GENOME-WIDE SCAN FOR LOCI AFFECTING IRON DEFICIENCY CHLOROSIS IN SOYBEAN

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Shireen Chikara

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Genomics and Bioinformatics

July 2010

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Genome-wide Scan for loci affecting iron deficiency Chlorosis in Soybean

By

SHIREEN CHIKARA

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

ABSTRACT

Chikara, Shireen, M.S., Genomics and Bioinformatics Program, College of Graduate and Interdisciplinary Studies, North Dakota State University, July 2010. Genome-wide Scan for Loci Affecting Iron Deficieny Chlorosis in Soybean. Major Professor: Dr. Phillip E. McClean.

Iron deficiency results in iron deficiency chlorosis (IDC) in soybean grown in the north central regions of the United States. Soybean plants display a variety of symptoms, ranging from slight yellowing of the leaves to interveinal chlorosis, and sometimes IDC is followed by stunted growth. In severe cases IDC may even lead to cell death. The objective of this project was to employ a whole genome association mapping approach to uncover the genomic regions associated with the iron deficiency trait in soybean. Golden gate assay technology was applied to expedite the screening of 1,536 single nucleotide polymorphisms in two different sets of soybean populations belonging to the year 2005 and 2006. The two soybean populations were screened for IDC at multiple locations in replicated field trials.

The experiment only considered marker loci with a minor allele frequency greater than 0.1. Probability-probability plot helped in selecting the appropriate general linear models, which controlled for only population structure, and mixed linear models, which controlled for both the population structure and the ancestry. For the 2005 population, three statistical approaches (PCA, PCA+K and PCA+K*) identified twelve marker/trait associations, and for the 2006 population, five statistical models (Q, PCA, Q+K, Q+K* and PCA+K*) resulted in the discovery of twenty-two such associations. Although none of the markers significantly associated with IDC was common to both the populations under study, similar regions of significance were observed between the two years. When the phenotypic and the genotypic data of the two populations were combined, 10 markers were significantly (pFDR < 0.01) associated with the IDC trait using the PCA and PCA+K* statistical models. Out of the 10 markers, six selected markers showed a significant phenotypic mean difference for the tolerant and susceptible alleles. A detailed analysis revealed that using a smaller set of combinations from these six markers can effectively identify IDC tolerant genotypes. The next step would be to verify the reproducibility of the selected set of marker combinations in another set of populations.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my major professor, Dr. Phillip E. McClean, for his valuable guidance and support throughout the graduate research project and in the preparation of this dissertation. It was indeed a pleasure working in his lab under his able tutelage. I am extremely grateful for all the support provided by him which made this study possible.

My sincere thanks are also due to the other members of my Graduate Committee: Drs. Jay Goos, Justin Faris, and Sharyar Kianian. I am indebted to all of them for their incredible help and guidance in the research, analysis and preparation of this dissertation. I am extremely grateful to Dr. Goos and his team for providing phenotypic data of the two populations studied and to Dr. Perry B Cregan, USDA, in whose laboratory the screening of the genotypic data was conducted.

I would like to thank my fellow colleagues, Sujan Mamidi and Rian Lee, for their help and valuable suggestions during the course of study, and to all my Indian and international friends, without whom I would have missed the fun and wonderful life in Fargo, North Dakota.

Last, but not least, I would also like to thank my parents and brother for their encouragement and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Biparental and association mapping are the two strategies employed to uncover marker-trait associations for quantitative traits with the help of molecular markers. These strategies utilize statistical analyses to correlate the phenotypic data with the selected polymorphic markers in a particular study. Biparental populations have undergone a limited number of recombination events in developing F2 or recombinant inbred line (RIL) population. In contrast to this, Association mapping (AM) utilizes numerous recombination events to enhance the resolution of physical linkage between the genetic variants and the gene(s) responsible for the particular genotype. Linkage disequilibrium (LD) refers to the non-random association of alleles at two or more loci on the same or different chromosomes. AM utilizes LD to enhance the ability to detect quantitative traits by utilizing all the meotic and recombination events that have occurred in multiple natural or breeding populations. AM has been conducted for plant traits such as tolerance to cold, flowering time, yield, pathogen resistance, nutrient deficiency and ecological adaptation in a wide variety of crops like Arabidopsis (Aranzana et al. 2007), maize (Thornsberry et al. 2001), rice (Agrama et al. 2007), barley (Cockram et al. 2008), soybean (Wang et al. 2008) and wheat (Tommasini et al. 2007).

Iron deficiency chlorosis (IDC) is one of the major yield-limiting factors in soybean cultivated on calcareous soils in the north central states of the U.S. (Hansen et al. 2004). Iron is taken up from the soil into the root system and is transported to the leaves via the xylem system. This process is regulated at several steps and multiple genes are involved in this process (Clemens et al. 2002). Previous QTL analysis studies on soybean have discovered markers linked with IDC in F2 and recombinant inbred (RIL) populations (Lin et al. 2000; Charlson et al. 2003). However, the results were non-reproducible in other studies. Wang et al. in 2008 conducted an AM study to discover SSR loci associated with IDC.

The primary aim of the present study is to identify SNPs having diagnostic potential in identifying IDC efficient genotypes with the ultimate goal of assisting the breeders in developing IDC tolerant varieties through marker assisted selection (MAS). For this purpose, two populations consisting of advanced breeding lines were first studied individually and later the combined data was evaluated to discover significant marker-trait associations. The notion behind using two different independent populations was to check the effectiveness and reliability of the markers in detecting IDC QTL across multiple populations. These lines were developed by public and private breeding programs for the north central states of the United States. The Golden Illumina Gate assay was used to genotype 1536 SNP markers having genome-wide distribution in the two populations. During the statistical analysis of these SNPs, factors such as population structure and kinship were taken into account as advocated by Pritchard et al. (2000); Price et al. (2006); Zhao et al. (2007).

LITERATURE REVIEW

Soybean

Importance of soybean

Soybean [*Glycine Max* (L.) Merr.], like several other legume species such as common bean (*Phaseolus vulgaris* L.), pea (*Pisum sativum*), peanut (*Arachis hypogarea* L.), lentil (*Lens culinaris* Medik) and chickpea (*Cicer arietinum* L.), has been an important agricultural crop since its introduction in North America in 1765 (Hymowitz 2003) and has been recognized for its nutritive value. It is a rich source of vegetable oil (20%) with high unsaturated and low saturated fatty acids and has a high protein content (40%). Soybean seeds are also rich in soluble-fiber and are a good source of phytochemicals called isoflavones. The plant has the ability to fix atmospheric nitrogen and is compatible with grasses for crop rotation (Keshun 1997; Messina 1999). In the United States, soybean is cultivated mainly for its high protein content and is used as animal feed and for extracting oil for edible and inedible uses. One of the most contemporary industrial uses of soybean oil is in the production of biodiesel, an alternative to fossil fuel, to power diesel engines (Gardner and Payne 2003).

In the United States, 80% of soybean is grown in the upper Midwest, Delta and southeast regions. In these regions, soybean accounts for approximately 90% of the vegetable oil production and is cultivated in rotation with corn. In 2005, the US Department of Agriculture listed soybean as the second most important crop in terms of farm value (Ash et al. 2006).

An increase in soybean acreage and production has been achieved by the cultivation of newer and improved seed varieties, use of fertilizers and improved soybean planting practices. These have resulted in a greater number of pods per acre, thereby boosting productivity (Ash et al. 2006).

Glycine soja-ancestor of soybean

The genomic size of soybean is 1115 Mbp/1C. The diploid ancestor (n=11) of soybean first underwent an aneuploid loss (n=10) and then allo and auto polyploidization events (n=20) (Armuganathan and Earle 1991; Singh and Hymowitz 1988). Soybean belongs to the pea family *Leguminosae* and is classified as genus *Glycine* which is further subdivided into subgenera *Glycine* and *Soja. Glycine max* (G. *max*) belongs to the subgenera G. *soja* (Hymowitz 1970). Morphologically G. *soja* is an annual weedy-form climber, and its pods shatter before the plant matures. It is prone to vining and lodging and lacks complete leaf abscission. Nutritionally, it is rich in seed-protein and linolenic acid with the concentration of the former ranging from 31% to 52%; however, the concentration of lipid, oil and oleic acid is on the lower side (Hymowitz 1970).

Domestication of soybean

The transition of G. *soja* to G. *max (*cultivated soybean) is a result of three genetic bottlenecks. First, the domestication in various parts of Eastern Asian countries led to the production of many Asian landraces. Second, the founding effect led to the selection of a few landraces and to their subsequent introduction in northern and southern US. Third, the selective breeding in the US led to the production of the present-day cultivars (Hyten et al. 2006).

The first genetic bottleneck occurred around 5000 years ago in Eastern Asian countries where the domestication of wild soybean (*Glycine soja*) (Sieb and Zucc.) initially commenced (Ohashi 1982). During this initial phase the classification of soybean cultivars was based on grain color, size and shape, and the time of planting (Smith 1995). The soybean cultivars from China were introduced in Europe in 1712, however, unfavorable climatic conditions in that region prevented its growth (Smith 1995).

The second genetic bottleneck occurred when soybean cultivars from East Asia were introduced in the United States. These soybean cultivars became the germplasm base for the production of subsequent cultivars. This resulted in the founder effect which is defined as a loss of genetic variations that occurs when a new population is established by a very small number of individuals from a larger population. W.J. Morse in 1918 (quoted by Smith in 1995) classified soybean germplasm into Northern and Southern germplasms. Morse grouped cultivars based on their maturity period as late (adapted to southern states), medium late, medium, and very early (adapted to northern states). In 1949, Morse proposed the presence of nine different soybean maturity groups (MG) ranging from 0, I to VIII. According to him, MG 0 and I cultivars were adapted to northern United States, while MG VIII cultivars were better suited to southern United States. Today, however, there are 13 MG ranging from 000, 00, 0, I to X.

The third genetic bottleneck occurred during the intensive breeding and selection process for good phenotypic traits between the Northern and the Southern cultivars in the United States (Gizlice et al. 1993). These three genetic bottlenecks have decreased the genetic diversity in the current soybean cultivars. In the 258 Northern cultivars released between 1947 and 1988, 35 ancestors contributed to more than 95% of all the alleles (Gizlice et al. 1993).

Important characteristics of the released soybean cultivars

Selective breeding by skilled soybean breeders has led to the development of cultivars better adapted to the US. After World War II, research on soybean cultivars in the United States focused on plant breeding, marketing, plant physiology and pathology, and soil and weed science (Windish 1981). Soybean breeding has been mainly aimed towards improving yield, seed size, seed protein and oil quality and quantity, shattering of the pods, emergence, plant height, and resistance to various crop limiting affections. The conditions that commonly affect soybean are yellow mosaic virus (Singh et al. 1974), phytophthora root rot (Kaufmann and Gerdemann 1958; Kenworthy1989), alfalfa mosaic virus (Horlock et al. 1997) and soybean cyst nematode (*Heterodera glycines* Ichinohe) (Riggs et al. 1998). In addition to these, research is also directed towards improving lodging resistance and mineral nutrient resistance, especially the iron deficiency chlorosis (IDC). These events are predominantly seen when soybean is grown in calcareous soil (Chen and Barak 1982).

