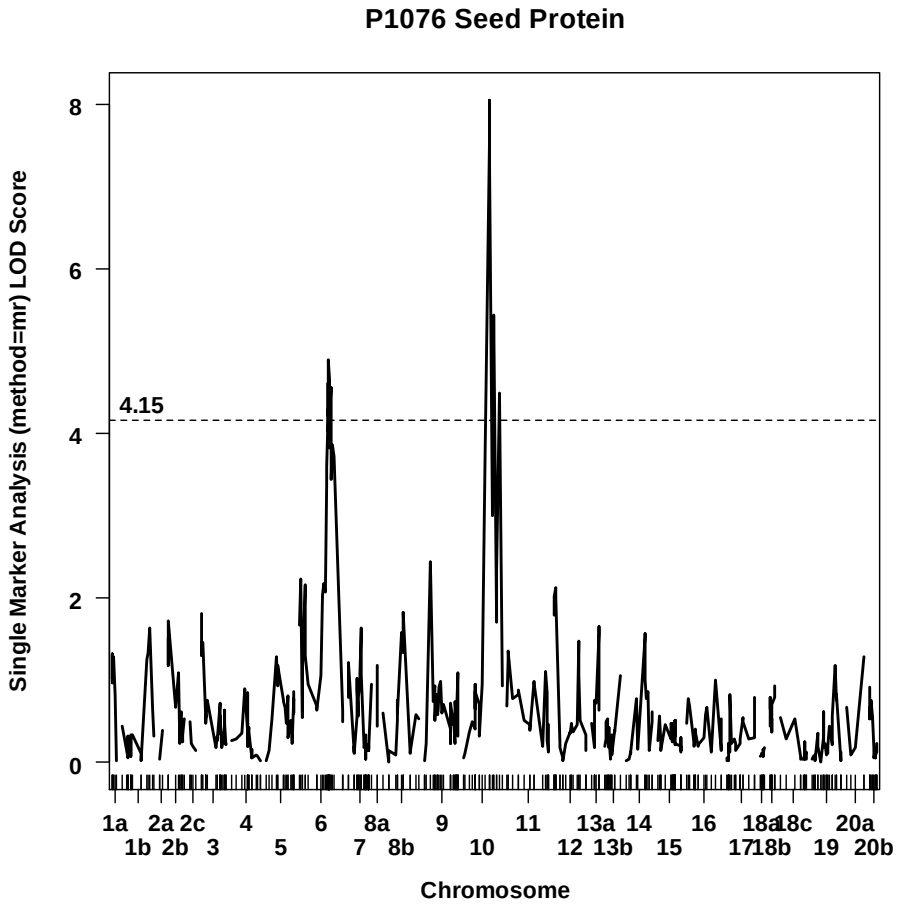


Fig. 5. Shown here are the genome-wide LOD score scans generated with the marker regression method with respect to the selectively genotyped $F_{2:3}$ progeny protein values in each of the six $F_{2:3}$ populations [1076, 1121, 1122, 1139, 1143, 1143w (moist), 1143d (dry), and 1146]. The LOD score criteria for significance (dashed line) in each population was determined by using the 95^{th} percentile of genome-wide maximum LOD scores obtained from 1000 replicates of stratified permutation.

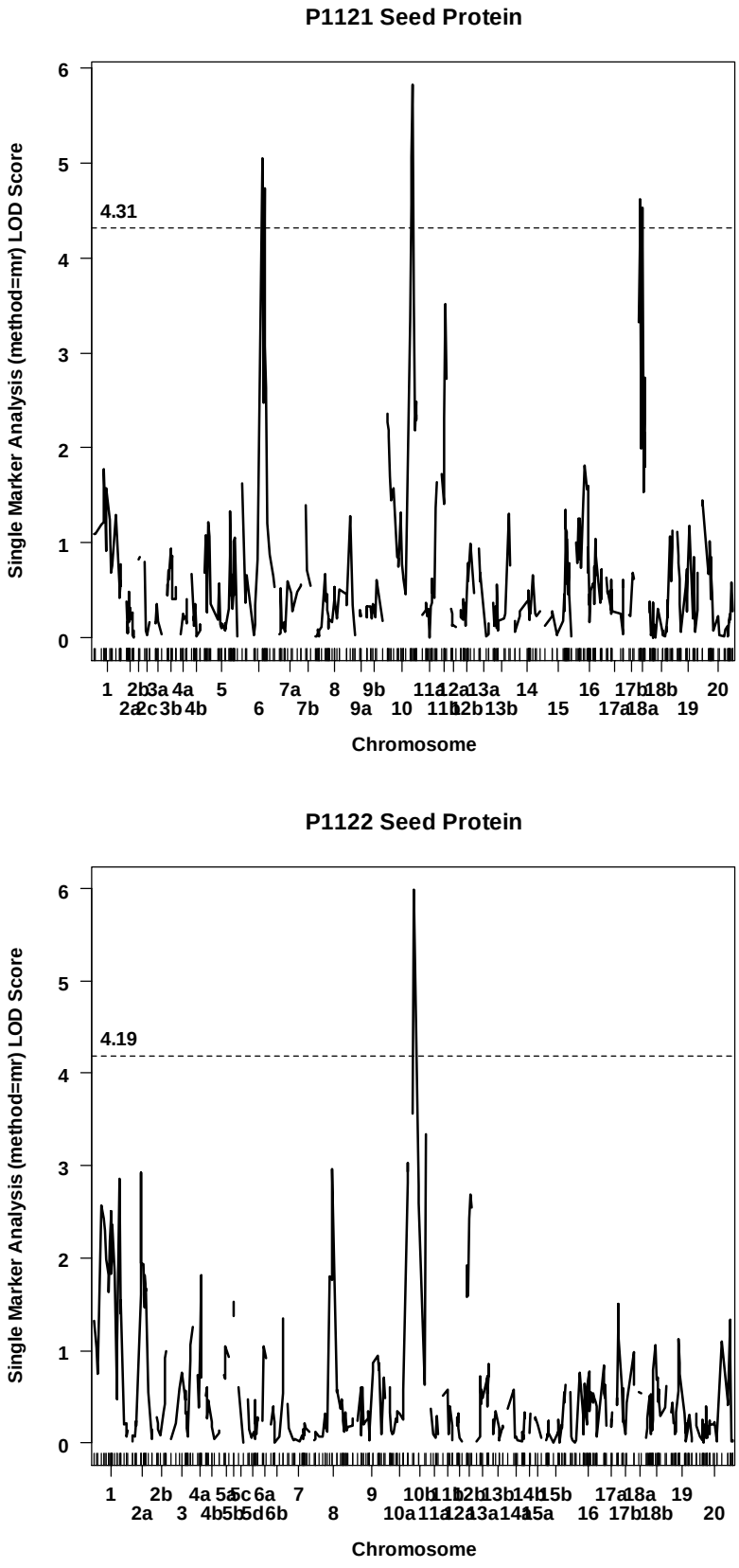

P1121 Seed Protein
Single Marker Analysis (method=mr) LOD Score
4.31
Chromosome
P1122 Seed Protein
Single Marker Analysis (method=mr) LOD Score
4.19
Chromosome

Fig. 5. (Cont.)

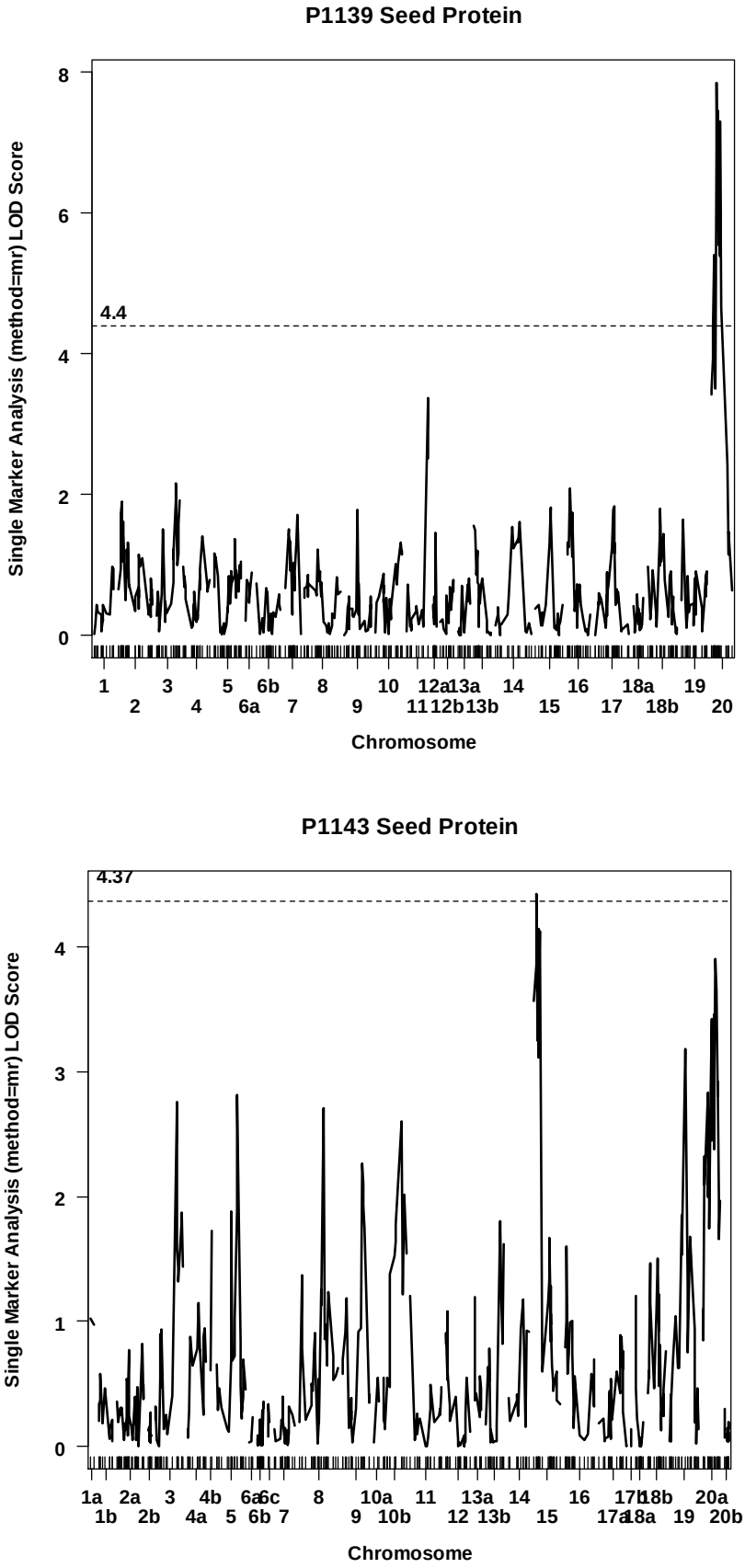

P1139 Seed Protein
Single Marker Analysis (method=mr) LOD Score
4.4
Chromosome
P1143 Seed Protein
Single Marker Analysis (method=mr) LOD Score
4.37
Chromosome

Fig. 5. (Cont.)

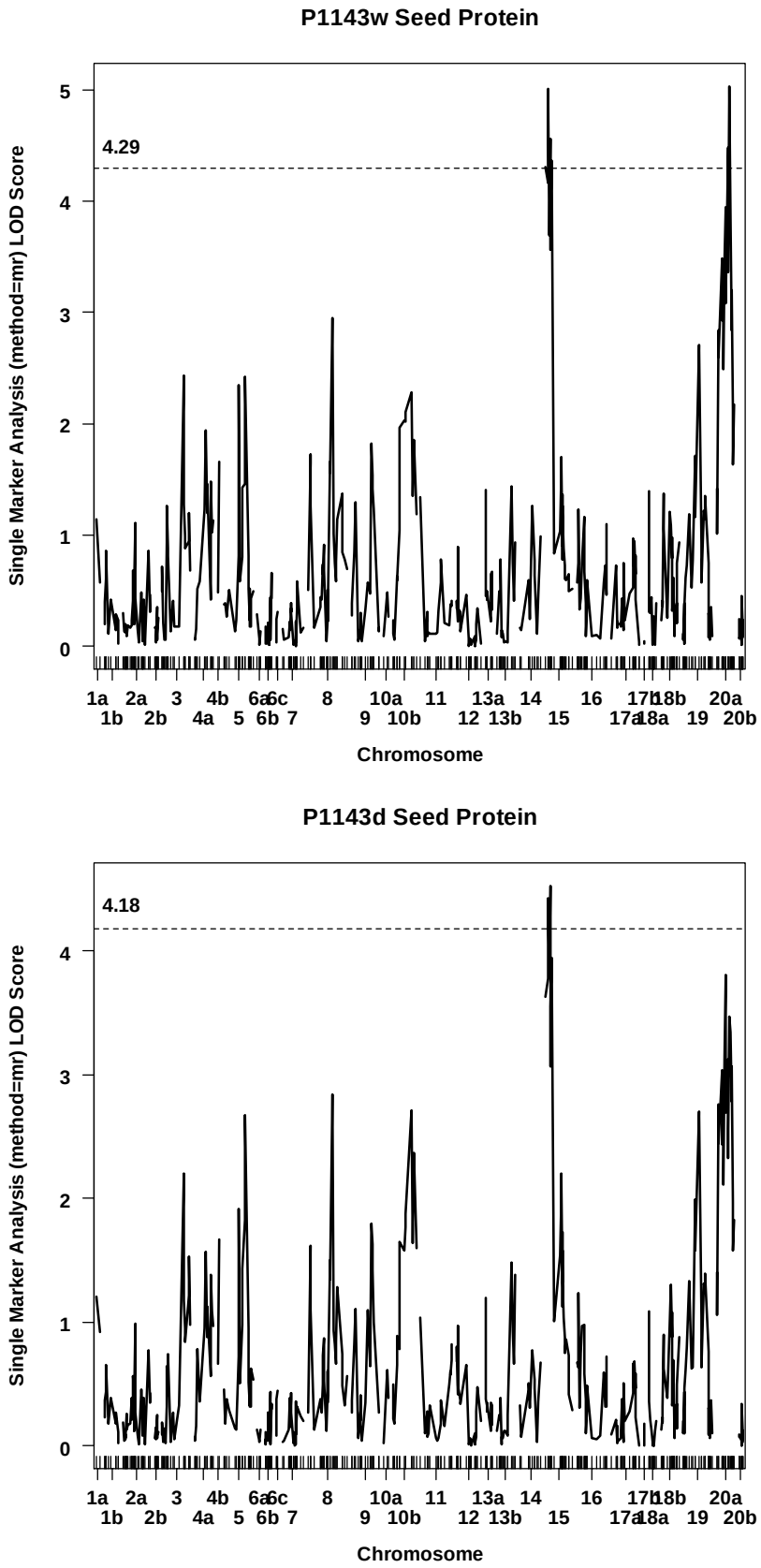

P1143w Seed Protein
Single Marker Analysis (method=mr) LOD Score
4.29
0
1
2
3
4
5
1a 1b 2a 2b 3 4a 4b 5 6a 6b 6c 7 8 9 10a 10b 11 12 13a 13b 14 15 16 17a 17b 18a 18b 19 20a 20b
Chromosome
P1143d Seed Protein
Single Marker Analysis (method=mr) LOD Score
4.18
0
1
2
3
4
1a 1b 2a 2b 3 4a 4b 5 6a 6b 6c 7 8 9 10a 10b 11 12 13a 13b 14 15 16 17a 17b 18a 18b 19 20a 20b
Chromosome

Fig. 5. (Cont.)