Iron Deficiency Chlorosis

Loss in soybean productivity

IDC is a very common and important yield-limiting factor affecting soybean grown on calcareous soil. Calcareous soil, with a relatively high percentage of calcium carbonate and soluble salts, is commonly present in the north-central regions of the US and extends from central Iowa to central Minnesota and further into southeast South Dakota (Franzen and Richardson 2000). In the US, the yearly loss of soybean yield due to IDC, has been estimated to be USD120 million. In Iowa and Minnesota alone, IDC leads to a loss of over ten million dollars due to decreased soybean production (Hansen et al. 2004).

Importance of Fe in plants and its storage

Iron (Fe) is the fourth most abundant element in the earth's crust, and its concentration in the soil ranges from 0.5% to 5% (Ma and Nomoto 1996). Fe is needed for various biological redox processes such as chlorophyll synthesis and photosynthesis in the leaves, plant DNA and hormone synthesis, and nitrogen fixation (Marschner 1995). As a constituent of porphyrin ring precursors, iron plays an important role in the formation of chlorophyll in the chloroplasts. Because it is able to form six coordinated links with the electron donor atoms like oxygen and nitrogen, Fe is associated with heme and Fe-S cluster proteins (Marschner 1995). Iron is also involved in enzymatic systems such as the prosthetic groups of cytochromes that enable electron transport, and cytochrome oxidase. It also plays a significant part in the terminal step of respiration chain (Audebert 2006).

Iron is stored in the plants in the apoplasmic spaces, vacuoles, and also as ferritin located in the plastids (Briat and Lobreaux 1998; Harrison and Arosio 1996). All ferritins have a conserved three dimensional structure having 24 protein subunits arranged in 432 symmetry with a hollow shell of about 80 Å diameter. The cavity is adept at storing up to 4500 Fe^{2+} ions as an inorganic complex (Harrison and Arosio 1996). Ferritins play an important role in iron homeostasis by storing iron in the seeds. It also alleviates environmental stresses by mobilizing iron stored in cotyledons to young seedling axis (Hyde et al. 1963).

Phenotypic indications of IDC

Photosynthesis takes place in chloroplast which accounts for about 80% of leaf cell iron content and is the major plant cellular machinery where chlorophyll production occurs (Smith 1984). Inadequate production of chlorophyll by the plant, and its deficiency in the leaves leads to chlorosis, i.e., yellowing of the leaves (Brown 1956). IDC can be due to several reasons such as a deficiency of iron in the soil during the growth of seedlings, an inability of the plant to mobilize the absorbed Fe from the roots to the leaves, or an inability of the leaves to utilize Fe. IDC symptoms range from slight yellowing of the leaves, with no differentiation between veinal and interveinal areas, to interveinal chlorosis in which the veins remain green, while the interveinal area becomes yellow. In severe cases, interveinal chlorosis is followed by a stunted growth or even cell death. As per the chlorosis rankings, described by Lin et al. (1977), plants are ranked on a scale of 1-5 where rank 1 indicates green plants with no chlorosis, while rank 5 denotes severe chlorosis with reduced plant growth followed by necrosis of some leaf tissues. Severe chlorosis may even lead to plant death (Froehlich and Fehr 1980). IDC symptoms are generally more visible in young leaves as Fe, a relatively immobile element bound in a complex with ferritin within the chloroplast, cannot move from the mature leaves to the younger ones (Waldo et al. 1995). In 1943, Weiss coined the terms Fe-efficient and Fe-inefficient plants. According to him, when grown on calcareous soil, plants which develop IDC are termed as Fe-inefficient and those which do not are termed as Fe-efficient.

Factors that impede the uptake of Fe present in the soil

Onken and Walker (1966) suggested that IDC in sorghum resulted not due to a lack of iron in the soil but due to factors that affect its solubility. Several plant and soil factors responsible for chlorosis have been identified by Brown 1959a, 1959b; Walter and Aldrich 1970; Coulombe et al. 1984; Nikolic and Romheld 1999; Barker and Pilbeam 2007. Some of the soil factors responsible for chlorosis include soil pH, which leads to the predominance of ferrous (Fe^{2+}) or ferric (Fe^{3+}) forms of iron, presence of bicarbonates, soil compaction, soil temperature, and soil water and heavy metals content. Plant factors include low root growth which controls Fe solubility in the soil solution and plant sap, and low Fe efficiency.

It has been postulated that for the proper growth of a plant, the iron concentration in the soil should be 10n*M* (Stephan 2002). Iron availability is governed by the soil redox potential and its pH. When the soil pH is low, ferric (Fe^{3+}) iron is reduced to ferrous (Fe^{2+}) form which is then readily available to the plant as per the following chemical process:

$$Fe(OH)_3 + 3H^+ + e^- = Fe^{2+} + 3H_20$$

When the soil pH is high, as seen in calcareous soil, ferrous iron is oxidized to ferric iron and is not readily available to plants. Under such conditions, the concentration of iron in the soil is not higher than 100p*M* (Stephan 2002). However, at the same time, an excess of Fe^{2+} under acidic or reducing soil conditions would be toxic for the plant if the plant is unable to control iron fluxes and subsequently protect itself from the active cell metabolism which results when iron is absorbed at the root level (Stephan 2002). Water-logging occurring due to poor soil structure may lead to an accumulation of bicarbonates in alkaline or calcareous soils. "Lime induced chlorosis" (iron deficiency) generally occurs as elevated concentration of bicarbonates (HCO_3^-) in the root apoplast impedes Fe^{3+}-chelate reductase activity by neutralizing the H^+ ions pumped into the cytosol to decrease the pH. Also, the uptake of nitrate/H^+ co-transport increases the pH of the soil, further hindering Fe^{3+} reduction. Bicarbonates also immobilize the movement of iron to young leaves once it is absorbed at the root level (Barker and Pilbeam 2007). However, at the same time, under acidic conditions in a water-logged field, anaerobic bacteria may efficiently reduce iron to Fe^{2+} and an excess of the same may result in iron toxicity.

The uptake of iron from the soil may be hindered due to its interaction under acidic soil conditions with other elements such as copper, calcium, magnesium, potassium and manganese. The high solubility of these elements suppresses the uptake of iron. Additionally, an excess of organic matter in the soil and/or compaction of the soil may cause poor root growth and generation of ethylene which hinders iron uptake, resulting in "ethylene-induced chlorosis" (Barker and Pilbeam 2007).

The identification of the causes of IDC in soybean is often complicated due to changing environmental conditions (Cianzio et al. 1979). It is further adversely influenced by infestation of the plant with soybean cyst nematodes (SCN, *Heterodero glycines* Ichinohe), a condition which often produces symptoms similar to IDC (Tylka 2001).

In iron deficiency conditions, a set of coordinated responses are triggered. These responses enable the plants to assimilate an optimal amount of iron from the soil, utilize iron stores, and systematize the regulation of iron in the intercellular and intracellular compartments for various important cellular processes.

Biochemistry involved in the uptake of Fe from the soil

Brown et al. (1958) performed a reciprocal graft experiment and discovered that the uptake of iron is controlled by factors at the root level. They grafted the Fe-inefficient T203 soybean tops on Fe-efficient Hawkeye (HA) and Fe-efficient HA tops on the Fe-inefficient T203 rootstocks. They observed that the former became iron efficient and turned green while the latter developed chlorosis. The reciprocal experiment was also conducted in tomato using Fe-inefficient T3238FER and Fe-efficient T328FER by Brown et al. (1971) and the same results were obtained.

Plants can uptake iron from the soil in both its ferrous or ferric forms. When the more soluble form of Fe, i.e., Fe^{2+} predominates in the soil, both the Fe-efficient and Fe-inefficient plants are able to absorb it from the soil into their root system. However, when Fe^{3+} predominates in the soil, Fe-efficient plants initiate a series of biochemical reactions to absorb Fe^{3+} from the soil, while Fe-inefficient plants fail to do so (Brown 1978).

Marschner et al. (1986) divided plants into two categories, strategy I and strategy II plants, based on their mechanism of response to Fe availability. Strategy I plants include all dicotyledons and nongraminaceous plants such as Arabidopsis (*Arabidopsis thaliana)*, pea (*Pisum sativum*) and soybean (*G. max*). In these plants, three steps regulate the uptake of Fe from the soil. First, there is a release of H^+ from the root surface by the proton pumping H^+ ATPase which lowers the pH in the soil rhizosphere. Acidification of the soil initiates the dissociation of $Fe(OH)_3$ complexes into ferrous ions. The lowering of pH by one unit increases the solubility of Fe^{3+} ions by a factor of thousands (Connolly and Guerinot 1998). Second, there is a reduction of Fe^{3+} by Fe^{3+} chelate reductase to the more soluble Fe^{2+}. At neutral pH ferrous ions are 10^6 times more soluble than ferric ions. Third, the iron transporters carry out the plasmalemma transport of Fe^{2+}. Strategy I plants also undergo changes in root morphology and biochemistry. The changes include an increase in root hair formation, thereby increasing the surface area available for iron uptake, and an increase in the citrate concentration in the phloem (Schmidt 1999).

Strategy II plants include all graminaceous plants like wheat (*Triticum aestivum*), rice (*Oryza sativa*), barley (*Hordeum vulgare*) and maize (*Zea mays*) (Mori 1999). Under iron deficiency conditions, these plants synthesize and secrete phytosiderophores (PS), a non-proteinogenic group of amino acids. PS belong to the mugineic acid family of compounds and act as ferric chelators in the apical zones of the roots (Takagi 1976). During the biosynthesis of PS, the activity of nicotianamine synthase, the first important enzyme involved in the pathway, increases (Higuchi et al. 1996). The Fe^{3+} -PS complex is then taken up by the iron uptake system from the soil (Mori 1999).

Translocation of absorbed Fe at root level in plants

It has also been observed that even when the concentration of Fe is high at the root level, the plants grown on calcareous soils may exhibit chlorosis (Mengel 1994). Along with an efficient root uptake system for the absorption of Fe from soil, an efficient root-to-shoot translocation of iron is imperative for the utilization of iron by the plant. The radial transport of iron either occurs through the symplast of rhizodermal cells or it is transported apoplasmically and introduced to the symplast of the cortex (Nikolic and Romheld 1999). It is imperative to balance iron homeostasis during its symplasmic transport because excess iron reacts with oxygen to produce free toxic radicals and decreases photosynthetic activity (Guerinot and Yi 1994; da Silveira et al. 2007). The reactive free radical can cause serious damage to cellular components leading to cell death (Guerinot and Yi 1994). Therefore, iron molecules are chelated with the help of non-proteinogenic amino acid nicotianamine (Stephan et al. 1996.). The Fe^{2+-}nicotianamide complex so formed is oxidized to Fe^{3+} in the symplast of the root to facilitate its transport to the leaves through the xylem system.

From the root symplast, iron can also be loaded into the xylem vessel as a ferric iron-citrate complex (Tiffin 1966). The re-transfer of iron from symplasm to apoplasm occurs with the help of a respiration dependent proton pump at the plasma membrane of xylem parenchyma cells (DeBoer et al. 1983). Kohler and Raschke (2000) based on their study on measurement of plasma membrane potential, doubted the loading of iron in xylem vessel as an energized process and concluded that it is a thermodynamically passive process occurring through an ion channel.

The phloem sap sends signals to the root tips about the level of iron in the shoot. As the pH of the phloem sap is more than 7, iron is chelated to avoid its precipitation (Stephan et al. 1996; Gerendas and Schurr 1999). However, other synchronous studies have shown that protein-bound iron transport is generally favored as proteins have higher binding affinity in the phloem (Marentes et al. 1997; Wang et al., 1999).

Methodology to avoid loss of soybean yield due to IDC

IDC is a common problem which causes reduction in soybean yield in the North-Central regions of the US. Soybean yield loss due to IDC can be prevented by: (i) developing high yielding chlorosis resistant cultivars through breeding programs and (ii) studying the genetic inheritance for iron utilization in soybean plants. Plant breeders follow a rigorous process of selection for several years to develop improved cultivars having a desired trait, while a geneticist tries to understand the underlying mechanism of inheritance and variations for the same trait (Bernardo 2008).

Development of an iron efficient cultivar

Breeders aim to develop high yielding cultivars to combat chlorosis when grown on calcareous soil. Through this approach, a partial improvement in IDC resistance in soybean cultivars has been achieved. On an average, a 20% increase in yield is expected with just one unit decrease in the chlorosis score. In breeding practices, the development of a desired cultivar involves the following steps: (i) crossing genotypes with desirable traits to generate variations in the segregating population (like F_1, F_2 populations and so on), (ii) selecting genotypes which show the desirable traits from both the parents, followed by (iii) pedigree breeding, recurrent selection, and repeated backcrossing to develop improved cultivars. Breeders have developed IDC resistant cultivars since 1970s. A2, released in 1978, was the highest yielding cultivar resistant to IDC. Iowa State released five IDC resistant germplasm lines: A11, A12, A13, A14 and A15 (Jessan et al. 1988)

Breeding programs of private seed companies such as Monsanto, Pioneer International Inc. (Helms et al. 2005 and Helms et al. 2009), Iowa State University (Cianzio 1991), and University of Minnesota (Orf and Denny 2004) are directed towards development of IDC resistant cultivars which may be made available commercially. Cultivars and germplasm with improved resistance to IDC are available for oat (*Avena byzantina* C.Koch), sorghum (*Sorghum bicolor* (L.) Moench), dry bean (*Phaseolusvulgaris* L.) and soybean (G. *max*).

However, at the same time, the soybean breeding programs can be hindered due to several factors. First, due to heterogeneous nature of the soil, IDC symptoms may vary from severe to nonexistent within a couple of meters; hence, variety screening to evaluate Fe efficiency is hard to perform (Diers et al. 1991; Lin et al., 1999). Second, visual analysis of chlorosis is unreliable as occasionally chlorotic symptoms observed during the early stages of plant development may disappear as the plant matures (Franzen and Richardson 2000). Third, variety breeding and selection is laborious, time consuming and expensive. It has been estimated that it takes at least six years of intensive breeding before a new plant variety is made commercially available. This means that soybean crosses made in 2005 will be able to reach preliminary trials in 2008 and then be commercially available in 2011 (quoted by Wang et al. 2005). Fourth, if the trait under investigation is a quantitative trait, it may at times be controlled by a few genes having a large effect and at other times by several genes having a small effect. Breeders usually select more than one trait in a breeding program, i.e., multiple QTL at one time. They have no control over the recombination events occurring during meiosis which segregates the desired QTL, thereby impeding the development of the desired genotype (quoted by Bernardo 2008).

Genetics of Fe uptake by plants

Identification of candidate genes involved in iron utilization in model organisms like Arabidopsis and *Saccharomyces cerevisiae* have helped in our understanding of the role of multiple genes in the biochemical mechanisms under Fe deficient conditions. Grusak and Pezeshgi (1996) studied the expression of ferric reductase oxidase (*FRO*) gene in pea (*Pisum sativum*) and put forward the idea that indications for Fe deficiency in the plant pass from the shoots to the roots. In strategy I plants (dicots and non-graminaceous plants), FRO protein regulates the release of Fe^{3+} chelate reductase which reduces Fe^{3+} to a more soluble Fe^{2+}. The Fe-regulated transporter1 IRT1 protein regulates the release of Fe^{2+} transporter, which helps in the uptake of Fe^{2+} from the soil into the root epidermis (Eide et al. 1996; Henriques et al. 2002). The genes for these proteins were cloned in Arabidopsis based on their sequence similarity with yeast homologue *FRE*1 and *FRE*2 genes (Eide at al. 1996, Robinson et al. 1999). Arabidopsis has seven members of the *FRO* gene family. Under Fe deficiency conditions, the expression of mRNA for *FRO2* and *FRO5* is elevated in the roots, while, for *FRO*3 it is increased in both the roots and the shoots, for *FRO8* it is increased in shoots, while for *FRO*6 and *FRO*7 the same is increased in photosynthetic tissue of the plants (Mukherjee et al. 2006). *FRO*2 is essential for the expression of root specific Fe^{3+} chelate reductase (Robinson et al. 1999). *FRO*1 has been identified in pea and tomato (Waters et al. 2002; Li et al. 2004). In pea, it is expressed in the root epidermis and cortex, in the nodules, and in the mesophyll and parenchyma of the leaves (Waters et al. 2002). After Fe^{3+} has been reduced to Fe^{2+}, it is transported into the roots by IRT1 regulated metal transporters. *IRT*1 gene was cloned in the Arabidopsis roots by functional complementation of a yeast mutant fet3/fet4 (Eide et al. 1996). Both IRT1 and IRT2 proteins are members of the *ZIP* (**Z**rt-**I**rt-like **p**roteins) gene family (Vert et al. 2001). *IRT*2 is expressed in the epidermal cells of the plant roots and has same function as *IRT1* (Vert et al. 2001). The induction of *FRO*2 and *IRT* under Fe deficiency conditions is best understood from the studies of the tomato mutant *fer* (Ling et al. 2004). In tomato the knockout mutant of AtbHLH29 showed Fe-deficiency symptoms due to a lack of *leIRT1* and *leFRO* gene expression (Ling et al. 2004). *FER*, a regulatory gene, encodes for a protein bHLH, which is involved in the mechanism of iron uptake in the roots of tomato. The protein AtbHLH29 (At2g28160) of Arabidopsis, showed high sequence similarity with the *FER* sequence at protein level in tomato. Protein AtbHLH29, encoded by *FIT1* or *FRU* gene is necessary to initiate the expression of Fe mobilization genes, *IRT1* and *FRO2*, in the roots of Arabidopsis. At2g28160 is named as FRU (**F**er-like **r**egulator of Iron **u**ptake or FIT1= **F**e-deficiency **i**nduced **t**ranscription factor1) (Jakoby et al. 2004; Bauer et al. 2007). Since over-expression of *FRU* in Arabidopsis initiates Fe uptake responses, this gene is considered to be conserved in strategy I plants (Jakoby et al. 2004). *FRU* gene AtbHLH29 is an ortholog to *FER* gene in tomato (Yuan et al. 2005). Arabidopsis has four Fe transporter gene families: *ZIP*, natural resistance associated-macrophage protein (*NRAMP*), yellow stripe like 1 (*YSL*1) and iron regulated exporter from gut (*IREG*) (Eide et al. 1996; McKie et al. 2000; Thomine et al. 2000; Le Jean et al. 2005). The *NRAMP* family consists of six genes that have been identified in Arabidopsis (Maser et al. 2001). Complementation studies in yeast Fe-uptake mutant identified that *AtNramp1*, *AtNramp3* and *AtNramp4* are expressed under Fe deficiency condition (Curie et al. 2000). *AtNramp1* also plays a role in the distribution of Fe in the cells as its over-expression provides resistance to Fe toxicity (Curie et al. 2000). Transporter AtYSL1 helps in the uptake of Fe^{2+}- nicotianamine (Fe-NA) chelate complexes into the seeds and also transports the same in the xylem. It is expressed in the vasculature and the intercostal regions of the leaves (Le Jean et al. 2005; Waters et al. 2006). YSL transporter protein in Arabidopsis shares homology with the strategy II plant maize (*Zea mays*). For the synthesis of chlorophyll and for the important role iron plays in photosynthetic electron transport, Fe has to reach the chloroplast. A permease gene in the chloroplast (*PIC1*) was identified in Arabidopsis (Duy et al. 2007). In pic1 mutant plants, they observed upregulation of *FER*1, *FER*4, *YSL*1 and *IRT*1, and the presence of ferritin clusters in plastids. However, as the transport of Fe was blocked at the inner chloroplasts, the plants developed chlorosis and had dwarf phenotypes.

In Strategy II plants, genes responsible for the release of PS are nicotianamine synthase (*NAS*), nicotianamine aminotransferase (*NAAT*) and iron deficiency specific gene (Mori 1999). *NAS* regulates the synthesis of nicotianamine (NA) which undergoes a deamination step by *NAAT* followed by another reduction step by deoxymugineic acid synthase (*DMAS*) (Bashir et al. 2006). Fe^{3+}-PS complex is transported within the roots by specific transporters like YS and YSL (Higuchi et al. 1996; Schmidt 2003). Rice, a strategy II plant, releases PS in Fe deficiency conditions. Takagi (1976) observed that PS secreted from the roots of the rice plant facilitates the uptake of Fe from the soil. The amount of PS released is proportional to the level of Fe deficiency in the soil. The Fe^{2+} transporter genes *OsIRT1* and *OsIRT2* in rice are homologous to the *IRT*1 in non-graminaceous plants (Bughio et al. 2002; Ishimaru et al. 2006). OsIRT1 has features identical to those of the ZIP metal transporter family (Bughio et al. 2002). However, both Os*IRT*1 and Os*IRT*2 in rice, and Le*IRT*1 and Le*IRT*2 in tomato also reverse the growth defects of the yeast copper uptake mutant ctr1 (Dancis et al. 1994). But, *IRT*1 gene is not able to do so in Arabidopsis (Eckhardt et al. 2001). NA is a mobile non-protein amino acid found in the root and the leaf cells, as well as, in the phloem sap (Hell and Stephen 2003) and it is important for translocation of Fe and its accumulation in developing seeds as studied by Takahashi et al. (2003) in NA-deficient transgenic tobacco. Two *NAAT* genes, previously identified in barley, when introduced in rice did not have much effect on the response of the plant to Fe deficiency. However, it was observed that there was an increase in the grain yield of these genetically modified rice plants when grown in alkaline soil with limited Fe availability (Takahashi 2001).