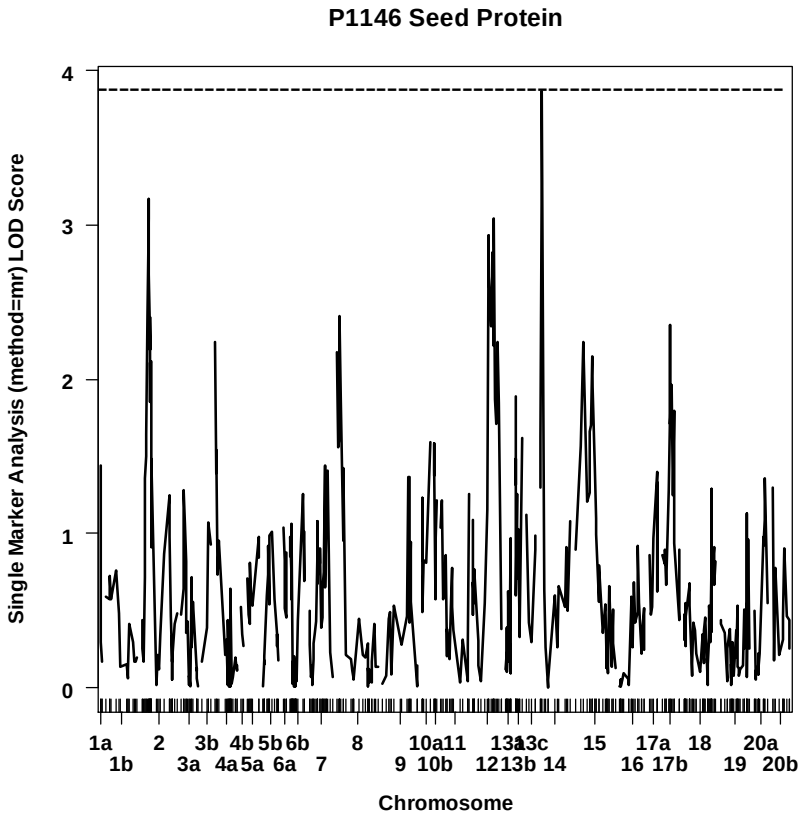

P1146 Seed Protein
Single Marker Analysis (method=mr) LOD Score
4
3
2
1
0
1a
1b
2
3a
3b
4a
4b
5a
5b
6a
6b
7
8
9
10a
10b
11
12
13a
13b
13c
14
15
16
17a
17b
18
19
20a
20b
Chromosome

Fig. 5. (Cont.)

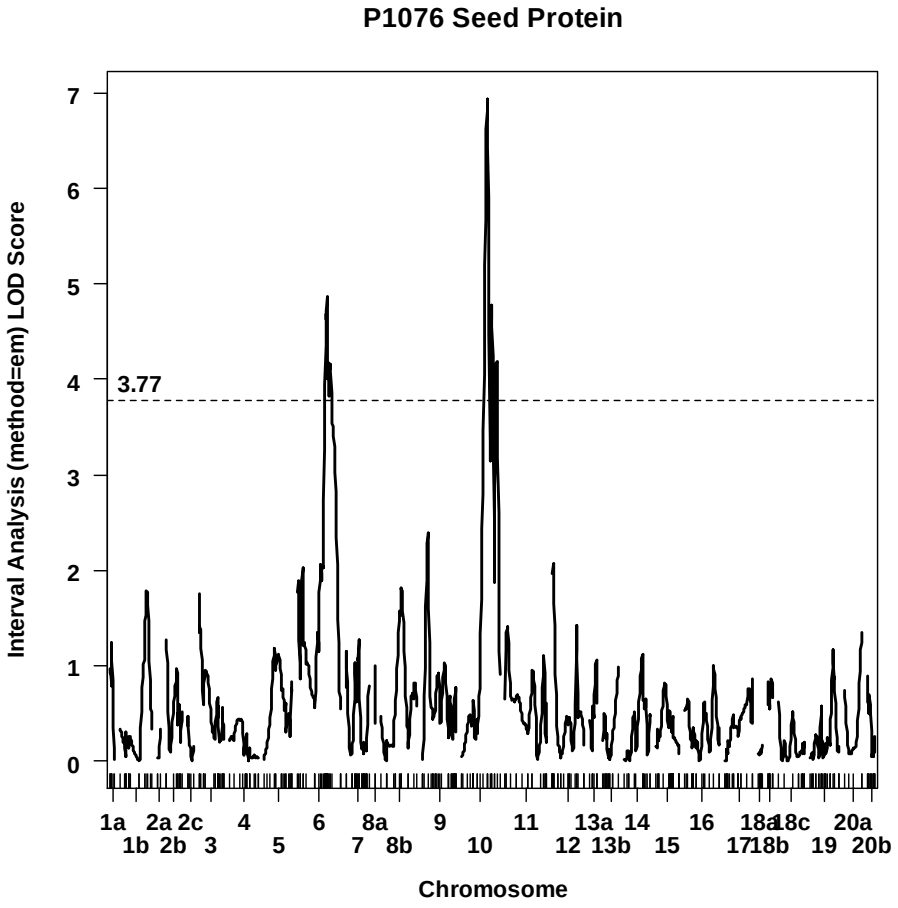


Fig. 6. Shown here are the genome-wide LOD score scans generated using the interval analysis method (i.e., maximum likelihood approach using the EM algorithm) with respect to the selectively genotyped $F_{2:3}$ progeny seed protein values in each of the six $F_{2:3}$ populations [1076, 1121, 1122, 1139, 1143, 1143w (moist), 1143d (dry), and 1146]. The LOD score criteria for significance (dashed line) in each population was determined by using the 95^{th} percentile of genome-wide maximum LOD scores obtained from 1000 replicates of stratified permutation.

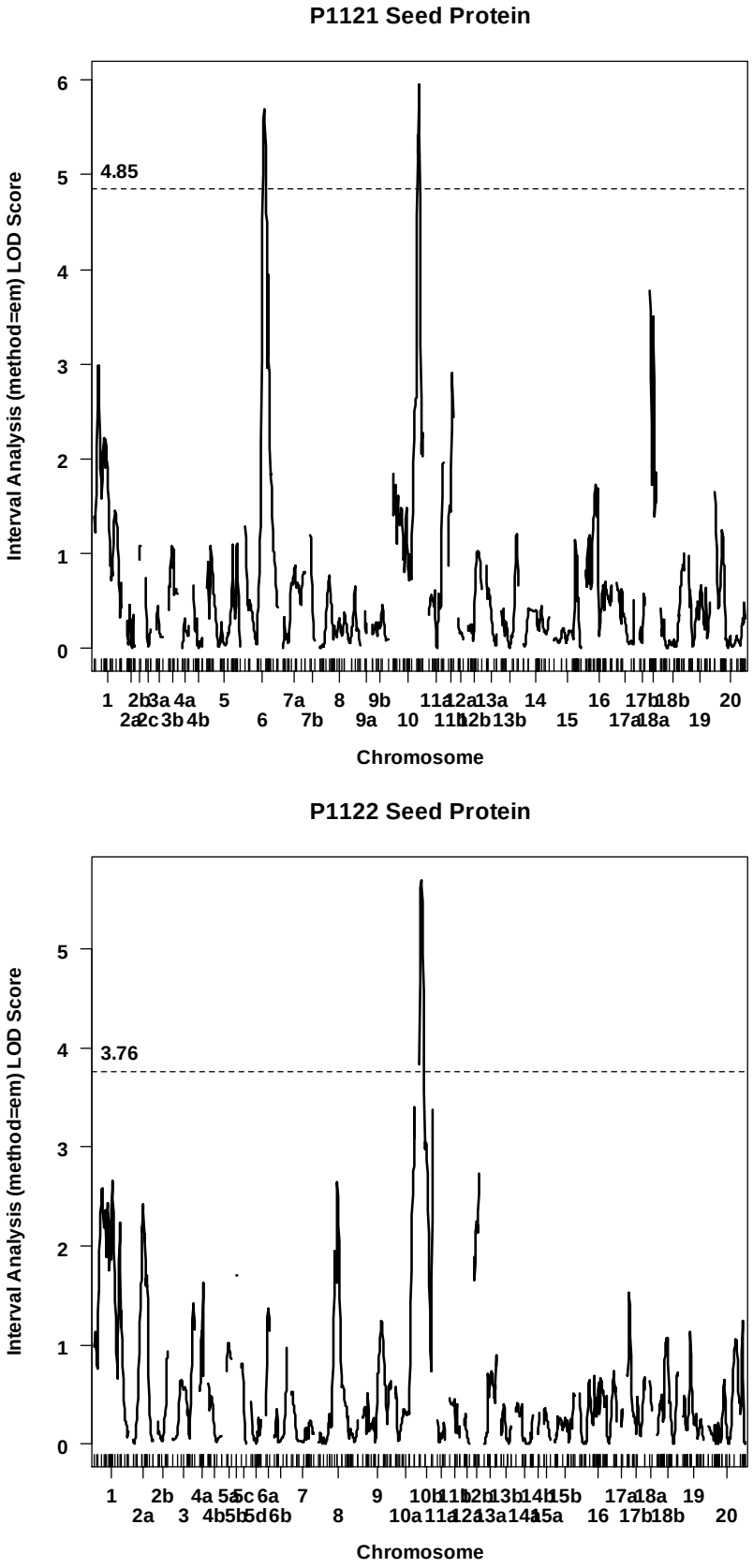

P1121 Seed Protein
Interval Analysis (method=em) LOD Score
4.85
Chromosome
P1122 Seed Protein
Interval Analysis (method=em) LOD Score
3.76
Chromosome

Fig. 6. (Cont.)

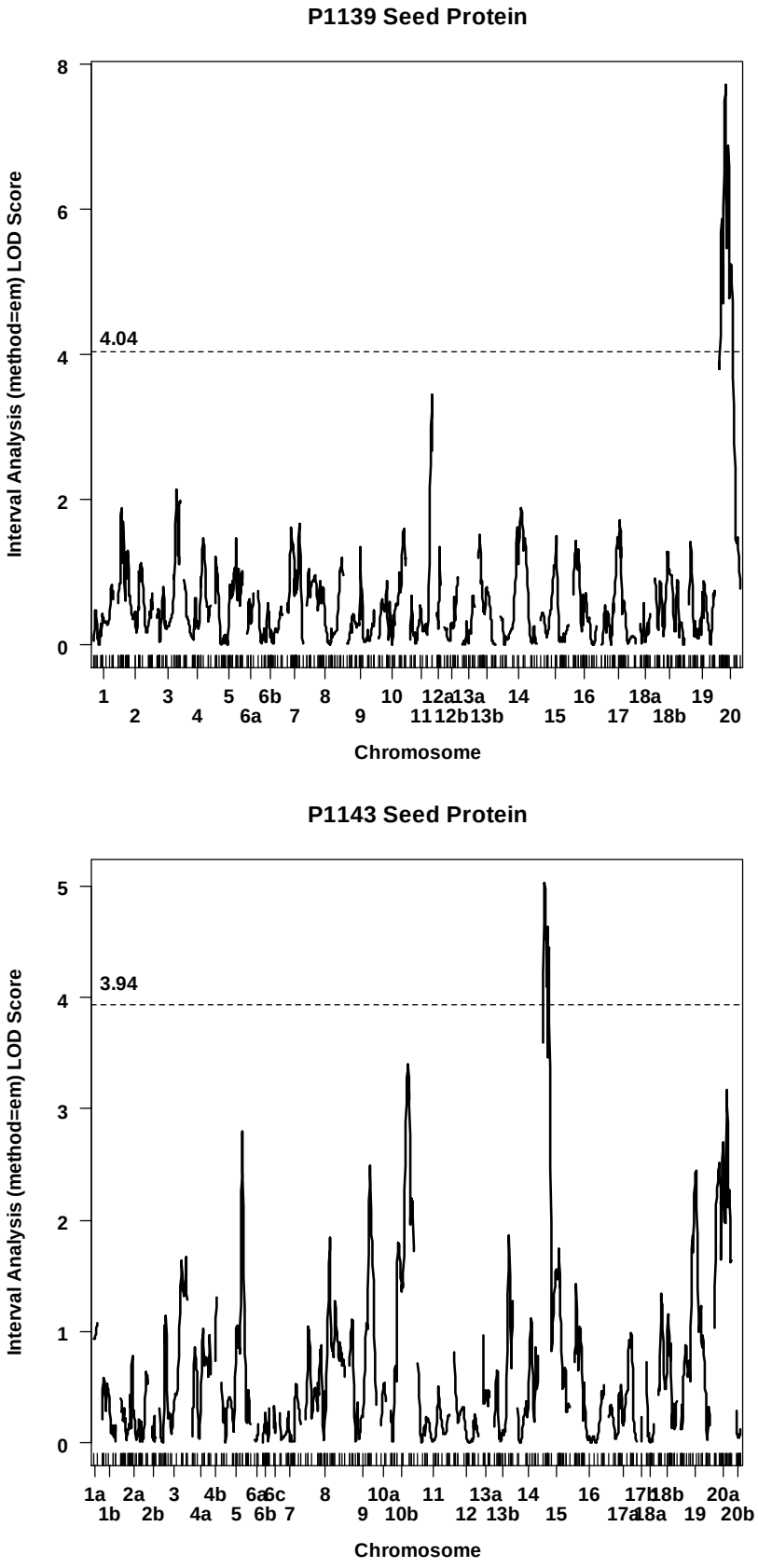

P1139 Seed Protein
Interval Analysis (method=em) LOD Score
4.04
Chromosome
P1143 Seed Protein
Interval Analysis (method=em) LOD Score
3.94
Chromosome

Fig. 6. (Cont.)

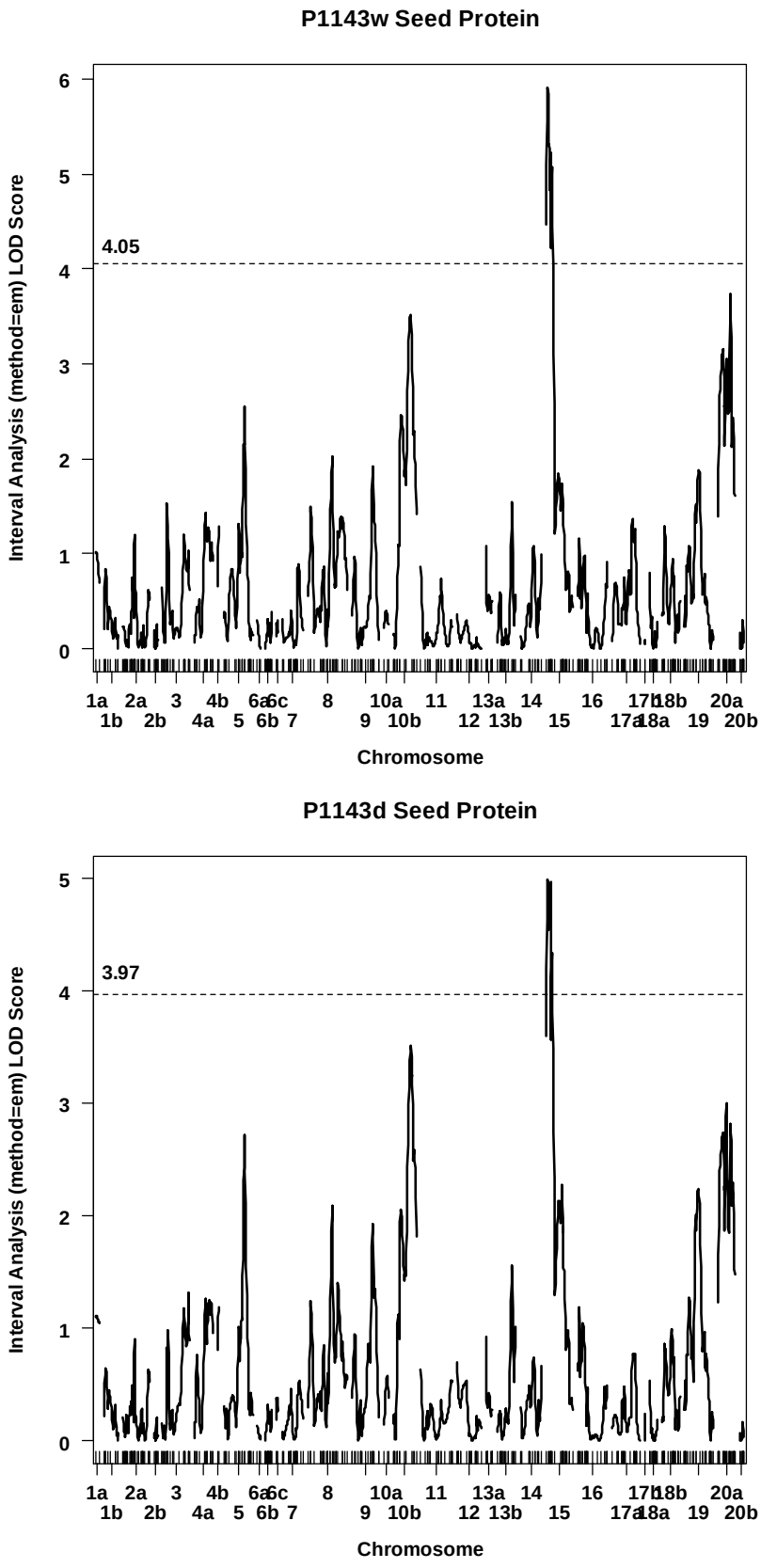

P1143w Seed Protein
Interval Analysis (method=em) LOD Score
4.05
0
1
2
3
4
5
6
1a 1b 2a 2b 3 4a 4b 5 6a 6b 6c 7 8 9 10a 10b 11 12 13a 13b 14 15 16 17a 17b 18a 18b 19 20a 20b
Chromosome
P1143d Seed Protein
Interval Analysis (method=em) LOD Score
3.97
0
1
2
3
4
5
1a 1b 2a 2b 3 4a 4b 5 6a 6b 6c 7 8 9 10a 10b 11 12 13a 13b 14 15 16 17a 17b 18a 18b 19 20a 20b
Chromosome

Fig. 6. (Cont.)