Like other dicotyledon plants, soybean, a strategy I plant, responds to iron non-availability in calcareous soil by inducing an active proton pump, a ferric reductase and an iron transporter mechanism. Several studies have been conducted to find candidate genes associated with IDC in soybean. Vasconcelos et al. (2006) introduced Arabidopsis *FRO2* gene into soybean through agro-bacterium mediated transfer to study the gene's heterologous expression. They observed that hydroponics studies, with Fe^{3+}-DTPA as a source of iron, showed a relatively reduced chlorosis response in *FRO2* transgenic soybean, while the non-transgenic control soybean plants had yellowing of the leaves. It was noticed that the transgenic soybean plants expressing Arabidopsis *FRO2* genes showed a higher concentration of chlorophyll in the leaves as compared to that observed in control soybean plants. A cDNA microarray study of soybean RNA, isolated from the roots of two near-isogenic lines, PI 548533 (Clark, iron efficient) and PI 547430 (IsoClark, iron inefficient), showed differences in iron efficiency. A total of 43 genes were differentially expressed, and while 24 of these genes showed sequence similarity to genes associated with Fe stress, the remaining 19 were unique to the soybean Fe response (O'Rourke et al. 2009).

IDC - A Quantitatively Inherited Trait

Quantitative traits show a continuous phenotypic distribution since they are controlled by more than one interacting loci and are under the influence of the environment (Falconer and Mackay 1996). The interacting loci may act in an additive, dominant, and epistatic fashion with each other (Mackay 1996, 2001). Also, the distribution of multiple genes over the genome may be in a fixed or a random order. QTL mapping helps to gain insight into the genetics of quantitative variation (Bechmann and Sollerm 1986).

Weiss (1943) studied the inheritance pattern for IDC in soybean and concluded that iron utilization in soybean is controlled by a single major gene with dominant allele (Fe) for iron efficiency and recessive allele (fe) for iron inefficiency with complete dominance. He ignored some variations among the inefficient cultivars in response to their degree of iron efficiency as the effect of the modifying genes was little as compared to that of the dominant gene. Studies by Bernard (1947), and Cianzio and Fehr (1980) further affirmed the single gene inheritance for IDC. However, Cianzio and Fehr (1982) working on a breeding program at Iowa State University observed a continuous distribution of chlorosis scores ranging from 1.2 to 4.8 with a mean of 2.8 in the F¬2 segregating population developed from a cross between Pride B-216 (a high yielding cultivar susceptible to IDC) and A2 (a low yielding cultivar resistant to IDC). Seven selected genotypes out of the 200 F_2 derived lines were backcrossed to the high yielding cultivar Pride B-216. None of the 280 BC_1F_2–derived lines were as IDC resistant as A2. Therefore, it was concluded that IDC was a quantitative trait and its inheritance was controlled by additive gene action.

Molecular Genetics as an Aid to Decipher QTL

The discovery of DNA based molecular markers has paved the way for intensive genetic research to decipher the location of individual genes controlling a quantitative trait. Molecular markers are genetic variations in the genomic sequences between genotypes. These variations may or may not have any direct effect on the phenotype and are not influenced by environment (Charlson et al. 2005). Molecular markers help in the construction of linkage maps, which enable linkage analysis of agronomically important traits.

For marker assisted selection (MAS) to be effective, the construction of a high density genetic map where markers are tightly linked to the trait of interest is preferred. A high resolution genetic map helps to restrict the location of genes to a narrow frame in the complex genomic organization and it can be used to statistically associate the marker variant with the trait under study. If the marker is in close proximity to the QTL, i.e., 2cM or less, the marker may be substituted for the gene itself in MAS and thus traversing from phenotype dependent selection to genotype dependent selection. For a molecular marker to be incorporated in the MAS, it should be: (i) easy to use, (ii) cost effective, (iii) able to be screened using high-throughput analysis and multiplexing, (iv) able to produce reproducible results, (v) highly polymorphic and (vi) co-dominant in nature, i.e., have an ability to detect heterozygotes.

Molecular markers are broadly classified into three groups: (i) the first generation molecular markers such as restriction fragment length polymorphism (RFLP) and random amplified polymorphic DNA (RAPD), (ii) the second generation molecular markers such as simple sequence repeat (SSR) and amplified fragment length polymorphism (AFLP), and (iii) the third generation molecular markers such as expressed sequence tags (EST) and single nucleotide polymorphism (SNP). Molecular markers selected for marker trait studies should employ a simple and inexpensive assay to detect polymorphism between genotypes (Vignal et al. 2002).

Restriction fragment length polymorphism (RFLP)

RFLPs are different sizes of fragments produced upon restriction by a restriction enzyme in different genotypes. After restriction, the DNA is electrophoresed on the agarose gel which separates the fragments of different sizes. These fragments are hybridized with probes and then visualized by autoradiography. The differences in the fragment size could be because of point mutations, deletions, insertions or transpositions. RFLP markers are co-dominant in nature since they can differentiate between heterozygotes and dominant homozygotes (Tanksley et al. 1989).

Random amplified polymorphic DNA (RAPD)

RAPD markers are based on the PCR-amplification of DNA segments with arbitrary random primer pairs. When a particular primer pair amplifies a genomic segment in one genotype and not in the other, a polymorphism is identified. RAPD markers are classified as a dominant marker class as they cannot differentiate between heterozygotes and dominant homozygotes (Williams et al. 1994).

Microsatellites or simple sequence repeats (SSRs)

SSR markers are sets of tandemly repeated DNA sequences. The repeated sequence may be a set of two, three, four or more nucleotides (Tautz and Renz 1984). The regions flanking each locus with nucleotide repeats are unique and primers are designed to amplify the intervening SSR in different genotypes. The amplification product is electrophoresed on the agarose gel or polyacrylamide gel. The length of a particular genomic region amplified in different genotypes depends upon the variability in the number of nucleotide repeats. This variability at a particular genomic location in different genotypes is an indicator of polymorphism. SSR markers are co-dominant in nature (Morgante et al. 1994).

Amplified fragment length polymorphism (AFLP)

AFLP markers employ the ability of restriction enzyme to restrict a genomic region along with polymerase chain reaction (PCR) to amplify that region after restriction. They detect the presence or the absence of a restriction fragment (Vos et al. 1995).

Single nucleotide polymorphism (SNP)

SNP refers to a single base pair change (point mutation) at a particular position in the DNA sequence, i.e., at each position any of the four nucleotide bases may be present (Vignal et al. 2002). SNPs are generally bi-allelic and co-dominant and their identification require prior DNA sequence data (Vignal et al. 2002). They are usually present in the non-coding regions of the genome, however, when present in coding, promotor or enhancer regions involved in gene expressions, they may or may not result in phenotypic difference among different genotypes. The distinguished quality of SNP molecular markers is their ability to genotype hundreds or more SNP in a large population through the high throughput genotyping technologies like Taqman, single-base extension based assay, MALDI-TOF mass spectrometry based systems, the invader assay and pyrosequencing etc. (Tsuchihashi and Dracopoli 2002). These genotyping methods employ different kinds of allele-specific discrimination methodology like differential hybridization, primer extension, ligation, allele-specific probe cleavage and methods of signal detection (Kwok 2001). SNP analysis has the added advantage that using Golden Gate Illumina Assay, an integrated SNP genotyping system, it offers a robust >1500 multiplexing assays on a bead array platform (Michael et al. 1998; Gunderson et al. 2004).

Soybean and molecular markers

RFLP markers were the first molecular markers used to construct the first genetic map of soybean. Keim et al. (1990) constructed F_2 segregating population from a cross between cultivated and wild soybean, and constructed a soybean genetic map. In this study, 150 RFLP markers covered 1200 cM and the genetic map included twenty six genetic linkage groups. The duplicated nature of soybean genome results in multiple banding patterns on hybridization (Shoemaker and Specht 1995).

Morgante and Olivieri (1993) studied the feasibility of using microsatellites as markers in plant genetics using cultivated and wild annual soybean. G. *soja* showed low level of diversity for the first generation RFLP markers. Presence of a large number of variations for both the dinucleotide and the trinucleotide SSR in soybean ascertained the significance of SSR in the construction of the linkage map. Other studies also confirmed that SSR markers are highly polymorphic and show a random distribution across the soybean genome with approximately 26 alleles per locus (Rongwen et al. 1995; Maughan et al. 1995; Powell et al. 1996). Cregan et al. (1999) constructed a genetic map of soybean with 606 SSR markers in which each marker mapped a single locus in the genome. Grimm et al. (1999) estimated the frequency of SNP in soybean to be 3.4 per kilobase in 18,000 bases of DNA sequence studied in 18 genotypes. According to Zhu et al. (2003) the frequency of SNP is two folds higher in the untranslated regions (UTRs), introns, and genomic regions close to coding sequence.

Choi et al. (2007) reviewed earlier studies which used SSR markers for mapping and observed that while a fairly extensive set of 1000 genetically mapped SSR markers was available to soybean breeders and geneticists, the current map had 138 gaps of 5 cM or more in which no SSR marker was present. Out of the 138 gaps, 26 were 10 cM wide. They doubted the efficacy of SSR markers in mapping since these regions or gaps having an absence of or a low SSR marker density may be gene rich. They constructed the first genetic transcript map of soybean. This map involved genetic mapping of one SNP in each of the 1141 genic regions in one or more of the three recombinant inbred mapping populations.

IDC related QTL discoveries

QTL discovery is facilitated by two types of genetic mapping studies: (i) linkage mapping using a bi-parental mapping population, and (ii) association mapping using natural populations or germplasm collections. The two genetic mapping studies are similar as they use recombination events within the genome and correlate polymorphic molecular markers with phenotypic variations. However, they differ in the number of recombination events occurring at every locus.

Bi-parental mapping and IDC QTL discovery

It involves making crosses between two parental genotypes which differ in the quantitative trait of interest. The two parental genotypes are screened with molecular markers to identify polymorphic ones. The segregating populations such as F_2, backcross (BC), double haploid (DH), recombinant inbred lines (RIL) and near isogenic lines (NIL) are generated and screened with the polymorphic markers (Collard et al. 2005). Polymorphic molecular markers in the segregating populations are ordered into linkage groups with their relative genetic distance (in cM) based on the recombination rates between the marker loci. These are then statistically correlated with the phenotypic trait and the location of QTL is identified between two marker loci (Abdurakhmonov and Abdukarimov 2008).

Diers et al. (1992) identified three markers out of the 272 mapped RFLP markers linked with QTL for Fe-efficiency in 13 F_2-derived lines developed from a cross between G. *max* (Fe-inefficient) and G.*soja* (Fe-efficient). The observation was significant ($P<0.01$) and it explained 31%, 25% and 17% of the phenotypic variations. However, these linkage associations were not reproducible in a second tester population.

Lin et al. (2000) mapped QTL for Fe-efficiency in two bi-parental crosses of *G. max* x *G. max*. The pride population was developed from a cross between Pride B216 (Fe-inefficient) x A15(Fe-efficient), and the Anoka population was developed from a cross between Anoka (Fe-inefficient) x A7 (Fe-efficient). The populations were scored visually and the chlorophyll concentration was measured in the laboratory. Ninety RFLP and ten SSR markers were used to construct the linkage map in Pride population of 120 F_2 plants. One morphological (hilum color) marker, eighty-two RFLP markers and fourteen SSR markers were used to construct the linkage map in Anoka population of 92 F_2 plants. Replicated field trials were conducted with the F_2 derived lines using randomized complete block design. Using the interval mapping method, a multi-genic model of inheritance was observed in the Pride population. Two QTL were mapped on chromosome 14 and one QTL was mapped on chromosomes 3 and 18. These QTL individually explained 7.7 to 10.8% of the phenotypic variations observed in IDC visual scores. These findings confirmed the observations of Cianzio and Fehr (1982) regarding the multiple gene action for Fe–efficiency. In the Anoka population, two QTL were mapped on linkage groups 3 and 5 and this explained 35.2% (LOD score =13.1) and 72.7% (LOD score =7.3) of the total IDC phenotypic variations. QTL on linkage group 3 was regarded as a major region because it was mapped with a high LOD score and contributed to a large number of phenotypic variations. This observation confirmed the findings of Cianzio and Fehr (1980). Markers from the Anoka population were evaluated in the Pride population and vice-versa to enhance the accuracy of the markers for MAS, however, neither of the markers could be scored in the other population.

Charlson et al. (2003) developed F_2 and $F_{2:4}$ populations using parent A97-770012 (Fe resistance and moderate yield) and Pioneer 9254 (P9254) (moderate resistance and superior yield). On calcareous soils, at two locations in Iowa, chlorosis scores were evaluated for parents and for F2:4 derived population in replicated field trails while the genotypic determination was conducted on F2 lines using SSR markers. Single factor analysis of variance identified that three SSR markers: Satt211, Satt481 and Sat104 were associated with IDC ($p<0.1$). Satt481 was the only marker which showed marker-trait association at both the locations, showing 12% of the total phenotypic variations for IDC resistance. Hence, it was concluded that Satt481 might serve as an indirect selection for the IDC QTL.

Charlson et al. (2005) used a population with acceptable traits but moderate IDC resistance. The mapping population was developed by crossing a high yielding cultivar, Pioneer 9254, and an advanced experimental line, A97-770012. Out of the 108 SSR markers previously identified, only 22% (24 markers) were polymorphic in the parents. The genotyping was performed on the F_2 lines, while the F_2-dervied lines ($F_{2:4}$ and $F_{2:5}$) were used to evaluate IDC scores. Three markers, namely, Satt211 (mapped on chromosomes, 5), Satt481 (mapped on chromosome, 19) and Sat104 (mapped on chromosome 20) showed association with IDC ($P \leq 0.5$) with r^2 value ranging from 3.9 to 11.5%. The objective of the study was to check the potential of Satt481 in MAS for early detection of IDC (Charlson et al. 2003). In this environment-independent study, Satt481 was the only marker which was consistently associated with IDC. Even though IDC scores were slightly better, QTL on chromosome 3, which explained approximately 70% of the phenotypic variations, went undetected.

Even though past researchers had discovered some major and minor QTL associated with IDC trait, none of the QTL-flanking markers discovered in one population have so far been reproduced in another population. This questions the efficiency of using these markers in MAS for quantitative traits. The economic value of Satt481 in MAS is dependent on its effectiveness as a detector of IDC trait in other breeding populations.

Association mapping in natural populations and QTL discovery

An alternative approach to bi-parental mapping for marker-trait co-relation is association mapping which discovers the marker-trait association with the help of linkage disequilibrium (LD) pattern in a large population (Risch 2000). Unlike linkage mapping in which the distance between two alleles at different loci is measured in cM, LD is defined as non-random associations between alleles at more than one loci and is a measure of physical distance between two alleles in a population (Flint-Garcia et al. 2003).

The following steps are involved in association mapping: (i) selecting a group of individuals from a natural population or germplasm collection representing the wide range of phenotypic variations, (ii) genotyping the population with molecular markers and phenotyping them for the traits using replicated trials in different environments, (iii) estimating LD decay with the help of molecular markers, (iv) estimating the population structure (the level of genetic differentiation among the groups in the population) and kinship (coefficient of relatedness between pairs of each individual within the population), and (v) applying appropriate statistical methods to identify marker loci in close proximity to the QTL of interest (Balding 2006).

Advantage of LD Based Association Mapping

There are several advantages of LD based association mapping. First, there is no pedigree or cross required as variations in the traits are studied not by the multiple segregation of loci between two the genotypes, but by the multiple segregation of loci in the entire population. Second, in a population of unrelated individuals, many rounds of recombination between alleles occur over several generations. In such a population, correlation between the QTL affecting the trait and the molecular markers closely linked to QTL will be retained, and the resolution of finding a marker in close proximity to QTL is higher (Mackay and Powell 2007). On the other hand, bi-parental mating results in F_2 population and recombinant inbred line (RIL). In a F_2 population a less amount of recombination has occurred while in a RIL population, homozygosity is achieved after a few generations. Third, multiple alleles for a locus can be identified in a random mating population, whereas in a bi-parental population there are just two segregating alleles for a single locus (Balding 2006).

Generation of LD in a population

There are two processes which regulate the pattern of LD decay observed in a population: (i) mutations which give rise to new alleles that might be linked to the QTL regions and (ii) recombinations which break the linkage between the new allele and the QTL region. The rate of LD decay has been reviewed by Hamblin et al. (2005) as a parameter of the rate of mutation (u) as $4N_e$u or Θ and as a parameter of the rate of recombination (r) as $4N_e$r or ρ, where N_e (effective population size) is a parameter of the historical size of the population, population structure, and mating system. When a random mating population is in equilibrium, LD is a simple function of Θ/ρ. However, equilibrium is just a concept, since, in nature a population is affected by a number of factors such as selection, genetic drift and mating nature of the population. Hence, LD decay is neither constant throughout the genome (Nordborg 2002), nor constant within the same genomic regions across multiple populations (Hyten et al. 2007).

Pattern of LD decay

LD decay occurs as follows: First, recombination shuffles the regions of chromosomes and the alleles located on the same chromosome. Recombination frequency depends on the degree of polymorphism among the homologous chromosomes and only the tight linkage between alleles in LD persists after several generations of recombination. Frudakis (2008) mentioned that under no selection pressure, the extent of LD decay over several generations is a function of the recombination rate between polymorphic markers and time. He summarized his observations in the following equation:

> $\Delta D = (1\text{-}r)^t$ *(where ΔD is the rate of LD decay, r is the recombination rate that is the function of the genetic distance between polymorphic markers, and t is the number of generations)*

Second, the range of LD depends on the population history of a crop species, i.e., population genetic bottlenecks followed by the geographical expansion and subpopulations in which it is measured (Rafalski and Morgante 2004). According to population genetics, a population is a group of individuals who can freely mate and hence there is no restriction to gene flow. A population has subpopulations derived from the founder individuals and gene flow between the subpopulations may be limited (Wright 1951). There are two sources of genetic variations among the newly formed subpopulations: (i) gene pool of the founders and (ii) new mutations, specific to a subpopulation, which result in a genetic drift, i.e., the random change in the allele frequency. In a large population, changes in allele frequency due to drift are small, but in a small population allele frequency may change drastically (Ellstrand and Elam 1993). A limited number of founders imply fewer multi-loci combinations of alleles on the chromosomes. Mating among individuals in a sub-population leads to homogenous genomic regions and as a result, a large extent of LD persists. Third, the mating system (selfing/outcrossing) also determines the genetic variations within a sub-population. The amount of gene flow in a selfing species is smaller than that of an outcrossing species. Therefore, a greater number of differences would be observed among subpopulations of the selfing species than among subpopulations of an outcrossing species. Fourth, population admixture results when two or more subpopulations are mixed together and allowed to interbreed. The different subpopulations have different allelic frequency at many loci (Flint-Garcia et al. 2003). This leads to changes in allele frequencies in the resulting admixed population, causing the generation of LD in previously unlinked loci. Random mating for several generations in the newly formed admixed population results in the breakage of LD (Flint-Garcia et al. 2003). Fifth, selection at a particular locus is expected to decrease the genetic diversity (same allelic combinations at different loci), thereby increasing LD in the surrounding region, a phenomenon known as selective sweep. During selection, LD is a function of recombination rate and distance (Morrell et al. 2005). Sixth, the introduction of new mutations can disrupt the LD between pairs of alleles. Any new mutation will be in LD with the genetic region (alleles) on the chromosome responsible for the trait of interest. Generations of crossing-over will generate LD between the old pairs of alleles and the new mutant allele (Mackay and Powell 2007).

LD studies in plants

The knowledge of the extent of LD decay is important for conducting association mapping in any plant species (Flint-Garcia et al. 2003). It is also important as recombination rates are not uniform across the physical distance (Gaut and Long 2003). LD decay studies have been conducted in many plant species such as Arabidopsis (Nordborg et al. 2002 and 2005), maize (*Zea. Mays ssp. mays*) (Thornsberry et al. 2001; Tenaillion et al. 2001; Remington et al. 2001; Jung et al. 2004), barley (*Hordeum vulgare subsp. vulgare*)

(Stracke et al. 2003; Morrell et al. 2005; Caldwell et al. 2006), rice (*Oryza. sativa*) (Garris et al. 2003; Rakshit et al. 2007), sorghum bicolor (Rooney and Smith 2000; Hamblin et al. 2005) and soybean (Zhu et al. 2003; Hyten et al. 2007).

Association mapping in a structured population

The desired cause of LD is physical linkage, however, LD can be present in a population due to the population structure. Population structure or stratification is an important factor which contributes to Type I error (declaring a false positive association). Population structure creates LD between unlinked loci. In a homogenous (unstructured) population, molecular markers associated with the putative QTL for the trait can be inferred by studying differences in the marker-allele frequencies among genotypes showing variation. However, a structured population (presence of subpopulations within a population) may have different allele frequencies and any such difference might be in LD with the other alleles in that population.

Population structure in soybean

Soybean genotypes adapted to the northern US belong to the maturity groups 00, 0 and I. Different soybean breeding programs aimed at improving the agricultural qualities of soybean in this region must exploit the variations present in a small sub-set of soybean population. In the different soybean breeding programs dedicated to improve the IDC resistance, breeders continually select breeding lines which show a good IDC score. This results in soybean lines having more differences in allele frequencies at a few gene loci responsible for the phenotype as compared to the differences in allele frequencies in other regions. Since private companies do not share pedigree information about their lines, various lines used in the association mapping may have the same pedigree although they may have been taken from different sources.

Statistical methods to avoid false positive associations due to population structure

Pritchard and Rosenberg (1999) and Pritchard et al. (2000) developed two model-based approaches to define populations with the help of unlinked markers: (i) no-admixture model, and (ii) admixture model. A package called STRUCTURE was developed by Pritchard et al. (2000) to perform association mapping for a structured population. The two models utilize the multi-locus genotypic information collected for a population to estimate the proportion of subpopulation membership. Both models assume that unlinked markers provide independent information about the individual's ancestry (Pritchard et al. 2000). The no admixture model assumed that individuals are derived from one of the K populations with no gene flow between the populations (Pritchard et al. 2000). The no-admixture model however is not accepted as individuals generally have had some common ancestor in more than one population. The second, admixture model, assumed that there is a certain amount of gene flow between the populations and individuals have mixed ancestry (Pritchard et al. 2000). This is described as admixture LD. The linkage model uses a Markov chain Monte Carlo (MCMC) method to accurately estimate the number of subpopulation, the allele frequency and the variations in ancestry (Falush et al. 2003). Once individuals are organized into separate subpopulations, marker-trait associations within each subpopulation are determined using general linear models (GLM) and mixed linear model (MLM) analyses. The linkage model introduced by Falush et al. (2003)

extends the admixture model and accounts for correlation between linked markers on the same chromosome.