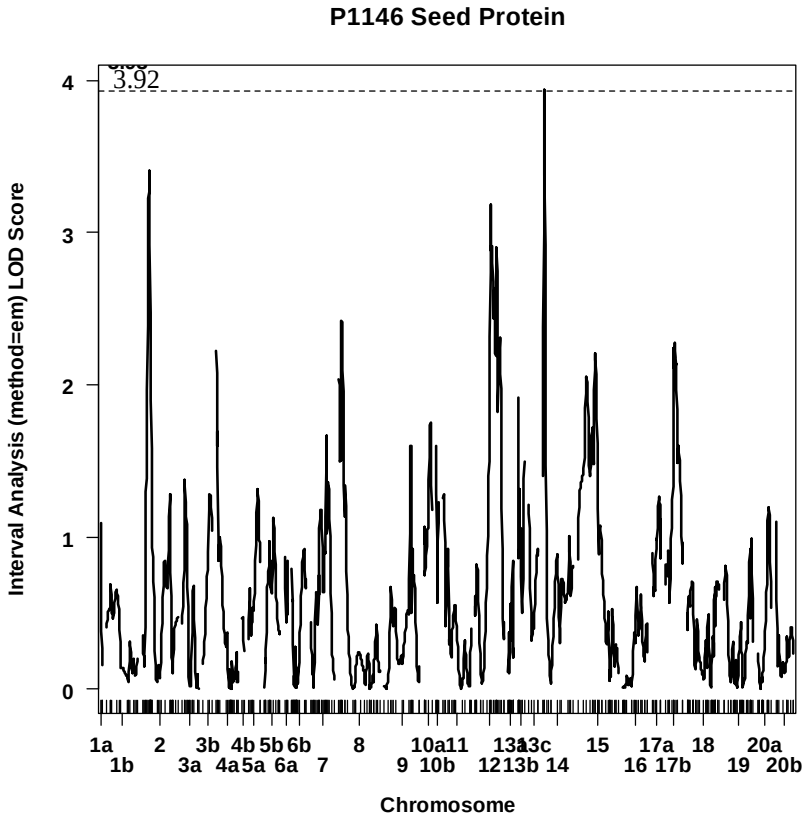

P1146 Seed Protein
3.92
Interval Analysis (method=em) LOD Score
0
1
2
3
4
1a
1b
2
3a
3b
4a
4b
5a
5b
6a
6b
7
8
9
10a
10b
11
12
13a
13b
13c
14
15
16
17a
17b
18
19
20a
20b
Chromosome

Fig. 6. (Cont.)

REFERENCES

Akkaya, M.S., A.A. Bhagwat, and P.B. Cregan. 1992. Length polymorphisms of simple sequence repeat DNA in soybean. Genetics 132:1131-1139.

Ayoub, M., and D.E. Mather. 2002. Effectiveness of selective genotyping for detection of quantitative trait loci: an analysis of grain and malt quality traits in three barley populations. Genome 45:1116-1124.

Bernardo, R. 2002. Mapping quantitative trait loci, p. 369, *In* R. Bernardo, ed. Breeding for quantitative traits in plants. Stemma Press, Woodbury, MN.

Brim, C.A. 1973. Quantitative genetics and breeding, p. 155–186, *In* B. E. Caldwell, ed. Soybeans: Improvement, production, and uses. ASA, Madison, WI.

Brim, C.A., and J.W. Burton. 1979. Recurrent selection in soybeans. II. Selection for increased percent protein in seeds. Crop Sci. 19:494-498.

Broman, K.W., and S. Sen. 2009. A guide to QTL mapping with R/qtl Springer, New York.

Broman, K.W., and T.P. Speed. 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. J. R. Statist. Soc. B 64: part4: 641-656.

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, J.W. 1987. Quantitative genetics: Results relevant to soybean breeding, p. 211-242, *In* J. R. Wilcox, ed. Soybeans: Improvement, production, and uses, 2nd ed. ASA, CSSA, and SSSA, Madison, WI.

Caldwell, B.E., C.R. Weber, and D.E. Byth. 1966. Selection value of phenotypic attributes in soybeans. Crop Sci. 6:249-251.

Chapman, A., V.R. Pantalone, A. Ustun, F.L. Allen, D. Landau-Ellis, R.N. Trigiano, and P.M. Gresshoff. 2003. Quantitative trait loci for agronomic and seed quality traits in an F2 and F4:6 soybean population. Euphytica 129:387-393.

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.

Cianzio, S.R. 2007. Soybean breeding achievements and challenges, p. 245-274, *In* S. K. Manjit and P. M. Priyadarshan, eds. Breeding major food staples, 1st ed. Blackwell Publishing Ltd, Ames, IA.

Cober, E.R., and H.D. Voldeng. 2000. Developing high-protein, high-yield soybean populations and lines. Crop Sci. 40:39-42.

Collard, B., M. Jahufer, J. Brower, and E. Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169-196.

Coryell, V.H., H. Jessen, J.M. Schupp, D. Webb, and P. Keim. 1999. Allele-specific hybridization markers for soybean. Theor. Appl. Genet. 98:690-696.

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, N. Kaya, T.T. VanToai, D.G. Lohnes, J. Chung, and J.E. Specht. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. 39:1464-1490.

Csanádi, 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. 1997. The effect of selective genotyping on QTL mapping accuracy. Mamm. Genome 8:67-68.

Darvasi, A., and M. Soller. 1992. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor. Appl. Genet. 85:353-359.

Darvasi, A., and M. Soller. 1994. Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics 138:1365-1373.

Darvasi, A., A. Weinreb, V. Minke, J.I. Weller, and M. Soller. 1993. Detecting Marker-QTL Linkage and Estimating QTL Gene Effect and Map Location Using a Saturated Genetic Map. Genetics 134:943-951.

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.

Doerge, R.W. 2002. Mapping and analysis of quantitative trait loci in experimental populations. Nat. Rev. Genet., 3: 43-52.

Dornbos, D.L., Jr., and R.E. Mullen. 1992. Soybean seed protein and oil contents and fatty acid composition adjustments by drought and temperature. J. Am. Oil Chem. Soc. 69:228-231.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. 5:435-445.

Falconer, D.S. 1981. Introduction to quantitative genetics, Longman, NY.

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.

Gallais, A., L. Moreau, and A. Charcosset. 2007. Detection of marker–QTL associations by studying change in marker frequencies with selection. Theor. Appl. Genet. 114:669-681.

Gunter, K., M. Andrea, T. Nigel, S. Richard, and B. Dirk van den. 2005. Single nucleotide polymorphisms: Detection techniques and their potential for genotyping and genome mapping, p. 75-107, *In* M. Khalid and K. Gunter, eds. The handbook of plant genome mapping: Genetic and physical mapping. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

Hartwig, E.E. 1973. Varietal development, p. 187–210, *In* B. E. Caldwell, ed. Soybeans: Improvement, production, and uses, Vol. 16. ASA, Madison, WI.

Hartwig, E.E., and K. Hinson. 1972. Association between chemical composition of seed and seed yield of soybeans. Crop Sci. 12:829-830.

Hartwig, E.E., and T.C. Kilen. 1991. Yield and composition of soybean seed from parents with different protein, similar yield. Crop Sci. 31:290-292.

Helms, T.C., and J.H. Orf. 1998. Protein, oil, and yield of soybean lines selected for increased protein. Crop Sci. 38:707-711.

Helms, T.C., C.R. Hurburgh, Jr., R.L. Lussenden, and D.A. Whited. 1990. Economic analysis of increased protein and decreased yield due to delayed planting of soybean. J. Prod. Agric. 3:367-371.

Howell, R.W., and J.L. Cartter. 1958. Physiological factors affecting composition of soybeans: II. Response of oil and other constituents of soybeans to temperature under controlled conditions. Agron. J. 50:664-667.

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

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., 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. USA 103:16666-16671.

Hyten, D.L., I.-Y. Choi, Q. Song, J.E. Specht, T.E. Carter, R.C. Shoemaker, E.-Y. Hwang, L.K. Matukumalli, and P.B. Cregan. 2010. A high density integrated genetic linkage map of soybean and the development of a 1536 Universal Soy Linkage Panel for quantitative trait locus mapping. Crop Sci. 50:960-968.

Jun, T.-H., K. Van, M. Kim, S.-H. Lee, and D. Walker. 2008. Association analysis using SSR markers to find QTL for seed protein content in soybean. Euphytica 162:179-191.

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.

Kearsey, M.J., and V. Hyne. 1994. QTL analysis: a simple 'marker-regression' approach. Theor. Appl. Genet. 89:698-702.

Kearsey, M.J., and H.S. Pooni. 1996. The genetical analysis of quantitative traits. Chapman&Hall, Boundary Row, London.

Kearsey, M.J., and A.G.L. Farquhar. 1998. QTL analysis in plants; where are we now? Heredity 80:137-142.

Keim, P., B.W. Diers, T.C. Olson, and R.C. Shoemaker. 1990. RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735-742.

Keim, P., J.M. Schupp, S.E. Travis, K. Clayton, T. Zhu, L. Shi, A. Ferreira, and D.M. Webb. 1997. A high-density soybean genetic map based on AFLP markers. Crop Sci. 37:537-543.

Lander, E., P. Green, J. Abrahamson, A. Barlow, M. Daly, S. Lincoln, and L. Newburg. 1987. MAPMAKER: aninteractive computer package for constructing primarygenetic linkage maps of experimental and natural populations. Genomics 1:174-181.

Lander, E.S., and D. Botstein. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199.

Lark, K.G., J. Orf, and L.M. Mansur. 1994. Epistatic expression of quantitative trait loci (QTL) in soybean [*Glycine max* (L.) Merr.] determined by QTL association with RFLP alleles. Theor. Appl. Genet. 88:486-489.

Lebowitz, R.J., M. Soller, and J.S. Beckmann. 1987. Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor. Appl. Genet. 73:556-562.

Lee, S.H., M.A. Bailey, M.A.R. Mian, T.E. Carter, 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.

Lincoln, S., M. Daly, and E. Lander. 1993. Constructing genetic maps with MAPMAKER/EXP version 3.0: a tutorial and reference manual:pp 97.

Mansur, L.M., J. Orf, and K.G. Lark. 1993a. Determining the linkage of quantitative trait loci to RFLP markers using extreme phenotypes of recombinant inbreds of soybean (*Glycine max* L. Merr.). Theor. Appl. Genet. 86:914-918.

Mansur, L.M., K.G. Lark, H. Kross, and A. Oliveira. 1993b. Interval mapping of quantitative trait loci for reproductive, morphological, and seed traits of soybean (Glycine max L.). Theor. Appl. Genet. 86:907-913.

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.

Martin, J.H., R.P. Waldren, and D.L. Stamp. 2006. Principles of field crop production. 4 ed. Pearson Education, Inc., Upper Saddle River, NJ.

Morgante, M., and A.M. Olivieri. 1993. PCR-amplified microsatellites as markers in plant genetics. Plant J. 3:175-182.

Muranty, H., and B. Goffinet. 1997. Selective genotyping for location and estimation of the effect of a quantitative trait locus. Biometrics 53: 629-643.

Narvel, J.M., W.R. Fehr, W.-C. Chu, D. Grant, and R.C. Shoemaker. 2000. Simple sequence repeat diversity among soybean plant introductions and elite genotypes. Crop Sci. 40:1452-1458.

Navabi, A., D. Mather, J. Bernier, D. Spaner, and G. Atlin. 2009. QTL detection with bidirectional and unidirectional selective genotyping: marker-based and trait-based analyses. Theor. Appl. Genet. 118:347-358.

Ngwako, S. 2008. Mapping quantitative trait loci using the marker regression and the interval mapping methods. Pakistan J. of Bio. Sci. 11(4): 553-558.

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.

Ooijen, J.W. 1992. Accuracy of mapping quantitative trait loci in autogamous species. Theor. Appl. Genet. 84:803-811.

Openshaw, S.J., and H.H. Hadley. 1984. Selection indexes to modify protein concentration of soybean seeds. Crop Sci. 24:1-4.

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

Pantalone, V.R., J.W. Burton, and T.E. Carter, Jr. 1996. Soybean fibrous root heritability and genotypic correlations with agronomic and seed quality traits. Crop Sci. 36:1120-1125.

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.

Paterson, A.H., S.D. Tanksley, and M.E. Sorrells. 1991 DNA markers in plant improvement. Adv. Agron. 46:39-90.

Piper, E.L., and K.J. Boote. 1999. Temperature and cultivar effects on soybean seed oil and protein concentrations. J. Am. Oil Chem. Soc. 76:1233-1241.

Pipolo, A.E., T.R. Sinclair, and G.M.S. Camara. 2004. Effects of temperature on oil and protein concentration in soybean seeds cultured *in vitro*. Ann. Appl. Biol. 144:71-76.

Podlich, D.W., C.R. Winkler, and M. Cooper. 2004. Mapping as you go: an effective approach for marker-assisted selection of complex traits. Crop Sci. 44:1560-1571.

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’×‘Essex’ population. Theor. Appl. Genet. 98:356-364.

Ritchie, R.A. 2003. High-protein plant introductions: Selective genotyping to detect soybean protein QTL. M.S. thesis, Univ. of Nebraska, Lincoln.

Saghai-Maroof, M.A., K.M. Soliman, R.A. Jorgensen, and R.W. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance,

chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. (USA) 81:8014-8018.

Sebern, N.A., and J.W. Lambert. 1984. Effect of stratification for percent protein in two soybean populations. Crop Sci. 24:225-228.

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.

Sen, S., F. Johannes, and K.W. Broman. 2009. Selective genotyping and phenotyping strategies in a complex trait context. Genetics 181:1613-1626.

Shannon, J.G., J.R. Wilcox, and A.H. Probst. 1972. Estimated gains from selection for protein and yield in the F4 generation of six soybean populations. Crop Sci. 12:824-826.

Shoemaker, R.C., and T.C. Olson. 1993. Molecular linkage map of soybean (*Glycine max* L. Merr.), p. 6131-6138, *In* S. J. O'Brien, ed. Genetic maps: locus maps of complex genomes. Cold Spring Harbor Laboratory Press, NY.

Shoemaker, R.C., and J.E. Specht. 1995. Integration of the soybean molecular and classical genetic linkage groups. Crop Sci. 35:436-446.

Shoemaker, R.C., R.D. Guffy, L.L. Lorenzen, and J.E. Specht. 1992. Molecular genetic mapping of soybean: Map utilization. Crop Sci. 32:1091-1098.

Simpson, A.M., Jr., and J.R. Wilcox. 1983. Genetic and phenotypic associations of agronomic characteristics in four high protein soybean populations. Crop Sci. 23:1077-1081.

Soares, T.C.B., P.I.V. Good-God, F.D.d. Miranda, J.B. Soares, I. Schuster, N.D. Piovesan, E.G.d. Barros, and M.A. Moreira. 2008. QTL mapping for protein content in soybean cultivated in two tropical environments. Pesquisa Agropecuária Brasileira 43:1533-1541.

Soller, M., T. Brody, and A. Genizi. 1976. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor Appl. Genet. 47:35-39.

Song, Q.J., L.F. Marek, R.C. Shoemaker, K.G. Lark, V.C. Concibido, X. Delannay, J.E. Specht, and P.B. Cregan. 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. 109:122-128.

Song, Q.J., G. Jia, Y. Zhu, D. Grant, R.T. Nelson, E.-Y. Hwang, D.L. Hyten, and P.B. Cregan. 2010. Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCSOYSSR_1.0) in soybean. Crop Sci. 50:1950-1960.

Soybase. 2010. SoyBase– a genome database for Glycine. Administered online by USDA-ARS and Iowa State University. http://soybeanbreederstoolbox.org/ (Verified 20 August 2010).

Soystats. 2010. Soy Stats: A reference guide to important soybean facts & figures. http://www.soystats.com/2010/Default-frames.htm (Verified 20 August 2010).

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.

Stuber, C.W., M.M. Goodman, and R.H. Moll. 1982. Improvement of yield and ear number resulting from selection at allozyme loci in a maize population. Crop Sci. 22:737-740.

Stuber, C.W., R.H. Moll, M.M. Goodman, H.E. Schaffer, and B.S. Weir. 1980. Allozyme frequency changes associated with selection for increased grain yield in maize (*Zea mays* L.). Genetics 95:225-236.

Sun, Y., J. Wang, J. Crouch, and Y. Xu. 2010. Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement. Mol. Breeding 26:493-511.

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. Breed. Sci. 53:133-140.

Tanksley, S.D. 1993. Mapping polygenes. Annu. Rev. of Genet. 27:205-233.

Thorne, J.C., and W.R. Fehr. 1970. Incorporation of high-protein, exotic germplasm into soybean populations by 2- and 3-way crosses. Crop Sci. 10:652-655.

Van, K., E.-Y. Hwang, M.Y. Kim, Y.-H. Kim, Y.-I. Cho, P.B. Cregan, and S.-H. Lee. 2004. Discovery of single nucleotide polymorphisms in soybean using primers designed from ESTs. Euphytica 139:147-157.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T.v.d. Lee, M. Hornes, A. Friters, J. Pot, J. Paleman, M. Kuiper, and M. Zabeau. 1995. AFLP: a new technique for DNA fingerprinting. Nucl. Acids Res. 23:4407-4414.

Waldroup, P.W. 2009. Soybean information center fact sheet: Soybean meal- demand. http://soymeal.org/pdf/soymealdemand.pdf (Verified 30 March 2009).

Wang, B., W. Guo, X. Zhu, Y. Wu, N. Huang, and T. Zhang. 2007. QTL Mapping of Yield and Yield Components for Elite Hybrid Derived-RILs in Upland Cotton. J. Genet. Genom. 34:35-45.

Wehrmann, V.K., W.R. Fehr, S.R. Cianzio, and J.F. Cavins. 1987. Transfer of high seed protein to high-yielding soybean cultivars. Crop Sci. 27:927-931.

Wilcox, J.R., and J.F. Cavins. 1995. Backcrossing high seed protein to a soybean cultivar. Crop Sci. 35:1036-1041.

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, R.F. 2004. Seed composition, p. 621-678, *In* H. R. Boerma and J. E. Specht, eds. Soybeans: Improvement, production, and uses, 3rd ed. ASA, CSSA, SSSA, Madison, WI.

Wolf, R.B., J.F. Cavins, R. Kleiman, and L.T. Black. 1982. Effect of temperature on soybean seed constituents: Oil, protein, moisture, fatty acids, amino acids and sugars. J. Am. Oil Chem. Soc. 59:230-232.

Xu, S. 2003. Theoretical Basis of the Beavis Effect. Genetics 165:2259-2268.

Xu, Y., and J.H. Crouch. 2008. Marker-assisted selection in plant breeding: From publications to practice. Crop Sci. 48:391-407.

Zeng, Z.B. 1994. Precision mapping of quantitative trait loci. Genetics 136:1457-1468.

Zhang, W.K., Y.J. Wang, G.Z. Luo, J.S. Zhang, C.Y. He, X.L. Wu, J.Y. Gai, and S.Y. Chen. 2004. QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. Theor. Appl. Genet. 108:1131-1139.

Zhi-Hong, Z., S. Li, L. Wei, C. Wei, and Z. Ying-Guo. 2005. A major QTL conferring cold tolerance at the early seedling stage using recombinant inbred lines of rice (Oryza sativa L.). Plant Sci. 168:527-534.

Zhu, T., L. Shi, J.J. Doyle, and P. Keim. 1995. A single nuclear locus phylogeny of soybean based on DNA sequence. Theor. Appl. Genet. 90:991-999.

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.

Appendix Table 1. QTLs reported in the literature for soybean seed protein, oil, and (if nearby yield). Data table from Richie (2003), and latest updated in 2009.

LG	Dist	Nearest marker	Trait	QTL Parent high allele	Progeny AA - BB	LOD or Prob	R^2	Parental mating	Pop type	Literature reference	Comments
A1	26.4	Satt382	Yield	BSR 101	101	<0.0001	0.12	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
A1	30.3	A329-2	Protein	Minsoy	6.6	<0.001	0.05	Minsoy X Noir 1	RIL	Mansur et al. (1996)	
A1	30.3	A329-2	Oil	Noir1	4.4	<0.001	0.05	Minsoy X Noir 1	RIL	Mansur et al. (1996)	In this Table, Mansur et al. QTLs are P < 0.001
A1	49.3	B030-2									
A1	51.8	A096-1									
A1	53.4	K400-1	Oil	not stated	na	0.02	0.14	A87-296011 X CX1039-99	F2:5	Brummer et al. (1997)	Note: AA/BB means not reported in Brummer et al
A1	56.2	T153-3									& only Brummer QTL sig. over Env reported here
A1	75.4	A975-1	Oil	not stated	na	0.009	0.11	C1763 X CX1159-49-1	F2:5	Brummer et al. (1997)	
A1	88.6	Satt174	Oil	Archer	na	4.9	0.10	Minsoy X Archer	RIL	Orf et al. (1999)	
A1	92.3	A104-1	Oil	not stated	na	0.02	0.19	M82-806 X HHP	F2:5	Brummer et al. (1997)	
A1	93.6	T155_1	Protein	Noir 1	na	4.2	0.15	Minsoy X Noir 1	RIL	Orf et al. (1999)	
A1	93.6	T155-1	Protein	Noir 1	6.6	<0.001	0.09	Minsoy X Noir 1	RIL	Mansur et al. (1996)	
A1	93.6	T155_1	Oil	Minsoy	na	3.4	0.13	Minsoy X Noir 1	RIL	Orf et al. (1999)	
A1	93.6	T155-1	Oil	Minsoy	5.4	<0.001	0.07	Minsoy X Noir 1	RIL	Mansur et al. (1996)	
A1	94.9	B170	Oil	Noir 1	4.5	4.6	0.05	Minsoy X Noir 1	RIL	Specht et al. (2001)	One of two QTL in Literat. with a +Pro+Oil allele,
A1	94.9	B170	Protein	Noir 1	6.8	5.0	0.04	Minsoy X Noir 1	RIL	Specht et al. (2001)	but the variation it accounted for was small.
A1	95.2	Satt225	Yield	LG82-8379	60	2.6	0.14	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
A2	48.8	I	Protein	Moshidou Gong 503	10.8	2.6	0.08	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
A2	48.8	I	Oil	Misuzudaizu	8.6	4.2	0.10	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
A2	50.4	T153-1	Oil	Noir 1	14.0	5.5	0.36	Minsoy X Noir 1	F5	Mansur et al. (1993)	The I / i^i / i (hilum/seedcoat color) locus near here
A2	67.3	A111	Oil	Misuzudaizu	11.0	2.9	0.09	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
A2	92.3	A104	Protein	Misuzudaizu	13.4	3.3	0.08	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
A2	132.3	A505-1	Protein	not stated	na	0.01	0.11	C1763 X CX1039-99	F2:5	Brummer et al. (1997)	These QTLs may thus be NIR artefacts.
A2	132.3	A505-1	Oil	not stated	na	0.03	0.09	C1763 X CX1039-99	F2:5	Brummer et al. (1997)	Also another grey sd ct locus near Lf1 (5-lf) here
A2	145.6	Satt409	Protein	LG82-8379	4.0	3.6	0.06	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	

LG indicates linkage group; **Dist** indicates approx distance (cM) from top of LG using linkage data of Cregan et al., 1999); **QTL high allele** identifies **parent** whose allele had **greater** trait value; **Differ AA - BB** is the difference between population AA and BB means (given as 10g/kg for protein and oil, or kg/ha for yield); **LOD** or **Prob** indicates the statistical significance of the QTL, with **Lod** in integers, but **Prob** in fractions; **R2** denotes the fraction of the variation controlled by the QTL; **Pop type** indicates the type of population evaluated.

Bold bordering means QTL allele has inverse pleiotropic effects on protein and oil (i.e., +Pro & -Oil). Light blue shading indicates otherwise.

Appendix Table 1. (Cont.)

LG	Dist	Nearest marker	Trait	QTL Parent high allele	Progeny AA - BB	LOD or Prob	R^2	Parental mating	Pop type	Literature reference	Comments
B1	25.2	A702-1	Protein	not stated	na	0.008	0.39	C1763 X CX1159-49-1	F2:5	Brummer et al. (1997)	This locus reported with A109 below
B1	29.2	A109-1	Protein	not stated	na	0.008	0.39	C1763 X CX1159-49-1	F2:5	Brummer et al. (1997)	See also A702 above
B1	29.2	A109-1	Protein	not stated	na	0.0001	0.08	McCall X PI 445815	F2:5	Brummer et al. (1997)	
B1	29.2	A109-1	Oil	not stated	na	0.03	0.31	C1763 X CX1159-49-1	F2:5	Brummer et al. (1997)	
B1	36.5	Satt251	Protein	Essex	10.0	<0.05	0.03	Essex X Williams	F4:6	Chapman et al. (2003)	
B1	156.8	GmKF082c-GmKF168b	Yield	Kefeng No. 1	226	4.4	0.09	Kefeng No. 1 X Nannong 1138-2	RIL	Zhang et al. (2004)	
B1	167.0	GmKF168b-Gmpti_D	Yield	Kefeng No. 1	238	5.4	0.10	Kefeng No. 1 X Nannong 1138-2	RIL	Zhang et al. (2004)	
B2	29.2	A352-1	Protein	PI 416937	6.0	>2.5	0.01	Young X PI 416937	F4	Lee et al. (1996)	Not confirmed by Fasoula et al. (2004)
B2	43.6	B142-1	Protein	PI 416937	5.0	>2.5	0.01	Young X PI 416937	F4	Lee et al. (1996)	Not confirmed by Fasoula et al. (2004)
B2	55.2	Satt168	Protein	LG82-8379	2.0	2.7	0.10	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
B2	55.2	Satt168	Yield	LG82-8379	67	6.7	0.16	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
B2	72.1	Satt020	Oil	Ma. Belle	2.0	0.0013	0.03	Ma. Belle X Proto	F2	Csanadi et al. (2001)	
B2	104.7	A519	Oil	Moshidou Gong 503	10.8	2.2	0.09	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
B2	104.7	A519	Oil	Moshidou Gong 503	7.6	2.0	0.08	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
B2	124.4	A953_1H-Satt560	Protein	Nannong 1138-2	12	3.5	0.12	Kefeng No. 1 X Nannong 1138-2	RIL	Zhang et al. (2004)	
C1	10.3	SOYGPATR	Protein	Noir 1	na	3.7	0.12	Minsoy X Noir 1	RIL	Orf et al. (1999)	
C1	10.3	SOYGPATR	Oil	Minsoy	na	3.3	0.11	Minsoy X Noir 1	RIL	Orf et al. (1999)	
C1	10.3	SOYGPATR	Oil	Archer	na	3.4	0.07	Minsoy X Archer	RIL	Orf et al. (1999)	
C1	21.0	A463-1	Protein	PI 416937	4.0	>2.5	0.07	Young X PI 416937	F4	Lee et al. (1996)	Not confirmed by Fasoula et al. (2004)
C1	33.3	K001-1	Oil	not stated	na	??	0.11	Minsoy X Noir 1	F5	Mansur et al. (1993)	
C1	33.3	K001-1	Oil	Noir 1	8.2	4.1	0.07	Minsoy X Noir 1	RIL	Specht et al. (2001)	
C1	65.1	Satt578	Protein	Minsoy	na	6.4	0.12	Minsoy X Archer	RIL	Orf et al. (1999)	
C1	90.7	A063-1	Protein	not stated	na	0.0004	0.17	A87-296011 X CX1039-99	F2:5	Brummer et al. (1997)	**N/n (norm/abnorm hilum) locus near here**
C1	90.7	A063-1	Protein	not stated	na	0.009	0.12	M84-492 X Sturdy	F2:5	Brummer et al. (1997)	**This mutant has a whitish hilum - NIR artefact?**
C1	90.7	A063-1	Oil	PI 97100	3.0	>2.5	0.13	PI 97100 X Coker 237	F2	Lee et al. (1996)	Confirmed by Fasoula et al. (2004)
C1	90.7	A063-1	Oil	PI97100	4.0	0.0011	0.08	PI 97100 X Coker 237	F2:4	Fasoula et al. (2004)	Confirmed QTL designated as cqOil-001
C1	123.8	Satt338	Protein	LG82-8379	5.0	<0.0001	0.16	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
C1	123.8	Satt338	Oil	BSR 101	2.0	<0.0001	0.05	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	

Appendix Table 1. (Cont.)