Yu et al. (2006) proposed the MLM approach to avoid false positives (Type I error). The unified mixed linear model gathers information from the random selected marker, the subpopulation (Q-matrix) obtained from Structure, and (the) relative kinship (K-matrix) obtained from the SPAGeDi. The kinship matrix, generated using SPAGeDi software, estimates the identity by descent between individuals by adjusting the probability of identity by state between two individuals with the average probability of identity by state between randomly selected individuals (Yu et al. 2006). The independent variables included in the MLM are considered as covariates in the regression model to correlate the relation between genotype and phenotype. They studied three traits, namely, flowering time, ear height and ear diameter in a diverse set of 277 maize inbred lines using different association models. The Q+K model gave the best distribution of the p-values as compared to K-model, the Q-model and the naïve model in which no population structure data or kinship matrix was taken into consideration. Yu et al. (2006) observed that without the correction of population structure, the distribution of p-values was skewed towards significance, a strong indication of type I error.

Zhao et al. (2007) found an artifact in the method of Kinship matrix estimation based on identity by descent and identity by state. In the absence of mutations, two genes identical by descent (IBD) have the same nucleotide sequence, while two homologous genes with the same nucleotide sequence are not necessarily IBD. Identity by state (IBS) implies IBD for markers with low mutation rates (like SNP) (Zhao et al. 2007). They estimated the fraction of shared fragment haplotypes from the population under study to determine the kinship matrix.

Another effective way to control population structure, other than the Q-matrix developed by SURUCTURE, is the principal component analysis (PCA) (Patterson et al. 2006; Price et al. 2006). It was initially applied in association mapping studies conducted in humans to infer worldwide axis of human variation from the allele frequencies of various populations. PCA reduces the original markers to a minimum number (component variables) to explain the variations observed in a population. These principle components explain the unobserved subpopulations variation from which individuals have originated. PCA matrix can be generated in SAS faster as compared to the complex STRUCTURE algorithm.

Association mapping for IDC in soybean

Wang et al. (2008) conducted association mapping in two independent advanced breeding populations of soybean using 24 SSR molecular markers (20 random and 4 IDC polymorphic markers). Populations were grown at 3-4 sites located in North Dakota in the year 2002 and 2003. Population structure (presence of subpopulations, where loci are in Hardy-Weinberg equilibrium and linkage equilibrium) estimated from STRUCTURE software identified five subpopulations in the two association mapping populations. The kinship coefficients (proportion of familial relatedness between individuals) were calculated by the procedure described by Loiselle et al. (1995) in SPAGeDi. Different statistical tests were also conducted to gain confidence in the identified putative marker-trait associations. Statistical tests such as single factor analysis (SFA), general linear model (GLM) with population structure (Q) information, mixed linear model(MLM) with kinship matrix (K) information model, and another mixed linear model (Q+K) with both population structure and kinship matrix information was used to detect significant maker-trait associations with averaged IDC ratings. The four identified SSR markers, namely, Satt020, Satt114, Satt199, and Satt239 were associated with IDC at different locations. However, only two SSR markers, namely, Satt114 mapped on chromosome 13 and Satt 239 mapped on chromosome 20 in the 2002 population were detected in the second confirmation population of the year 2003.

MATERIALS AND METHODS

Materials

Association mapping populations

The two populations under study comprised of two independent advanced soybean breeding lines developed from different public and private breeding programs in Northern states of US during the years 2005 and 2006. These lines were unique except for a few standard lines. The two populations consisted of 143 and 141 soybean plants/lines for the years 2005 and 2006.

Phenotypic analysis

The 2005 population was grown at five sites near Arthur, Ayr, Chaffee, Colfax and Galesburg in North Dakota (ND). The soil at these sites had a pH varying from 7.8 to 8.1, salinity (EC) from 4.0 to 1.9 mmho/cm and $CaCO_3$ contents ranging from 2 to 11%. Thirty five seeds were planted in 5inch rows on 30-inch centers. The experimental design was a randomized complete block design with four replications at each site. Two visual observations were made at each location at the 2-3 and 5-6 trifoliolate stages. The visual ratings were ranked on a scale of 1-5, where 1= no chlorosis and plants were normal and green; 2= a slight yellowing of the upper leaves with no differentiation in color between the leaf veins and interveinal areas; 3= interveinal chlorosis of the upper leaves (veins were green and interveinal area was chlorotic) without any obvious stunting of growth or death (necrosis) of leaf tissue; 4= interveinal chlorosis of the upper leaves with some apparent stunting of growth or necrosis of plant tissue; and 5= severe chlorosis with stunted growth and necrosis of the young leaves and plant death in some instances. Ten standard varieties were also planted. These control varieties, listed here in descending order of their IDC resistance were: Iowa State ISU A11, Seeds 2000 2070RR, Traill, Council, Asgrow0801 and Peterson PFS 0202RR , Glacier, and Mycogen 5072,Stine 0480 and NuTech 0505RR.

The second independent population was grown in 2006 at five different locations at Arthur, Colfax, Galchutt, Galesburg and Prosper, ND. The soil at these sites had pH varying from 8.1 to 8.3, salinity (EC) from 0.2 to 0.8 mmho/cm and $CaCO_3$ contents ranged from 2 to 8%. The experimental design and the IDC rating scales were the same as that for the year 2005. The visual observations were made at 2-3 trifoliolate and 5-6 trifoliolate stages, and also two weeks later. Only two observations could be made at Prosper due to recovery of the plants from chlorosis. The crop at Galchutt site was lost due to an unusually heavy white grub infestation. The same 2005 standard lines were used as a control for this experiment. Dr. Jay Goos and his research group at the Department of Soil Science, North Dakota State University, performed the field experiments and IDC visual ratings.

DNA isolation

For DNA extraction, the varieties under study for the year 2005 and 2006 were grown in greenhouse, and the young leaves were harvested and stored at -80 °C. Two to three gram of the frozen leaf tissue was later ground in liquid nitrogen in a chilled mortar. The powdered leaf tissue was then transferred to a 50ml plastic capped centrifuge tube and preserved again at -80 °C. 10 ml of preheated CTAB (Doyle and Doyle, 1990) isolation buffer was added to each tube and mixed gently. The tubes were then incubated in a water bath at 60°C for 30 minutes during which they were shaken every 10 minutes. Following incubation, 10 ml of chloroform: isoamylalcohol (24:1) was added to each tube and the contents were centrifuged for 15 minutes at 3500 rpm. The aqueous phase from each tube was transferred to another clean 50 ml centrifuge tube. To precipitate the DNA, cold isopropyl alcohol was added to the tubes in 1:1 ratio with the aqueous phase. The tubes were kept overnight at 4^{o}C for maximal precipitation of the DNA and the next day, the samples were again centrifuged for 15 minutes at 3500 rpm to obtain a DNA pellet. The DNA pellet was washed twice with 70% ETOH and then dried. This dried DNA pellet was then re-suspended with ~ 400ul TE with RNase (Doyle and Doyle, 1987; Doyle and Doyle, 1990).

SNP genotyping: selection of SNP and golden gate assay technology

SNP molecular markers were designed from the Universal Soy Linkage Panel 1.0. 1536 genome-wide SNPs were scored using the Illumina's Golden Gate assay technology at Beltsville Agricultural Research Center-West, MD under the guidance of Dr. Perry Cregan.

Statistical Analysis

Genetic diversity analysis

Out of the 1536 SNP markers studied in the two populations, 1265 markers were found to be polymorphic. A subset of 881 SNP markers for the 2005 population and 913 SNP markers for 2006 population with a minor allele frequency (MAF) over 10% estimated using PowerMarker (Liu and Muse, 2005) were selected for analysis. On combining these 881 and 913 SNP markers, 847 markers were common between the two populations. Polymorphic information content (PIC) and the total number of alleles for each SNP locus were estimated separately for each population.

The expected PIC values were calculated as, PIC = 1- Σ $(P_i)^2$, where P_i is the proportion of the population carrying the i^{th} allele. It measures the polymorphism for the marker loci, a value ranging between 0 (monomorphic markers) to 1 (highly polymorphic marker). It is calculated by taking into consideration the number of alleles present and the relative frequency of each allele at that particular locus.

SNP marker imputation

FastPHASE 1.3 (Scheet and Stephens 2006) was used to impute missing 881 and 913 SNP marker loci in the two populations using the “likelihood” based imputation. The imputation was included to avoid eliminating individuals with missing loci in association analyses. The software clustered together similar haplotypes characterized by similar allele frequency. For a missing allele at a marker locus, the probability of it being one allele or the other is estimated as a function of the haplotype cluster origin and the allele frequency for the marker in each cluster (Scheet and Stephens, 2006). We used the default parameters for the analysis.

Pairwise linkage disequilibrium and LD decay

The extent of LD was estimated as a square of allele frequency correlation estimates r^2 for the 2005, 2006 and the combined 2005 and 2006 population data in TASSEL *ver*.2.1 (http://www.maizegenetics.net). r^2measures the proportion of sample variance explained by the presence of the polymorphic allelic state at two polymorphic loci

LD decay determines the resolution of the association mapping. LD decay graphs were plotted with genetic (cM) and physical distance (bp) on the x-axis and r^2 on the y-axis for each marker pair locus located on the same chromosome using a nonlinear regression described by Remington et al. (2001). The expected decay of LD was estimated according to the following equation,

$$E(r^2) = [\frac{10+pd}{(2+pd)(11+pd)}][1+\frac{(3+pd)(12+12pd+(pd)^2)}{n(2+pd)(11+pd)}]_n$$

The above equation was described by Pyhajarvi et al. (2007), where n denotes the number of sequences $\rho = 4N_e c$ between adjacent sites d is the distance between the two sites of a pairwise comparison and *c* is the recombination rate(Hill and Weir 1988). We fitted this equation into a non linear regression using NLIN procedures in SAS *ver.* 9.1.3 ® (SAS Institute, Cary, NC). The analyses were performed for individual chromosomes in both the populations.

Population structure

Estimation of population structure and ancestral (kinship) relationship were derived using a set of markers that had pairwise $r^2<0.5$. In the 2005 population 312 marker loci, in the 2006 population 356 marker, and in the combined population 334 marker loci met this criterion.

STRUCTURE *ver* 2.2 was used to estimate the subpopulation membership of the different lines in these two populations individually (Pritchard et al. 2000). Structure analysis was conducted with 312 SNP markers for the 2005 mapping population and 356 SNP markers for the 2006 mapping population. STRUCTURE was run using the linkage model with correlated allele frequency (Pritchard et al. 2000). The program was run with a burn-in-length of 100,000 Markov chain Monte Carlo (MCMC) and 500,000 iterations for estimating the parameters for each independent population. The optimal number of subpopulation (K) was set from 1 to 15 with 10 runs. For every K the posterior probability of the individuals was estimated. To find the best number of K, Wilcoxon two- sample test was used to compare adjacent sub populations (K1 vs. K2, K2 vs. K3 and so on) using NPAR1WAY procedure in SAS 9.1.2 as described by Wang et al. (2008). The smaller K value in a pairwise comparison for the first non-significant p-value was chosen as the best number of sub populations.