LG	Dist	Nearest marker	Trait	QTL Parent high allele	Progeny AA - BB	LOD or Prob	R^2	Parental mating	Pop type	Literature reference	Comments
C2	38.1	Satt432	Oil	Archer	na	3.3	0.11	Noir1 X Archer	RIL	Orf et al. (1999)	
C2	40.3	Satt281	Protein	Moshidou Gong 503	11.0	3.7	0.06	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
C2	40.3	Satt281	Yield	Minsoy	196	5.3	0.04	Minsoy X Noir 1	RIL	Specht et al. (2001)	
C2	56.6	A338-1	Protein	PI 416937	5.0	>2.5	0.10	Young X PI 416937	F4	Lee et al. (1996)	Not rechecked by Fasoula (1993?)
C2	97.2	L148-1	Oil	not stated	na		0.09	A87-296011 X CX1039-99	F2:5	Brummer et al. (1997)	This locus reported with A538-1 below
C2	98.1	Satt363	Protein	BSR 101	3.0	<0.0001	0.05	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
C2	98.1	Satt363	Yield	LG82-8379	60	<0.0001	0.10	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
C2	107.6	Satt277	Yield	Noir 1	na	6.1	0.11	Noir1 X Archer	RIL	Orf et al. (1999)	
C2	112.2	Satt557	Yield	Kefeng No. 1	262	6.9	0.13	Kefeng No. 1 X Nannong 1138-2	RIL	Zhang et al. (2004)	
C2	113.4	Satt489	Yield	Minsoy	na	3.3	0.06	Minsoy X Noir 1	RIL	Orf et al. (1999)	E1/e1 (late/early matutity) locus near here
C2	114.8	A109-2	Yield	Minsoy	474	2.4	0.24	Minsoy X Noir 1	F5	Mansur et al. (1993)	
C2	117.9	Satt079	Yield	Minsoy	218	<0.001	0.07	Minsoy X Noir 1	RIL	Mansur et al. (1996)	
C2	122.0	Sct_028	Protein	Proto	4.0	0.0069	0.07	Ma. Belle X Proto	F2	Csanadi et al. (2001)	
C2	123.4	A538-1	Oil	not stated	na	0.03	0.09	A87-296011 X CX1039-99	F2:5	Brummer et al. (1997)	See L148-1 above
C2		Satt205	Yield	Minsoy	378	17.3	0.13	Minsoy X Noir 1	RIL	Specht et al. (2001)	
C2	?-107.59	K11_3T-Satt277	Yield	Kefeng No. 1	246	5.9	0.12	Kefeng No. 1 X Nannong 1138-2	RIL	Zhang et al. (2004)	
C2	107.59-117.77	Satt277-Satt460	Oil	Williams	5.0	3.7	0.09	Essex X Williams	RIL	Hyten et al. (2004)	
C2	107.59-126.24	Satt277-Satt202	Protein	Williams	11.0	9.8	0.28	Essex X Williams	RIL	Hyten et al. (2004)	
C2	113.42-?	Satt319-K11_3T	Yield	Kefeng No. 1	258	6.0	0.12	Kefeng No. 1 X Nannong 1138-2	RIL	Zhang et al. (2004)	
C2	44.66-40.30	Satt422-Satt281	Protein	BARC-8	22.0?	15.44	0.14	BARC-8 X Garimpo	RIL	Soares et al. (2008)	
D1a	6.4	A398-1	Protein	not stated	na	0.009	0.28	LN83-2356 X PI 360843	F2:5	Brummer et al. (1997)	
D1a	40.8	A691-1	Protein	not stated	na	0.02	0.10	C1763 X CX1039-99	F2:5	Brummer et al. (1997)	
D1a	69.9	Satt468	Oil	Noir 1	6.2	4.7	0.09	Minsoy X Noir 1	RIL	Specht et al. (2001)	
D1a	77.5	Satt077	Oil	Ma. Belle	4.4	0.0009	0.05	Ma. Belle X Proto	F2	Csanadi et al. (2001)	
D1a	17.52-56.20	Satt184-Satt179	Oil	Williams	5.0	3.3	0.08	Essex X Williams	RIL	Hyten et al. (2004)	
D1b	27.1	Wc	Oil	Minsoy	3.6	2.62	0.07	Minsoy X Noir 1	RIL	Specht et al. (2001)	
D1b	37.1	Satt157	Protein	BSR 101	4.0	4.3	0.14	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
D1b	37.1	Satt157	Oil	LG82-8379	3.0	5.1	0.10	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
D1b	116.4	Satt274	Oil	N87-984-16	37.0	3.0	0.12	N87-984-16 X TN93-99	RIL	Panthee et al. (2005)	
D1b	na	Wp	Protein	LN89-5322-1	na	na	na	RM55 X LN89-5322-2		Hegstad et al. (2000)	Not a conventional QTL, but wp affects protein
										Stephens et al. (1993)	Note: Some D1b RFLPs are LG-I homoeologues
D1b+W		A481V-A725_3V	Oil	Nannong 1138-2	6.0	2.5	0.07	Kefeng No. 1 X Nannong 1138-2	RIL	Zhang et al. (2004)	

Appendix Table 1. (Cont.)

LG	Dist	Nearest marker	Trait	QTL Parent high allele	Progeny AA - BB	LOD or Prob	R^2	Parental mating	Pop type	Literature reference	Comments
D2	10.0	A064	Protein	Minsoy	8.0	6.1	0.11	Minsoy X Noir 1	RIL	Specht et al. (2001)	
D2	47.7	Satt002	Yield	Archer	na	3.9	0.08	Noir 1 X Archer	RIL	Orf et al. (1999)	
D2	73.5	K258-1	Oil	Young	3.0	>2.5	0.09	Young X PI 416937	F4	Lee et al. (1996)	These 3 not confirmed by Fasoula et al. (2004)
D2	93.7	Satt301	Protein	Moshidou Gong 503	10.4	3.1	0.06	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
D2	105.5	Satt186	Yield	LG82-8379	67	3.2	0.21	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
D2	107.5	Satt310	Protein	Moshidou Gong 503	13.2	3.2	0.07	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
D2	?	cr142-1	Oil	Young	3.0	>2.5	0.13	Young X PI 416937	F4	Lee et al. (1996)	Linked to K258 above
D2	?	cr326-1n	Oil	Young	2.0	>2.5	0.09	Young X PI 416937	F4	Lee et al. (1996)	Linked to K258 above
D2	24.52-57.07	Satt458-Satt154	Oil	Williams	5.0	3.0	0.08	Essex X Williams	RIL	Hyten et al. (2004)	
E	6.3	SAC7-1	Protein	PI 468916	17.0	0.003	0.24	A81-356022 XPI 468916	F2	Diers et al. (1992)	**Strong evidence of a pro/oil QTL in this 20-40 cM**
E	6.3	SAC7-1	Oil	A81-356022	17.0	0.0001	0.43	A81-356022 XPI 468916	F2	Diers et al. (1992)	**region of LG-E**
E	11.0	A242-2	Protein	PI 468916	12.0	0.004	0.19	A81-356022 XPI 468916	F2	Diers et al. (1992)	
E	11.0	A242-2	Oil	A81-356022	14.0	0.0001	0.39	A81-356022 XPI 468916	F2	Diers et al. (1992)	
E	13.6	Pb	Oil	A81-356022	13.0	0.001	0.27	A81-356022 XPI 468916	F2	Diers et al. (1992)	
E	19.3	Satt384	Protein	Moshidou Gong 503	12.6	3.6	0.07	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
E	19.3	Satt384	Protein	Moshidou Gong 503	12.8	4.7	0.08	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
E	21.1	A053-1?	Protein	PI 468916	14.0	0.01	0.16	A81-356022 XPI 468916	F2	Diers et al. (1992)	
E	21.1	A053-1?	Oil	A81-356022	14.0	0.006	0.32	A81-356022 XPI 468916	F2	Diers et al. (1992)	
E	26.0	A517-1	Protein	Young	5.0	>2.5	0.07	Young X PI 416937	F4	Lee et al. (1996)	
E	27.0	A458-1	Protein	not stated	na	0.02	0.11	McCall X PI 445815	F2:5	Brummer et al. (1997)	Reported with B174 above
E	28.3	K229-1	Oil	A81-356022	11.0	0.001	0.22	A81-356022 XPI 468916	F2	Diers et al. (1992)	
E	30.9	B174-1	Protein	not stated	na	0.02	0.11	McCall X PI 445815	F2:5	Brummer et al. (1997)	Linked to A458 above
E	30.9	A454-1	Protein	PI 97100	7.0	>2.5	0.09	Young X PI 416937	F2	Lee et al. (1996)	Confirmed by Fasoula et al. (2004)
E	30.9	A454-1	Protein	PI 97100	11	0.0001	0.12	PI 97100 X Coker 237	F2:4	Fasoula et al. (2004)	Confirmed QTL designated as cqProt-001
E	30.9	A454-1	Oil	A81-356022	10.0	0.0008	0.23	A81-356022 XPI 468916	F2	Diers et al. (1992)	
E	34.6	A203-1	Oil	A81-356022	10.0	0.006	0.18	A81-356022 XPI 468916	F2	Diers et al. (1992)	
E		A069-3	Oil	Young	2.0	>2.5	0.07	Young X PI 416937	F4	Lee et al. (1996)	
E	19.30-8.67	Satt384-Sat_112	Protein	BARC-8	23.8?	9.68	0.10	BARC-8 X Garimpo	RIL	Soares et al. (2008)	

Appendix Table 1. (Cont.)

LG	Dist	Nearest marker	Trait	QTL Parent high allele	Progeny AA - BB	LOD or Prob	R^2	Parental mating	Pop type	Literature reference	Comments
F	47.6	K002-1	Protein	not stated	na	0.03	0.09	A87-296011 X CX1039-99	F2:5	Brummer et al. (1997)	W1/w1 (purple/white flower) locus near here
F	71.4	Satt510	Protein	LG82-8379	5.0	<0.0001	0.16	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	for the +/- presence of tri-hydroxylated B-ring
F	71.4	Satt510	Oil	Noir 1	4.8	3.3	0.06	Minsoy X Noir 1	RIL	Specht et al. (2001)	anthocyanins, the black pigment in Impblk seed coats
F	80.8	A245-1	Protein	PI 468916	10.0	0.01	0.12	A81-356022 XPI 468916	F2	Diers et al. (1992)	
F	102.1	Satt144	Yield	Archer	na	7.0	0.13	Noir 1 X Archer	RIL	Orf et al. (1999)	
F	105.8	B148-1	Protein	Essex	40.0	0.0001	0.17	Peking X Essex	F2.3	Qui et al. (1999)	
F	114.1	A566-2	Protein	PI 97100	8.0	>2.5	0.14	PI 97100 X Coker 237	F2	Lee et al. (1996)	
F	114.1	A566-2	Oil	Coker 237	4.0	>2.5	0.10	PI 97100 X Coker 237	F2	Lee et al. (1996)	
F	142.4	Sat_074	Yield	Noir 1	175	3.4	0.02	Minsoy X Noir 1	RIL	Specht et al. (2001)	
F	77.70-102.08	Satt335-Satt144	Protein	Essex	8.2	4.4	0.18	Essex X Williams	RIL	Hyten et al. (2004)	
F		OP-AN09-OP-AC02	Protein	Garimpo	16.4?	7.76	0.07	BARC-8 X Garimpo	RIL	Soares et al. (2008)	
G	12.7	Satt570	Protein	N87-984-16	6.2	3.5	0.20	N87-984-16 X TN93-99	RIL	Panthee et al. (2005)	
G	43.4	Satt394	Yield	LG82-8379	67	6.0	0.28	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
G	65.6	A584-1	Oil	not stated	na	0.01	0.19	C1763 X CX1039-99	F2:5	Brummer et al. (1997)	
G	65.6	A584-1	Oil	not stated	na	0.009	0.11	M84-492 X Sturdy	F2:5	Brummer et al. (1997)	
G	67.5	A816-1	Protein	not stated	na	0.005	0.12	A87-296011 X CX1039-99	F2:5	Brummer et al. (1997)	
G	67.5	A816-1	Oil	not stated	na	0.007	0.11	A87-296011 X CX1039-99	F2:5	Brummer et al. (1997)	Linked to A374 below
G	96.6	Satt191	Protein	Minsoy	8.0	4.7	0.11	Minsoy X Noir 1	RIL	Specht et al. (2001)	
G	96.6	Satt191	Yield	BSR 101	81	<0.0001	0.27	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
G	97.2	A235-1	Protein	not stated	na	0.04	0.08	C1763 X CX1039-99	F2:5	Brummer et al. (1997)	
G	97.2	A235-1	Oil	Coker 237	1.0	>2.5	0.15	PI 97100 X Coker 237	F2	Lee et al. (1996)	See also pink-highlighted ones immediately below
G	97.7	L002-1	Oil	Coker 237	1.0	>2.5	0.14	PI 97100 X Coker 237	F2	Lee et al. (1996)	Linked to pink-highlighted one below
G	97.7	L002-1	Oil	Minsoy	6.0	5.3	0.08	Minsoy X Noir 1	RIL	Specht et al. (2001)	
G	99.3	L154-2	Oil	Coker 237	0.0	>2.5	0.17	PI 97100 X Coker 237	F2	Lee et al. (1996)	Linked to pink-highlighted one above
G	109.5	A378	Protein	Moshidou Gong 503	8.0	2.1	0.03	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
G		(B216-?)	Protein	not stated	na	0.003	0.16	C1763 X CX1159-49-1	F2:5	Brummer et al. (1997)	
G		A374-3	Oil	not stated	na	0.007	0.11	A87-296011 X CX1039-99	F2:5	Brummer et al. (1997)	Reported with A816 above
G		(B216-?)	Oil	not stated	na	0.004	0.15	M81-382 X PI 423949	F2:5	Brummer et al. (1997)	
G	62.16-52.94	Satt199-Satt594	Protein	Garimpo	19.0?	9.56	0.10	BARC-8 X Garimpo	RIL	Soares et al. (2008)	

Appendix Table 1. (Cont.)

LG	Dist	Nearest marker	Trait	QTL Parent high allele	Progeny AA - BB	LOD or Prob	R^2	Parental mating	Pop type	Literature reference	Comments
H	33.2	A069-1	Protein	not stated	na	0.05	0.07	C1763 X CX1159-49-1	F2:5	Brummer et al. (1997)	
H	33.2	A069-1	Oil	not stated	na	0.0003	0.18	C1763 X CX1159-49-1	F2:5	Brummer et al. (1997)	
H	69.1	Satt314	Yield	Minsoy	168	3.76	0.03	Minsoy X Noir 1	RIL	Specht et al. (2001)	
H	86.5	Satt142	Protein	BSR 101	3.0	2.8	0.03	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
H	86.5	Satt142	Yield	LG82-8379	148	2.7	0.35	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
H	89.5	Satt317	Oil	TN93-99	32.4	2.6	0.09	N87-984-16 X TN93-99	RIL	Panthee et al. (2005)	
H	111.4	A566-2	Protein	PI 97.100	7.0	NA	0.09	PI 97.100 X Coker 237	F2	Lee et al. (1996)	Not confirmed by Fasoula et al. (2004)
H	111.4	A566-2	Oil	Coker	4.0	NA	0.10	PI 97.100 X Coker 237	F2	Lee et al. (1996)	Confirmed by Fasoula et al. (2004)
H	111.4	A566-2	Oil	Coker 237	2.0	0.0008	0.08	PI 97100 X Coker 237	F2:4	Fasoula et al. (2004)	Confirmed QTL designated as cqOil-002
H	124.1	B072-1	Protein	Essex	50.0	0.0018	0.32	Peking X Essex	F2.3	Qui et al. (1999)	**One of two QTL in Lit. with a lrg effect +pro+oil allele!!!**
H	124.1	B072-1	Oil	Essex	20.0	0.002	0.21	Peking X Essex	F2.3	Qui et al. (1999)	**(However, A & B allele coding of pro & oil QTLs in the paper is not consistent with +pro+oil: so maybe not!)**
I	22.8	Satt562	Oil	Proto	5.5	0.004	0.06	Ma. Belle X Proto	F2	Csanadi et al. (2001)	
I	32.4	A688-1	Protein	PI 468916	18.0	0.001	0.25	A81-356022 XPI 468916	F2	Diers et al. (1992)	
I	32.4	A144-1	Protein	PI 468916	18.0	0.0007	0.24	A81-356022 XPI 468916	F2	Diers et al. (1992)	
I	32.4	A144-1	Protein	PI 468916	20.0	0.0001	0.44	BC3 Line X Parker	F3.4	Sebolt et al. (2000)	341kg/ha yield dbl-allelic diff in 1998 for A515-1
I	32.4	A144-1	Protein	PI 468916	19.0	0.0001	0.41	BC3 Line X Kenwood	F3.4	Sebolt et al. (2000)	210kg/ha yield dbl-allelic diff in1998 for A688-1
I	32.4	A144-1	Protein	not stated	na	0.0002	0.28	M82-806 X HHP	F2:5	Brummer et al. (1997)	Linked to A407 below
I	32.4	A144-1	Oil	A81-356022	9.5	0.0001	0.39	A81-356022 X PI 468916	BC3F4.6	Sebolt et al. (2000)	
I	32.4	A144-1	Oil	Parker	9.0	0.0008	0.15	BC3 Line X Parker	F3.4	Sebolt et al. (2000)	
I	32.4	A144-1	Oil	Kenwood	9.0	0.0001	0.23	BC3 Line X Kenwood	F3.4	Sebolt et al. (2000)	
I	35.2	BLT002	Oil	Minsoy	6.0	6.1	0.10	Minsoy X Noir 1	RIL	Specht et al. (2001)	
I	35.4	Satt127	Yield	A81-356022	212	0.0005	0.26	A81-356022 X PI 468916 F2	BC3F4.6	Sebolt et al. (2000)	Only QTLs sig. over Env in Sebolt reported here
I	35.4	Satt127	Oil	Misuzudaizu	10.6	3.4	0.09	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
I	35.4	Satt127	Protein	PI 468916	21.0	0.0001	0.65	A81-356022 X PI 468916 F2	BC3F4.6	Sebolt et al. (2000)	
I	36.9	Satt239	Protein	Moshidou Gong 503	22.2	9.3	0.20	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
I	36.9	Satt239	Protein	Moshidou Gong 503	24.4	16.0	0.30	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
I	36.9	Satt239	Oil	Misuzudaizu	7.0	3.0	0.07	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
I	38.1	K011-1	Protein	PI 468916	24.0	0.0001	0.42	A81-356022 X PI 468916	F2	Diers et al. (1992)	
I	38.1	K011-1	Oil	A81-356022	4.0	0.0002	0.27	A81-356022 X PI 468916	F2	Diers et al. (1992)	
I	39.4	A407-1	Protein	PI 468916	22.0	0.0001	0.39	A81-356022 X PI 468916	F2	Diers et al. (1992)	
I	39.4	A407-1	Protein	not stated	na	0.0002	0.28	M82-806 X HHP	F2:5	Brummer et al. (1997)	Reported with A144 above
I	39.4	A407-1	Oil	A81-356022	4.0	0.0005	0.28	A81-356022 XPI 468916	F2	Diers et al. (1992)	

Appendix Table 1. (Cont.)