Principle component analysis (PCA) was also used to control for population structure in the two populations individually, and in the combined data for 2005 and 2006. Those principal components (eigenvectors/combination of SNP markers) which collectively explained 50% of the variation present in both the populations and in the combined population were selected to minimize the false marker-trait associations. 312 and 356 SNP markers from the 2005 and the 2006 population respectively, and 334 SNP markers for the combined 2005 and 2006 populations were used for the PCA performed in SAS *ver*.9.1.3 using PRINCOMP procedure.

Population kinship

A pairwise Kinship coefficient matrix (K-matrix) which estimates the probability of recent co-ancestry between genotypes in the 2005 and 2006 mapping populations was developed in SPAGeDi 1.2 (Hardy and Vekemans 2002) using the formula as follows:

$F_{ij} = (Q_{ij} - Q_m)/ (1\text{-}Q_m) =\sim \Theta_j,$

Here Θ_{ij} is the pairwise kinship coefficient, F_{ij} is an estimator of the coefficient, Q_{ij} is the probability of the identity by state between random loci for genotypes i and j, and Q_m is the average probability of identity by state for loci from random genotypes in the population used to draw i and j (Loiselle et al. 1995). The F_{ij} was calculated for all pairwise combinations for 143 genotypes in 2005 and 141 genotypes in 2006 populations. Negative values for the kinship matrix were set to zero as described by Yu et al. (2006).

The second kinship coefficient (K^*- matrix) (Zhao et al. 2007) was estimated as the proportion of shared alleles for all pairwise comparisons for the individuals in the 2005, 2006 and the combined 2005 and 2006 populations using PowerMarker.

Marker-trait association model-testing

Nine different linear regression models were tested for marker-trait association using the MIXED procedure in SAS 9.1.3 (Table 1). Six mixed-linear models (MLM) considered both fixed and random effects while the remaining three general linear models (GLM) considered only the fixed effects. The performance of various models was tested by plotting a probability-probability (P-P) plot with observed p-values on the x-axis and cumulative p-values on the y-axis to accurately correct for the Type I error. The uniform distribution of the p-values for all the nine models helped in the assessment of the models which best controlled the type I error.

In these models, y is a vector for phenotypic observations, α signifies the fixed effects related to the SNP marker, β is a vector of the fixed effects related to the population structure, v is a vector of the random effects related to the relatedness among the individuals, and e is a vector of the residual effects. X is genotypes of the SNP markers, P

is the matrix of the principle components, i.e., 13 principal components for 2005 and 12 principal components for 2006 populations, K is the kinship matrix developed in the SPAGeDi *ver* 1.2, and K* is the kinship matrix developed in PowerMarker. The variances of the random effects were estimated as Var(u)= $2KV_g$ and Var(e)= IV_R, where K is a kinship matrix, I is an identity matrix with off-diagonal elements being 0 and diagonal elements being reciprocal of the number of the observations for which the phenotypic data was obtained, V_g is the genetic variance and V_R is the residual variance.

Table 1. Summary of the regression models used for plotting the P-P plot

Model	Statistical model	Information captured in the model
Naïve	$y= X\alpha +e$	y is related to X, without any correction for Q or K
K	$y= X\alpha + Kv + e$	y is related to X, along with K-matrix estimated in SPAGeDi
K*	$y= X\alpha + K^*v + e$	y is related to X, with K*-matrix estimated in Powermarker
Q	$y= X\alpha + Q\beta + e$	y is related to X with Q-matrix generated from STRUCTURE
PCA	$y= X\alpha + P\beta + e$	y is related to X with PCA instead of Q-matrix,
Q+K	$y= X\alpha + Q\beta + Kv + e$	y is related to X, along with Q-matrix and K-matrix
Q+K*	$y= X\alpha + Q\beta + K^*v + e$	y is related to X, along with Q-matrix and K*-matrix
PCA+K	$y= X\alpha + P\beta + Kv + e$	y is related to X, along with PCA-matrix and K-matrix
PCA+K*	$y= X\alpha + P\beta + K^*v + e$	y is related to X, along with PCA-matrix and K*-matrix

Marker trait association analysis

Three different models were used for marker-trait association in the 2005 population. These comprised of the PCA, a GLM approach, and PCA+K and PCA+K*, both of which are MLM approaches. In the 2006 population, five different models were studied for marker-trait association analysis. These included Q and PCA, which are GLM approaches, and Q+K, Q+K* and PCA+K*, which are MLM approaches. For the combined 2005 and 2006 population, two different models, PCA +K* and PCA were studied for marker-trait association analysis.

For all the above described GLM and MLM analyses, a total of 881 and 913 marker-trait pairs for the 2005 and the 2006 populations respectively, and a total of 847 marker-trait pairs for the combined 2005 and 2006 population were evaluated in SAS using the MIXED procedures (SAS Institute 1999). F-test with the denominator degree of freedom as determined by the Satterthwaite method was implemented to assess the importance of the marker effect of each marker-trait pair as described by Weber et al. (2009). For the selected model in each population, the positive false discovery rate (pFDR) was estimated using the MULTTEST procedures in SAS 9.2 to correct for multiple marker trait association. A union dataset for each of the population is based on a Q-value (pFDR) of 0.1 or less for each model (Weber et al. 2009).

RESULTS

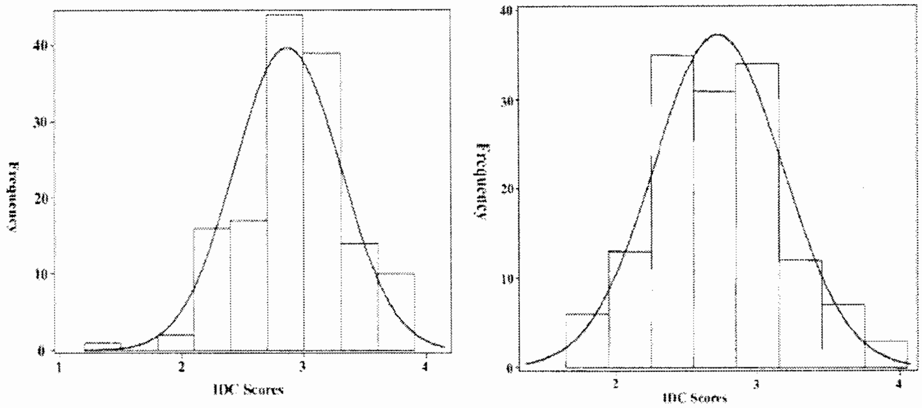
Phenotypic Analysis

IDC Scores in the two soybean populations

IDC, a quantitative trait, is controlled by both environmental and genetic factors. For the present study, field observations were made for IDC because they best mimic production environments. The correlation coefficients of visual IDC ratings averaged across different locations and sites indicated that the different ratings were highly correlated with each other at each location in the 2005 and the 2006 populations. The average IDC rating for each line was used for further data analysis. The visual IDC scores for the 2005 population ranged from 1.5 to 3.8 with an average of 2.9, while the scores for the 2006 population ranged from 1.6 to 3.8 with an average of 2.7.

The distribution of IDC scores in the two populations under study, (Figure.1a and 1b) were tested for normality using the Kolmogorov-Smirnov (KS) Test at a significance level of $p < 0.05$. The null hypothesis is that the phenotypic data is normally distributed. The KS values were found to be 0.07 and 0.09 for the 2005 and the 2006 independent populations respectively. Since the observed p-values were greater than the significance level ($p < 0.05$), the null hypothesis was accepted, and it was concluded that the phenotypic data was normally distributed for both the populations.

The variance in chlorosis scores for the 2005 and the 2006 populations showed a significant location effect, as well as a significant line by location interaction effect (Table 2). This shows that the environment influenced the IDC scores. Broad sense heritability on an entry mean basis was also deduced from the analysis of variance. The broad sense heritability values were 0.99 for 2005 population and 0.97 for 2006 population. These values demonstrate the demonstrate the consistency of the IDC rating.



(a) (b)

Figure 1. Normal distribution of IDC scores for the individual soybean lines for the year (a) 2005, and (b) 2006.

Table 2. Analysis of variance mean squares value for IDC ratings for the two soybean populations grown at different locations

	Populations			
	2005		2006	
Source of variation	Df	MS	df	MS
Location	4	431.16***	3	535.88***
Line	143	17.37***	140	12.99***
Location x line	568	1.06***	420	1.33***
Replication/location	12	3.19	9	3.08
Error	2322	0.84	1705	0.71

***Significant difference $p \leq 0.001$

Genotypic Analysis

SNP marker analysis

SNP marker information was collected in each population at 1265 loci using the Illumina Golden Gate Assay technology. Out of the 1265 SNP marker loci, 881 markers in the 2005 population and 913 markers in the 2006 population had a MAF > 10%. The Wilcoxon two-sample test was not significant (p = 0.3439) for comparing the major allele frequency in the two populations. The expected heterozygosity is generally low for SNP markers owing to their bi-allelic nature and the selfing nature of *G. max*. Gene diversity for the 2005 genotypes ranged from 0.1301 to 0.500 with an average of 0.38, and for the 2006 genotypes it ranged from 0.1195 to 0.500 with an average of 0.37. The markers in both populations were polymorphic with PIC values ranging from 0.1216 to 0.3750 for the 2005 population and from 0.1124 to 0.3750 for the 2006 population.

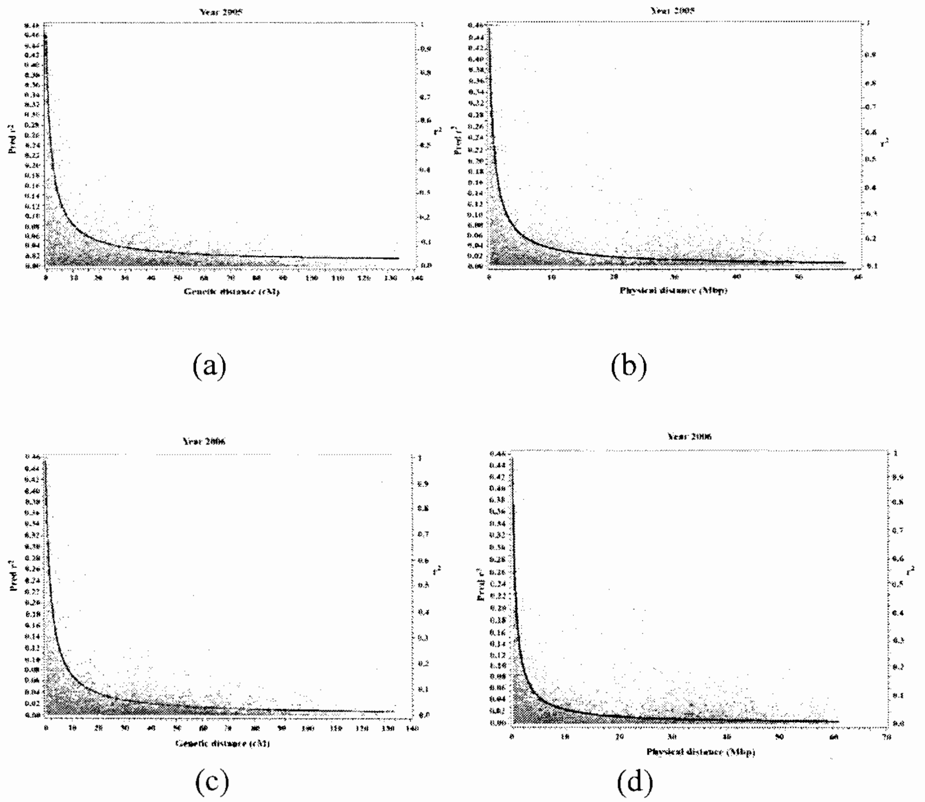
Initially, marker-trait associations were detected in both the 2005 and the 2006 populations and subsequently the genotypic and the phenotypic data for these populations were combined for further analysis. Because the breeding lines may have been related, it was imperative to control for population structure of the two independent populations and the combined 2005 and 2006 population dataset.