LG	Dist	Nearest marker	Trait	QTL Parent high allele	Progeny AA - BB	LOD or Prob	R^2	Parental mating	Pop type	Literature reference	Comments
I	52.0	A955-1	Yield	Minsoy	144	2.9	0.03	Minsoy X Noir 1	RIL	Specht et al. (2001)	
I	54.8	L048-1	Yield	Asgrow A3733	na	??	0.20	Minsoy X Noir 1	F5	Mansur et al. (1993)	
I	112.7	Satt440	Yield	BSR 101	114	<0.0001	0.15	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
I	31.94-46.22	Satt614-Satt354	Protein	PI 468916	12.0-17.0	<.001		A81-356022 X PI 468916	BC4	Nichols et al. (2006)	
I	31.94-46.22	Satt614-Satt354	Yield	A81-356022	179-309	<.05-.001		A81-356022 X PI 468916	BC4	Nichols et al. (2006)	
I	31.94-46.22	Satt614-Satt354	Oil	A81-356022	8.0-11.0	<.001		A81-356022 X PI 468916	BC4	Nichols et al. (2006)	
I	36.4-36.9	Satt496/239	Protein	PI 437.088A	19.4	16.9	0.45	A81-356022 X PI 468916	RIL	Chung et al. (2003)	Strong evidence of this QTL in this finemap region!!!
I	36.4-36.9	Satt496/239	Protein	Asg A3733	11.4	10.5	0.28	Minsoy X Noir 1	RIL	Chung et al. (2003)	
I	36.4-36.9	Satt496/239	Oil	Asgrow A3733	222	13.31	0.19	Minsoy X Noir 1	RIL	Chung et al. (2003)	
I	36.94-?	Satt239-ACG9b	Protein	PI 468916	12.0-21.0	<.001		A81-356022 X PI 468916	BC5	Nichols et al. (2006)	
I	36.94-?	Satt239-ACG9b	Yield	A81-356022	319	0.05-0.001		A81-356022 X PI 468916	BC5	Nichols et al. (2006)	
I	36.94-?	Satt239-ACG9b	Oil	A81-356022	12.0-21.0	<.001		A81-356022 X PI 468916	BC5	Nichols et al. (2006)	
J	25.5	Satt285	Oil	LG82-8379	3.0	<0.0001	0.16	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
J	27.6	B166-1	Protein	PI 416937	5.0	>2.5	0.08	Young X PI 416937	F4	Lee et al. (1996)	
J	28.1	K384	Oil	Moshidou Gong 503	9.0	2.4	0.06	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
J	57.2	B122-1	Oil	Young	3.0	>2.5	0.07	Young X PI 416937	F4	Lee et al. (1996)	
J	67.8	Satt547	Yield	BSR 101	74	<0.0001	0.08	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
J	73.7	G815	Yield	Minsoy	142	2.7	0.02	Minsoy X Noir 1	RIL	Specht et al. (2001)	
J	78.57	Satt431	Protein	HP	NA	<0.0001	NA	Association Mapping Population	AMP	Jun et al. (2008)	

Appendix Table 1. (Cont.)

LG	Dist	Nearest marker	Trait	QTL Parent high allele	Progeny AA - BB	LOD or Prob	R^2	Parental mating	Pop type	Literature reference	Comments
K	28.7	A315-1	Oil	Noir 1	10.0	2.9	0.24	Minsoy X Noir 1	F5	Mansur et al. (1993)	
K	?	(Q043-1?)	Protein	Coker 237	6.0	>2.5	0.10	PI 97100 X Coker 237	F2	Lee et al. (1996)	These three RFLPs are linked to each other but
K	31.8	R051-3?	Protein	Coker 237	7.0	>2.5	0.10	PI 97100 X Coker 237	F2	Lee et al. (1996)	their genomic map positions on LG-K are not
K	133.1	A065-1	Protein	Coker 237	7.0	>2.5	0.11	PI 97100 X Coker 237	F2	Lee et al. (1996)	certain.
K	40.9	Satt178	Protein	Minsoy	5.2	3.4	0.03	Minsoy X Noir 1	RIL	Specht et al. (2001)	
K	43.4	Satt544	Protein	BSR 101	2.0	3.7	0.03	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
K	43.4	Satt544	Yield	LG82-8379	114	5.9	0.38	BSR 101 X LG82-8379	F5	Kabelka et al. (2004)	
K	49.5	Satt326	Yield	Noir 1	137	3.0	0.02	Minsoy X Noir 1	RIL	Specht et al. (2001)	
K	98.9	K387-1	Oil	not stated	na	0.002	0.16	C1763 X CX1039-99	F2:5	Brummer et al. (1997)	
K	104.8	Satt196	Protein	Ma. Belle	3.2	0.0009	0.05	Ma. Belle X Proto	F2	Csanadi et al. (2001)	**Authors have AA & BB coding reversed, Proto donated BB so should be higher in protein R/r locus (anthocyanin present/absent) hilum/seedcoat also located near here.**
K	104.8	Satt196	Oil	Proto	3.2	0.0004	0.07	Ma. Belle X Proto	F2	Csanadi et al. (2001)	
K		GM195	Protein	Moshidou Gong 503	13.0	2.9	0.08	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
K	1.8-30.28	Satt539-Satt102	Protein	Essex	10.8	4.3	0.24	Essex X Williams	RIL	Hyten et al. (2004)	
L	30.6	Satt398	Oil	PI 416937	5.0	0.0025	0.08	Young X PI 416937	F2:4	Fasoula et al. (2004)	linked to RFLP/A023-1
L	34.5	Satt313	Oil	PI 416937	4.0	0.0049	0.07	Young X PI 416937	F2:4	Fasoula et al. (2004)	linked to RFLP/A023-1
L	36.7	A023-1	Oil	Young	2.0	>2.5	0.07	Young X PI 416937	F4	Lee et al. (1996)	
L	56.1	Satt156	Protein	Moshidou Gong 503	12.0	2.2	0.06	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
L	66.5	Satt166	Protein	Archer	na	3.3	0.11	Noir 1 X Archer	RIL	Orf et al. (1999)	
L	86.6	G173-1	Oil	Noir 1	3.8	3.7	0.07	Minsoy X Noir 1	RIL	Specht et al. (2001)	
L	89.1	Dt1	Yield	Noir 1	285	11.5	0.09	Minsoy X Noir 1	RIL	Specht et al. (2001)	**Dt1/dt1 (indeterm/determ) stem habit locus**
L	92.0	Satt006	Protein	Minsoy	6.4	3.2	0.07	Minsoy X Noir 1	RIL	Specht et al. (2001)	**Very near Dt1/dt1 locus (pseudo pro/oil QTL !!!!)**
L	92.0	Satt006	Protein	Minsoy	8.3	<0.001	0.08	Minsoy X Noir 1	RIL	Mansur et al. (1996)	**Very near Dt1/dt1 locus (pseudo pro/oil QTL !!!!)**
L	92.0	Satt006	Oil	Noir 1	5.7	<0.001	0.09	Minsoy X Noir 1	RIL	Mansur et al. (1996)	**Very near Dt1/dt1 locus (pseudo pro/oil QTL !!!!)**
L	95.4	A489_1	Oil	Noir 1	na	6.1	0.19	Noir 1 X Archer	RIL	Orf et al. (1999)	
L	107.2	Satt373	Protein	Essex	12.0	<0.05	0.04	Essex X Williams	F4:6	Chapman et al. (2003)	It is nignificant w hen maturity is used as covariate
L		OP-AS07-OP-P09	Protein	BARC-8	18.0?	7.11	0.07	BARC-8 X Garimpo	RIL	Soares et al. (2008)	
L	66.51-89.13	Satt166-Dt1	Oil	Williams	5.0	3.4	0.08	Essex X Williams	RIL	Hyten et al. (2004)	
L	93.89-107.24	Satt229-Satt373	Oil	Williams	3.0	3.3	0.08	Essex X Williams	RIL	Hyten et al. (2004)	
M	7.8	Satt590	Yield	Kefeng No. 1	182	3.5	0.07	Kefeng No. 1 X Nannong 1138-2	RIL	Zhang et al. (2004)	
M	18.6	Satt150	Yield	Noir 1	na	11.0	0.19	Minsoy X Noir 1	RIL	Orf et al. (1999)	
M	18.6	Satt150	Yield	Noir 1	623	36.6	0.38	Minsoy X Noir 1	RIL	Specht et al. (2001)	
M	33.5	Satt567	Protein	Minsoy	17.0	12.8	0.33	Minsoy X Noir 1	RIL	Specht et al. (2001)	**Noir 1 En (late mat) locus near here**
M	33.5	Satt567	Yield	Noir 1	530	31.7	0.27	Minsoy X Noir 1	RIL	Specht et al. (2001)	**Minsoy has the en (early mat) allele**
M	39.0	R079_1	Protein	Minsoy	na	3.2	0.06	Minsoy X Archer	RIL	Orf et al. (1999)	
M	39.0	R079-1	Yield	Noir 1	302	<0.001	0.13	Minsoy X Noir 1	RIL	Mansur et al. (1996)	
M	58.5	A584-3	Yield	Noir 1	482	3.2	0.20	Minsoy X Noir 1	F5	Mansur et al. (1993)	

Appendix Table 1. (Cont.)

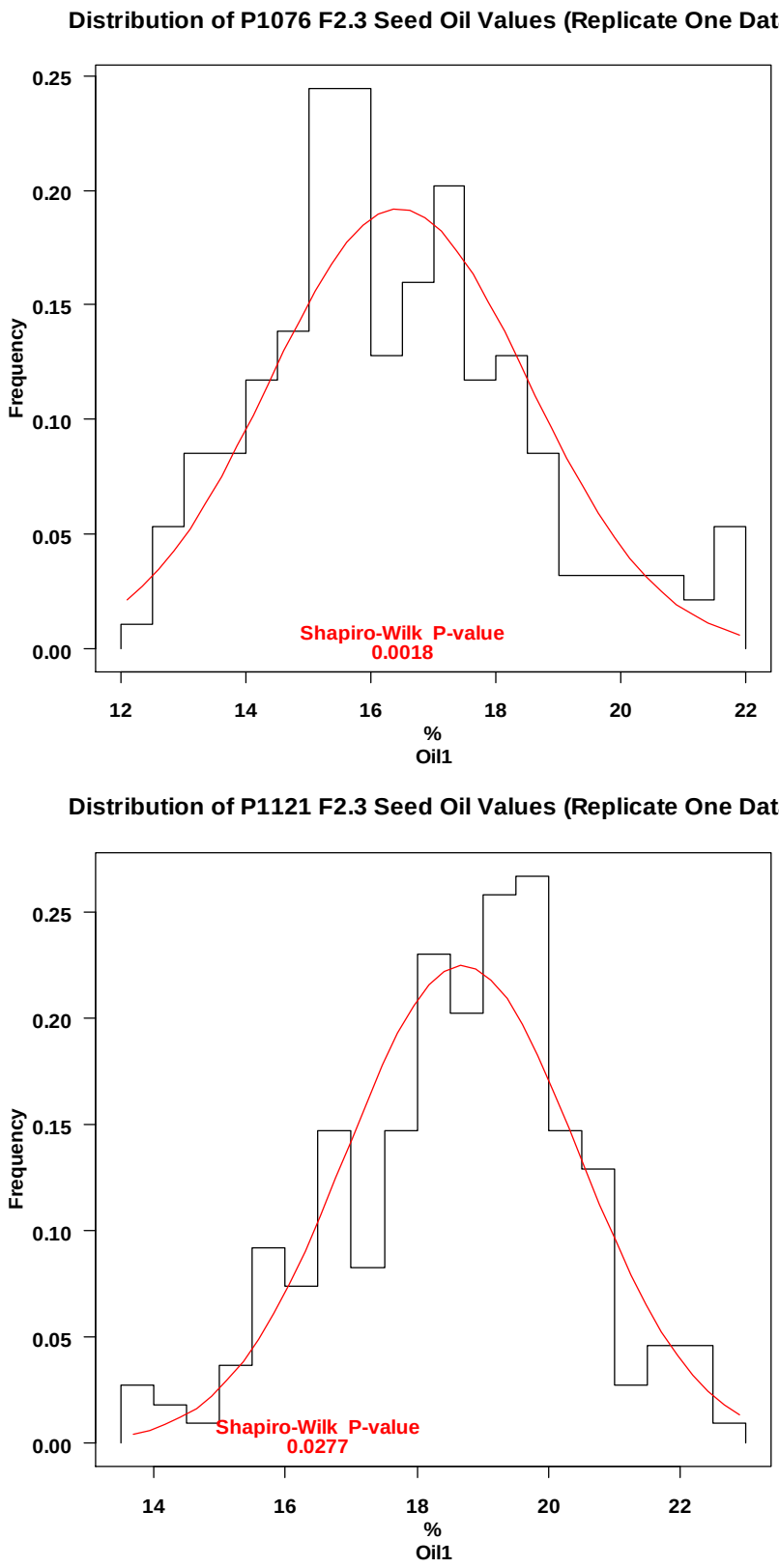
LG	Dist	Nearest marker	Trait	QTL Parent high allele	Progeny AA - BB	LOD or Prob	R^2	Parental mating	Pop type	Literature reference	Comments
M	95.45	Satt551	Protein	HP	NA	<0.0001	NA	Association Mapping Population	AMP	Jun et al. (2008)	
M	107.7	Satt250	Oil	Moshidou Gong 503	9.0	2.2	0.06	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
M	130.8	Satt308	Yield	LG82-8379	87	<0.0001	0.18	BSR101 X LG82-8379	F5	Kabelka et al. (2004)	
M	130.8	Satt308	Protein	LG82-8379	3.0	<0.0001	0.10	BSR101 X LG82-8379	F5	Kabelka et al. (2004)	
M	130.8	Satt308	Oil	Misuzudaizu	7.2	3.3	0.07	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
M		Satt567	Protein	Ma. Belle	3.9	0.0	0.07	Ma. Belle X Proto	F2	Csanadi et al. (2001)	
M	?-18.58	A60V-Satt150	Yield	Kefeng No. 1	222	3.8	0.10	Kefeng No. 1 X Nannong 1138-2	RIL	Zhang et al. (2004)	
M	35.85-50.10	Satt540-Satt463	Protein	Essex	7.0	3.0	0.13	Essex X Williams	RIL	Hyten et al. (2004)	
M	35.85-50.10	Satt540-Satt463	Oil	Williams	3.0	3.6	0.12	Essex X Williams	RIL	Hyten et al. (2004)	
N	9.7	A071-1?	Protein	PI 416937	5.0	>2.5	0.11	Young X PI 416937	F4	Lee et al. (1996)	Near the Rps1 & Rps7 Phytop root rot resistance locus
N	53.3	Satt387	Yield	BSR101	128	3.1	0.15	BSR101 X LG82-8379	F5	Kabelka et al. (2004)	
N	55.4	Rpg4	Yield	Minsoy	181	4.4	0.04	Minsoy X Noir 1	RIL	Specht et al. (2001)	
N	65.5	Satt521	Protein	Noir 1	5.8	2.3	0.05	Minsoy X Noir 1	RIL	Specht et al. (2001)	
N	75.9	Satt339	Protein	LG82-8379	5.0	4.4	0.13	BSR101 X LG82-8379	F5	Kabelka et al. (2004)	
N	75.9	Satt339	Yield	BSR101	67	3.2	0.12	BSR101 X LG82-8379	F5	Kabelka et al. (2004)	
N		Q026	Protein	Moshidou Gong 503	9.0	2.1	0.04	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
N	70.60-36.86?	Satt549-Satt084	Protein	BARC-8	16.2?	7.53	0.07	BARC-8 X Garimpo	RIL	Soares et al. (2008)	
O	5.4	Satt358	Protein	LG82-8379	3.0	3.0	0.09	BSR101 X LG82-8379	F5	Kabelka et al. (2004)	
O	5.4	Satt358	Yield	LG82-8379	47	3.0	0.14	BSR101 X LG82-8379	F5	Kabelka et al. (2004)	
O	49.7	Satt420	Yield	Minsoy	124	2.5	0.02	Minsoy X Noir 1	RIL	Specht et al. (2001)	
O	49.7	Satt420	Oil	N87-984-16	3.0	3.5	0.15	N87-984-16 X TN93-99	RIL	Panthee et al. (2005)	
O	54.2	Satt479	Oil	N87-984-16	2.8	3.1	0.12	N87-984-16 X TN93-99	RIL	Panthee et al. (2005)	
O	71.1	Satt478	Protein	Noir 1	8.0	3.0	0.06	Minsoy X Noir 1	RIL	Specht et al. (2001)	
O		GM214b	Oil	Moshidou Gong 503	7.2	2.3	0.07	Misuzudaizu X Moshidou Gong 503	RIL	Tajuddin et al. (2003)	
UNK		A132-4	Protein	Coker 237	6.0	0.0116	0.06	PI 97100 X Coker 237	F2:4	Fasoula et al. (2004)	Confirmed QTL designated as cqProt-002
UNK1	?	A199-3	Protein	PI 416937	7.0	>2.5	0.14	Young X PI 416937	F4	Lee et al. (1996)	not confirmed by Fasoula et al. (2004)
UNK2	?	A132-4	Protein	Coker 237	5.0	3.4	0.13	PI 97100 X Coker 237	F2	Lee et al. (1996)	Confirmed by Fasoula et al. (2004)
Cytoplasmic		A006b	Yield	Minsoy	216	<0.001	0.06	Minsoy X Noir 1	RIL	Mansur et al. (1996)	slight chlorophyll deficiency seen in early growth

Appendix Table 2. Summary of seed oil QTL peak LOD scores > 3.00, ordered by chromosome, then positions, that were identified by marker regression (MR) and standard interval mapping using EM algorithm. A stratified permutation test of 1000 replicates was conducted in each popualtion to provide a genome-wide 95th percentile LOD score to serve as a statistical significance criterion for evaluating any given QTL LOD score peak. The additive (a) and dominance (d) effects were calculated on the basis of the substitution of a female high protein low oil parent allele for a male low protein high oil parent allele at the indicated marker locus.

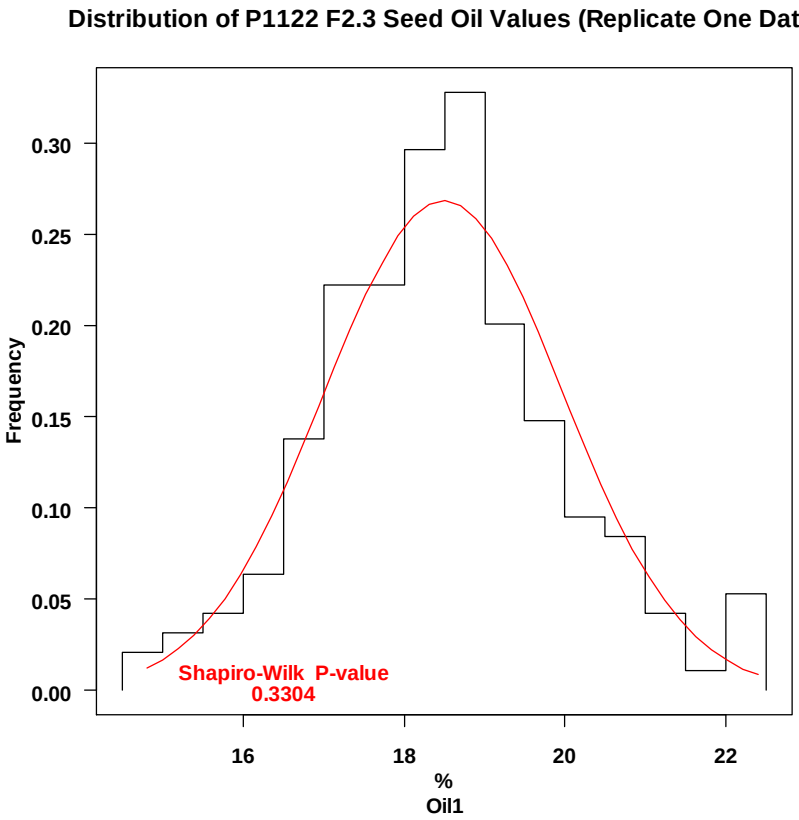
Pop.	Hi-Pro	Lo-Pro	Chr.	LG	Marker or	USLP 1.0	QTL peak LOD (if >3.0)		Permutation-based LOD score		QTL effect		R^2
no.	female parent	male parent	No.	name	nearest marker	map pos.	MR	EM	MR	EM	a‡	d‡	EM
						cM					------g kg^{-1}------		--%--
1122	PI 360843	PI 597387	1	D1a	S24079†	42.01	-	3.72	-	-	6.4	-1.9	9
1122	PI 360843	PI 597387	1	D1a	S30194	59.03	3.36	-	-	-	-	-	8
1139	PI 407788A	PI 606748	3	N	S13702†	86.21	-	3.61	-	-	-7.8	-4.2	9
1076	PI 437112A	PI 597386	5	A1	S16326†	17.65	-	3.21	-	-	9.0	1.3	8
1121	PI 398672	PI 597387	6	C2	S17861	97.81	5.16	4.18	4.11	-	4.5	3.9	8
1122	PI 360843	PI 597387	8	A2	S12625	73.55	4.68	4.37	4.24	3.92	7.3	-2.0	10
1122	PI 360843	PI 597387	10a	O	S26243	58.49	3.08	-	-	-	-	-	7
1076	PI 437112A	PI 597386	10	O	S19004	96.44	6.21	5.20	4.13	4.72	-8.8	3.3	12
1121	PI 398672	PI 597387	10	O	S19004	96.44	3.17	-	-	-	-	-	7
1121	PI 398672	PI 597387	10	O	S15265†	99.69	-	3.19	-	-	-6.2	-0.7	7
1121	PI 398672	PI 597387	11b	B1	S24429	115.55	3.36	-	-	-	-	-	7
1146	PI 407823	PI 606748	14	B2	S30533	13.10	3.26	3.36	-	-	8.3	0.4	8
1143	PI 398704	PI 606748	15	E	S20164	19.80	-	3.06	-	-	-10.5	-0.3	7
1143 (moist)	PI 398704	PI 606748	15	E	S20164	19.80	4.01	4.38	-	4.05	-8.2	1.0	10
1143 (dry)	PI 398704	PI 606748	15	E	S20164	19.80	3.38	3.46	-	-	-9.3	-1.0	8
1121	PI 398672	PI 597387	18a	G	S12541	9.13	4.40	-	4.11	-	-	-	9
1139	PI 407788A	PI 606748	20	I	S17070†	30.00	-	8.32	-	4.17	-13.3	-0.1	21
1139	PI 407788A	PI 606748	20	I	S13577	43.99	8.07	-	4.33	-	-	-	20

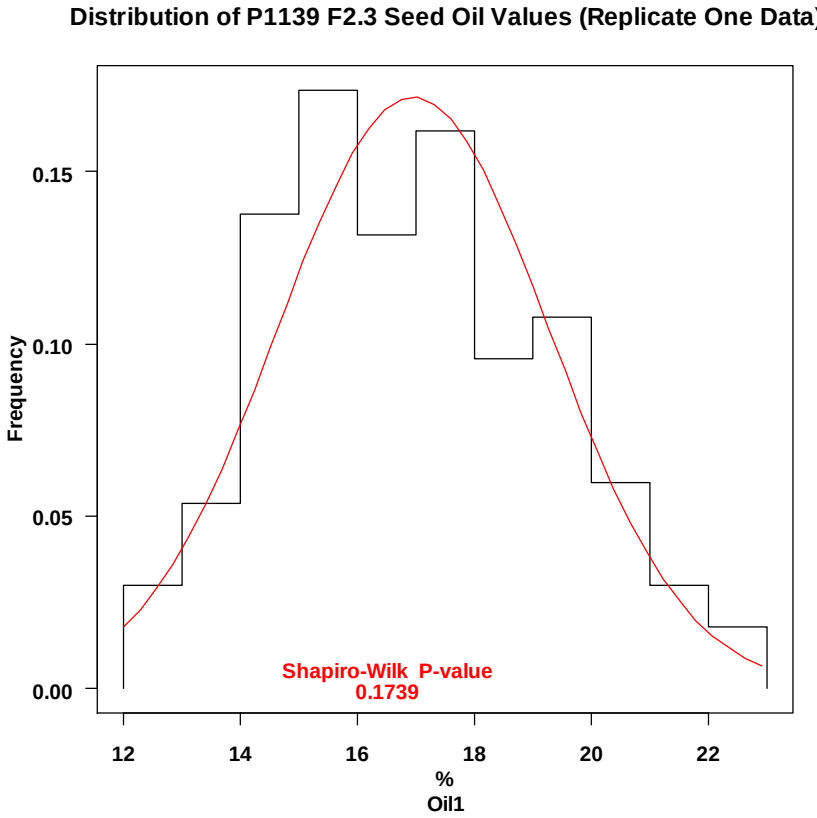
† nearest marker.

‡ If the effect is negative, the high protein low oil parent marker allele depresses seed protein content.

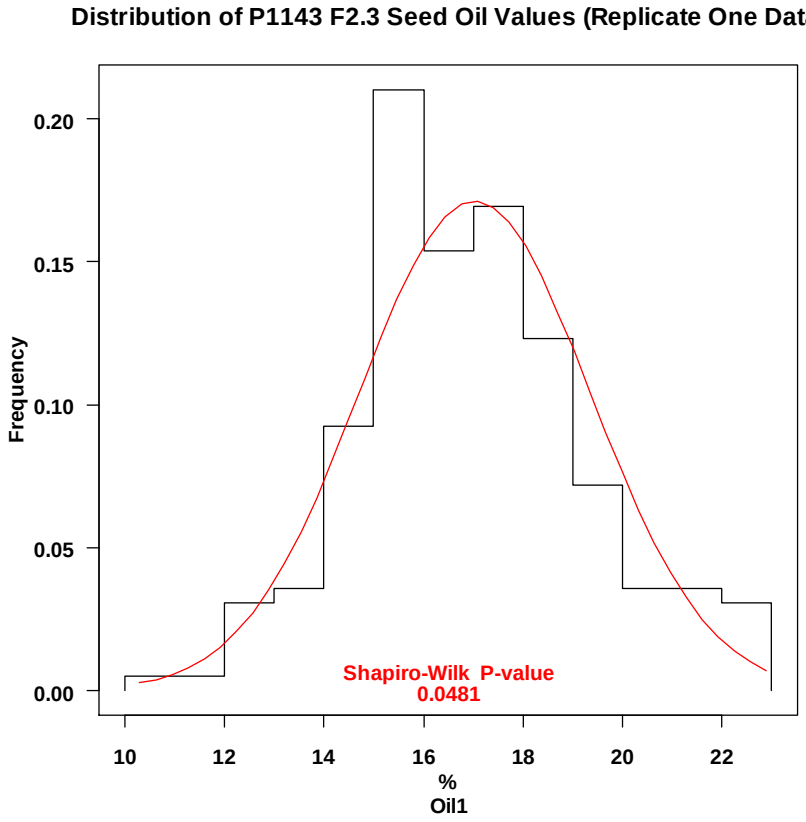


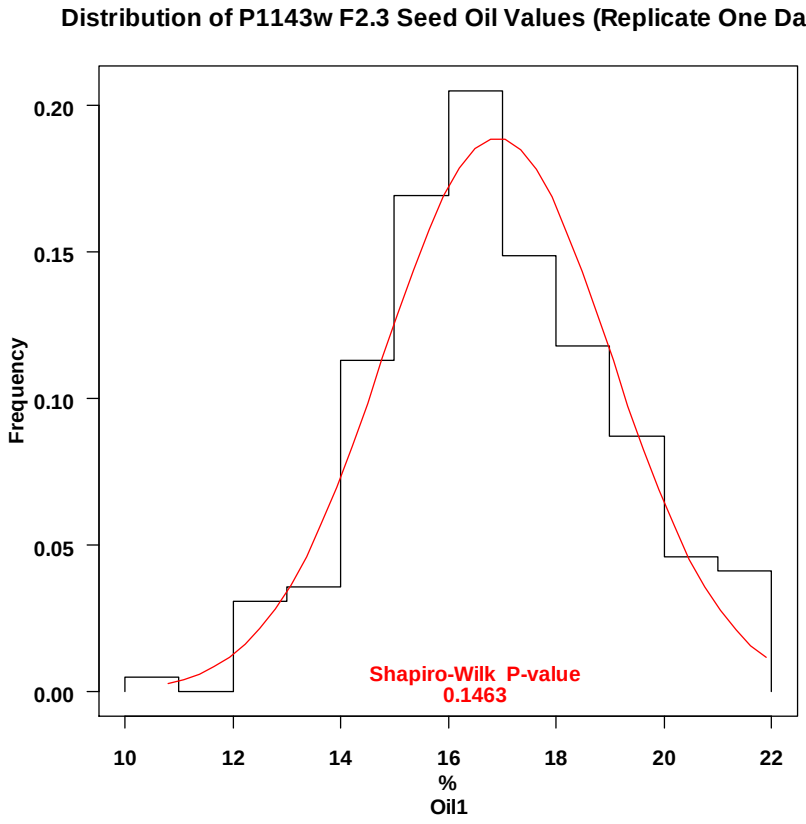
Appendix Fig. 1. Histogram distributions for seed oil phenotype in each of the six $F_{2:3}$ populations. The solid line is showed normal distribution curve.



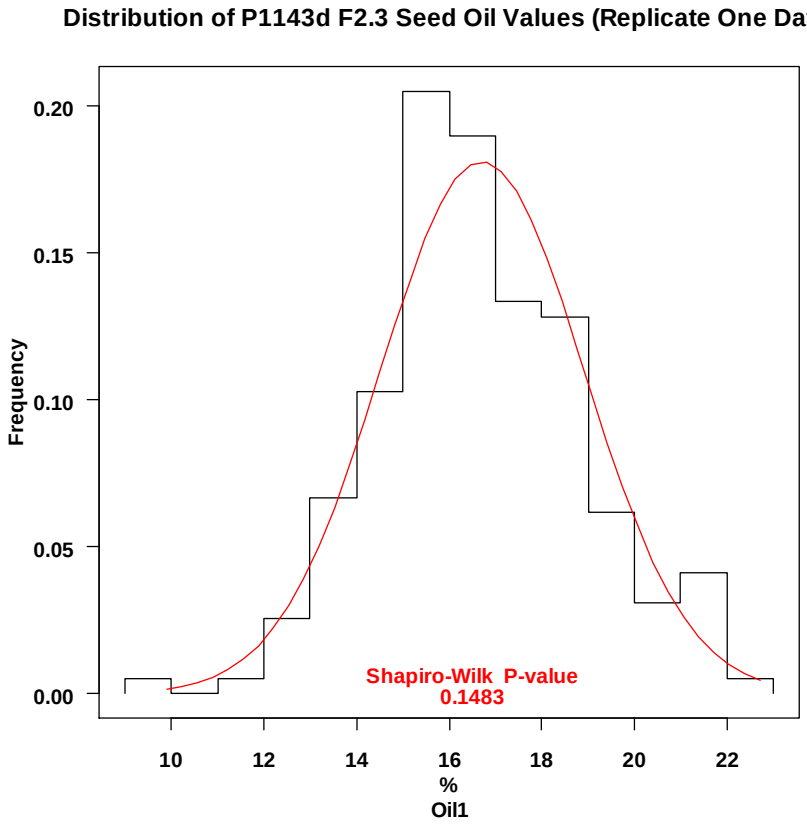


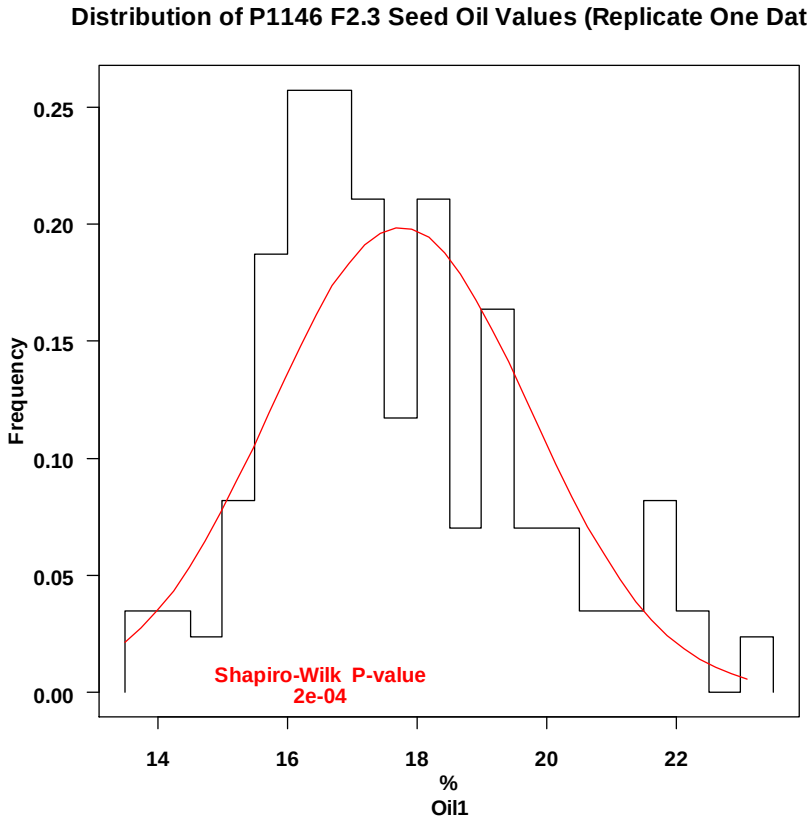
Appendix Fig. 1. (Cont.)



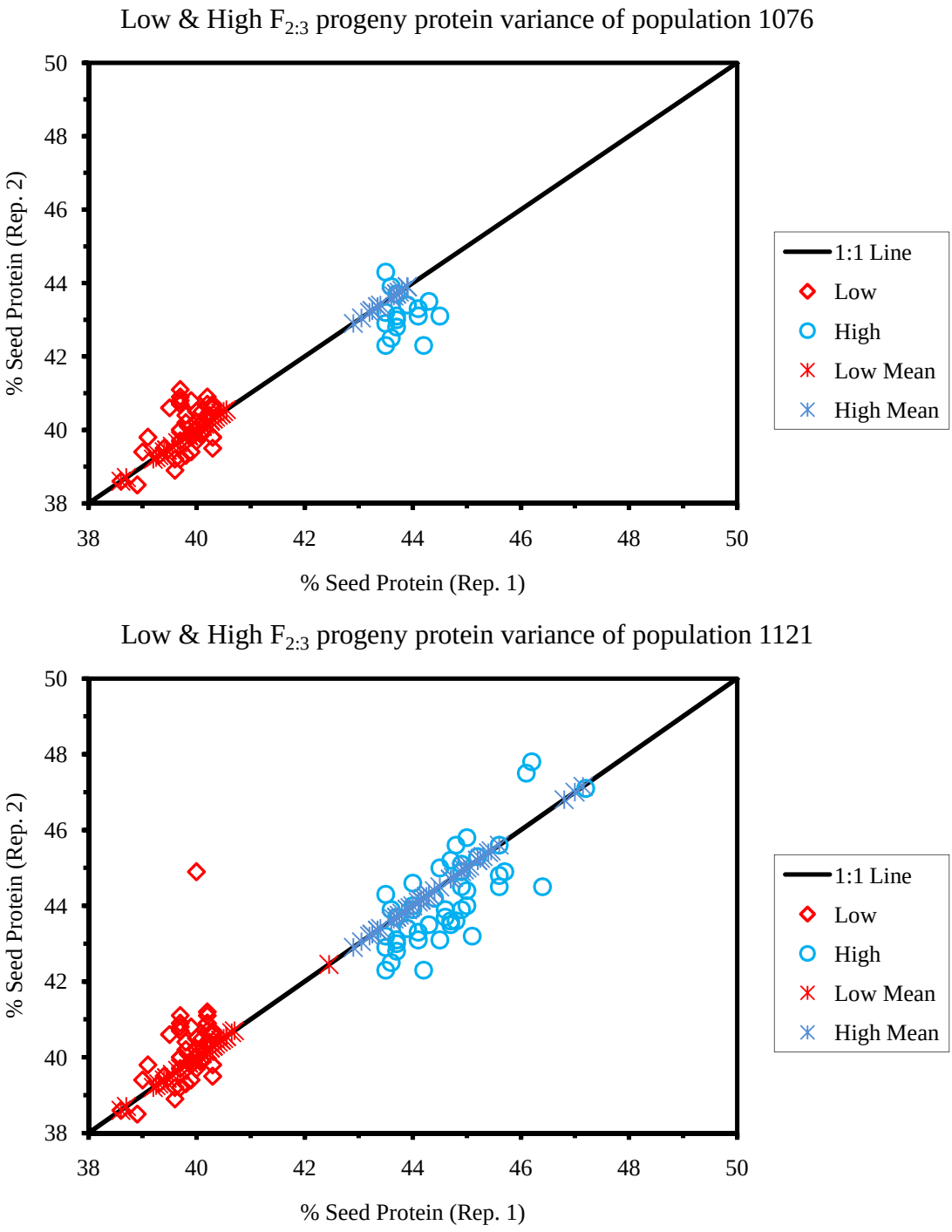


Appendix Fig. 1. (Cont.)

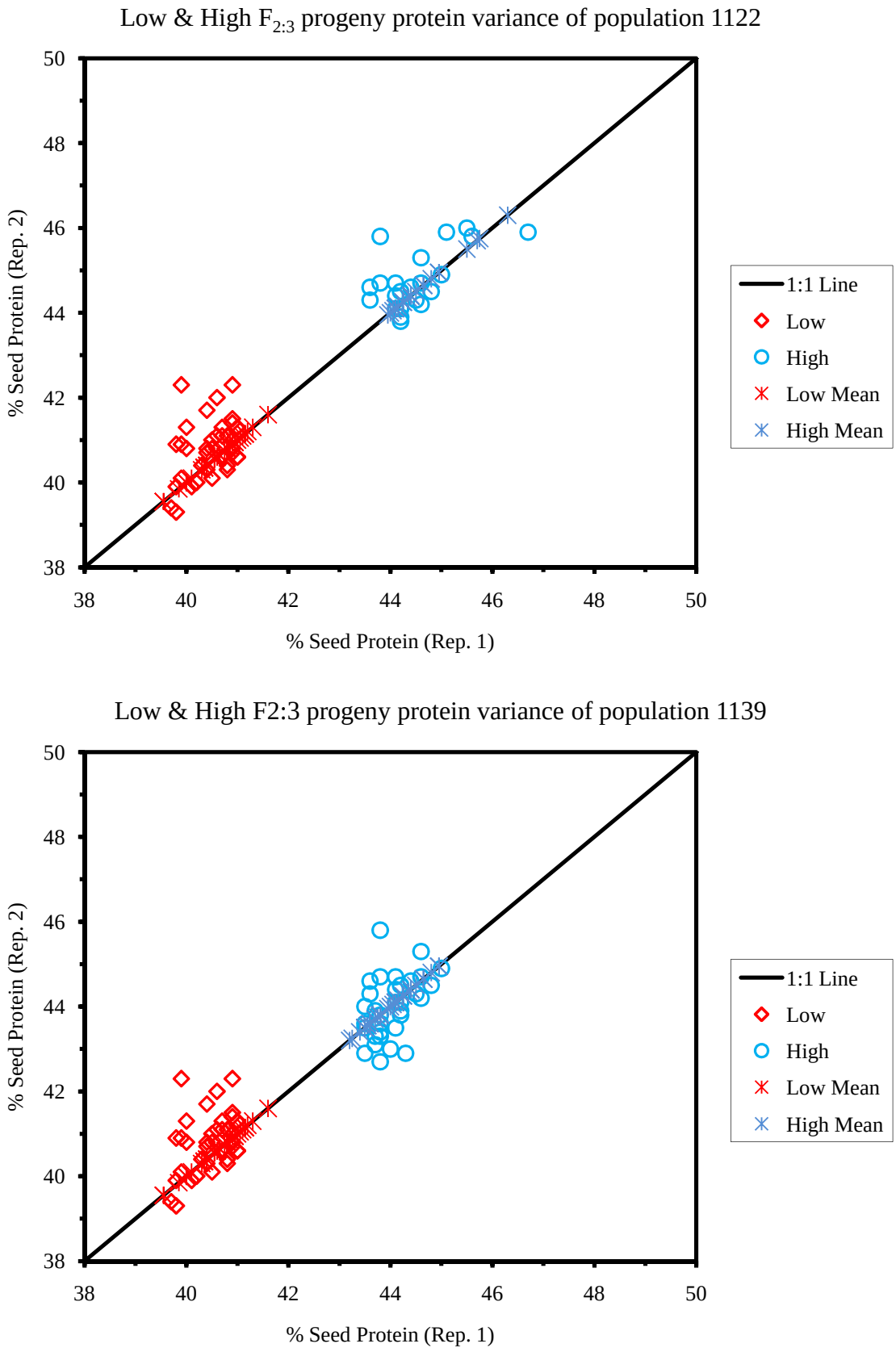




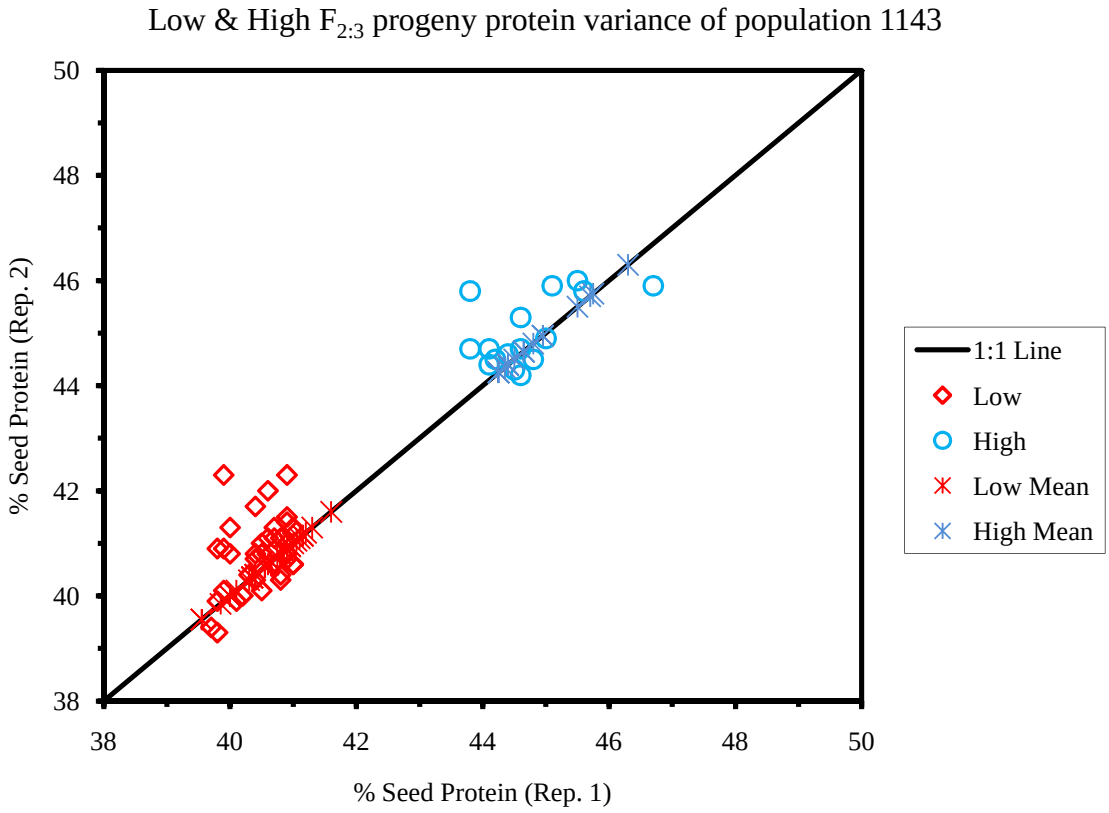
Appendix Fig. 1. (Cont.)

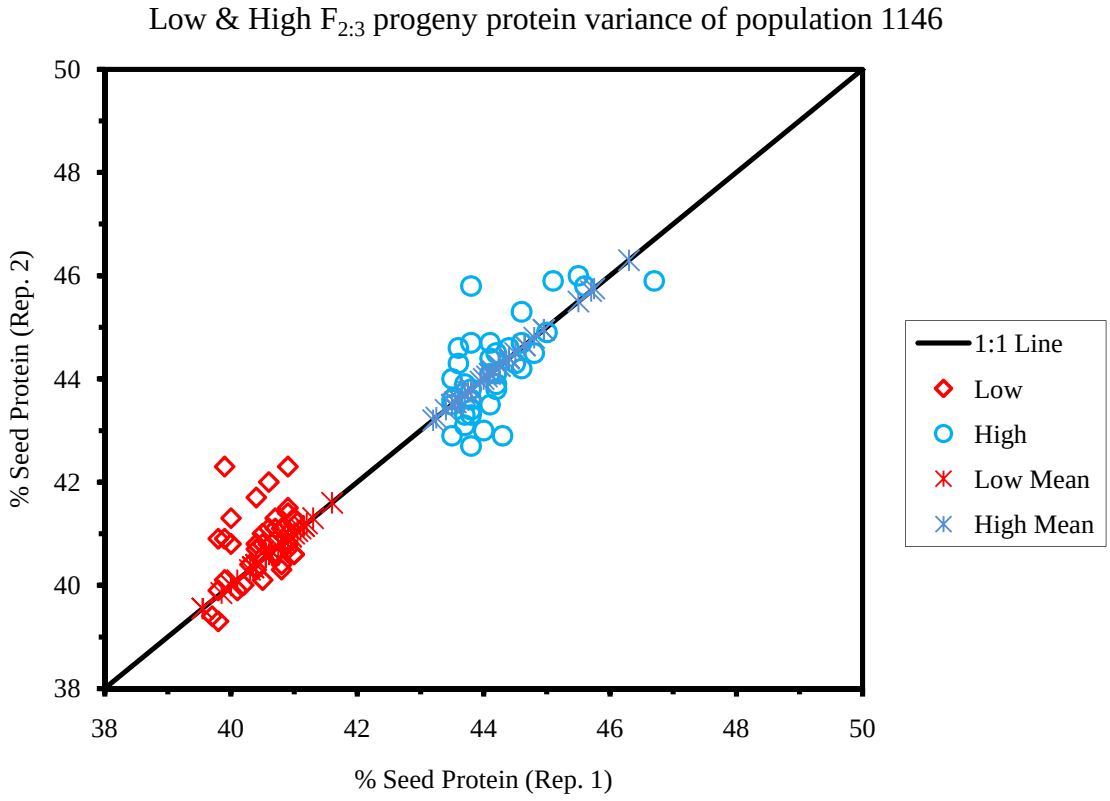


Appendix Fig. 2. The coordinate plots of the replicate one and replicate two low and high quintile $F_{2:3}$ selections in the six $F_{2:3}$ populations. The two-replicate mean seed protein contents of each progeny were also plotted as criss-cross symbols along the 1:1 line.



Appendix Fig. 2. (Cont.)





Appendix Fig. 2. (Cont.)