LD decay estimation (2005 and 2006 populations)

The non-linear regression model for estimating the decay of LD with distance was determined using a genome-wide LD decay graph. Using a pair-wise analysis for all the 881 and 913 SNP loci in the 2005 and the 2006 populations respectively, LD decay graphs were plotted with r^2 values on the y-axis, and with genetic distance in cM and physical distance in Mbp on the x-axis (Fig 2). The average decay of LD in terms of physical distance declined to an $r^2 < 0.05$ within 7 Mbp and 5 Mbp in the 2005 and the 2006 populations respectively. The average decay of LD in terms of genetic distance declined to an $r^2 < 0.05$ within 20 cM in the 2005 population and within 12 cM in the 2006 populations.

Population structure and kinship analysis

Two measures, r^2 and D' were used to estimate LD. While D' gives an estimate of the historical recombination by taking into account the allelic association, r^2 is the squared correlation coefficient between loci based on allele frequencies and is influenced by the mutations.



(a) (b)

(c) (d)

Figure 2. Genome-wide LD decay plot for the two populations. LD, measured as r^2, between pairs of polymorphic marker loci is plotted against the gentic distance (cM) and physical distance (Mbp) between the loci: (a) 2005, cM distance vs. r^2, (b) 2005, Mbp distance vs. r^2, (c) 2006, cM distance vs. $r^2_{,}$ and (d) 2006, Mbp distance vs. r^2 .

Based on 312 markers in the 2005 dataset, 311,899 SNP marker-pair comparisons had $r^2<0.5$. Similarly, for the 2006 population based on 356 markers, 332,378 SNP marker-pair comparisons had $r^2<0.5$. These two marker sets were then used to decipher population structure and kinship.

Two different types of analyses were conducted to infer the number of subpopulations present in the 2005 and 2006 populations. The methods were implemented in the software programs STRUCTURE and Principal Component Analysis (PCA). The PCA analysis is an attractive approach compared to the STRUCTURE algorithm as it is computationally less demanding.

Population structure was estimated in the software STRUCTURE using the linkage based model approach for the multilocus genotype data (Pritchard et al. 2000). The number of subpopulations (*k*) was entered in the STRUCTURE input file and the software program assigned genotypes by estimating the number of loci in Hardy-Weinberg equilibrium and linkage equilibrium in each sub-population. In the linkage model approach, the software determined the posterior probabilities for each run of an assigned *k*-value. In the present study, a total of 10 runs were made for *k*-values ranging from 1 to15 for both the 2005 and the 2006 independent populations respectively. As described by Wang et al. (2008), the Wilcoxon two-sample test was used to compare the posterior probabilities averaged over all the 10 runs for a given *k* value, i.e., $k=1$ vs. $k = 2$; $k = 2$ vs. $k= 3$, and so on. The smaller *k* value in the first non-significant Wilcoxon two-sample test was considered to be the most accurate estimate of the population structure. Structure analysis showed that the 2005 population consisted of eight subpopulations, while the 2006 population was comprised of twelve subpopulations.

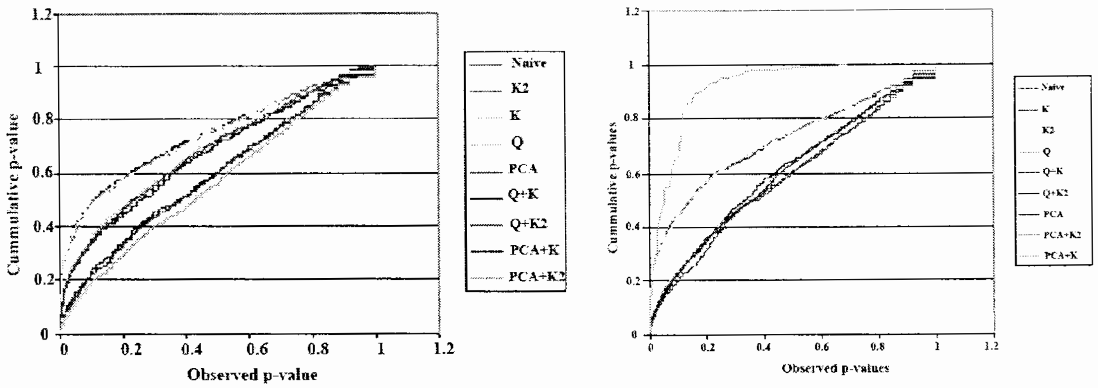
The PCA analysis, another approach to control population structure, was also implemented in this study. 50% of the molecular variance was explained by 13 principal components for the 2005 population and 12 principal components for the 2006 population. The first principal component explains the linear combination of the observed data with the greatest variance and the subsequent components maximize the variance subject to being uncorrelated with the preceding components. The first principal component accounted for approximately 10% of variance present in the 2005 and the 2006 populations.

To control for recent co-ancestry, a pairwise n x n kinship (K) coefficient matrix was developed for both the 2005 and the 2006 populations using the method employed by Loiselle et al. (1995). Another kinship matrix, K*, was generated using the shared haplotypes information in the Powemarker software for the two populations.

Selection of the mixed linear regression model for marker-trait associations

For all the nine different linear regression models (Table 1) independent marker-trait associations were conducted using 881 and 913 markers for the 2005 and the 2006 populations respectively. A P-P plot showing the distribution of raw *p*-values by cumulative *p*-values was developed for the nine linear regression models for the 2005 and the 2006 population (Figure 3a and 3b). For each population, the naive linear regression model, which did not consider population structure or co-ancestry, had the highest inflation of *p*-values. *P*-values were not uniformly distributed and 40.8% and 35.2% of the *p*-values were under the 5% threshold for both the populations (Table 3 and 4). The incidence of type I error was high in the naive model due to undetected population structure or kinship between individual genotypes. The different regression models which took into account population structure or kinship or both resulted in a decrease in the incidence of false positives marker-trait associations.

For the 2005 population, the distribution of *P*-values for PCA, PCA+K and PCA+K* models resembled a uniform distribution. Similarly, for the 2006 population, the distribution of *P*-values for Q, PCA, Q+K, Q+K* and PCA+K* models depicted a uniform distribution (Figure 3).



(a) (b)

Figure 3. Graphical representation of the distribution of the *p* values for the naive, general and mixed models: (a) using 881 SNP markers and average IDC scores for the 143 genotypes in the 2005 population, (b) 913 SNP markers and average IDC scores for the 141 genotypes in the 2006 population.

Table 3. Performance of nine different models to account for Type I error for the 2005 population and estimation of such an error for all the different regression models used for marker-trait associations

Models Tested	p-value ≤ 0.001	p-value ≤ 0.01	p-value ≤ 0.05
Naïve	11.50%	24.80%	40.80%
K	11.40%	24.10%	39.50%
K*	11.30%	24.30%	38.10%
Q	4.90%	13.30%	26.60%
PCA	1.80%	4.70%	14.90%
Q+K	5.30%	13.50%	25.20%
Q+K*	4.20%	12.70%	25.50%
PCA+K	1.40%	5.00%	13.40%
PCA+K*	1.50%	3.10%	11.60%

Table 4. Performance of nine different models to account for Type I error for the 2006 population and estimation of such an error for all the different regression models used for marker-trait associations

Models Tested	p-value ≤ 0.001	p-value ≤ 0.01	p-value ≤ 0.05
Naïve	9.30%	19.80%	35.20%
K	10.00%	19.93%	34.90%
K*	10.40%	20.40%	34.70%
Q	1.80%	7.10%	17.60%
PCA	0.98%	6.70%	15.00%
Q+K	2.12%	5.40%	13.60%
Q+K*	1.65%	6.90%	15.80%
PCA+K*	1.00%	5.80%	13.80%

Marker/trait associations

Associations between 881 and 913 SNP markers for the 2005 and the 2006 independent populations respectively and IDC visual observations were evaluated using the GLM and MLM analysis. In the 2005 population, out of the 881 SNP markers, 15% of markers in the PCA model, 13% in the PCA+K-model and 10% markers in the PCA+K* model were observed to be significantly associated with the IDC trait ($p < 0.05$) for the average IDC visual observations at all the locations. These markers effectively covered all soybean chromosomes. However, a significant reduction in the number of marker-trait pair associations was observed after correction of multiple testing using the pFDR procedures at a significance level of Q < 0.10. Only 4% of the markers in the PCA model, 3% in the PCA+K model and 2% of the markers in the PCA+K* model revealed marker-trait associations. In the 2005 data, averaged over all the locations, 17, 28 and 31 significant SNP marker-trait associations were detected in one or more than one of the PCA, PCA+K and PCA+K* models respectively. One SNP marker on each of the chromosomes of 3, 4 and 10, two markers on chromosome 7, three on 8 and four markers on chromosome 19 significantly associated with the IDC trait were observed to be significantly associated with the IDC trait in the three models studied (Table 5).

Association analysis was conducted with the 2006 population to check whether markers previously associated with IDC could be used to differentiate between IDC efficient and inefficient soybean genotypes in a second independent population. Out of the 913 markers, approximately 17% of the markers in the Q-model, 15% in the PCA model, 11% in the Q+K model, 15% in the Q+K* model and 13% markers in the PCA+K* model were observed to be significantly associated with the IDC trait ($p < 0.05$) for the average IDC visual observations at all the locations. pFDR, significantly reduced the number of marker-trait pair associations observed as a result of multiple comparisons. At a significance level of Q < 0.10, only 7% of the markers in the Q model, 4% in the PCA model, 0.04% in the Q+K model, 7% in the Q+K* model and 3% of the markers in the PCA+K* model revealed marker-trait associations. In the 2006 data, averaged over all the locations, 25, 34, 40, 63 and 64 significant SNP marker-trait associations were detected in one or more than one of the Q, PCA, Q+K, Q+K* and PCA+K* models respectively. One SNP marker on chromosomes 2, 3, 5, 10, 11, 16, 18 and 20, and three markers on chromosome 1 and 19 were significantly associated with the IDC trait (Q<0.10). These observations were shared among the five linear regression models (Table6)

Needle graph explicitly shows the multipoint marker trait association analysis result for the two population for all the 20 linkage groups. The vertical axis shows that $\log_{10}(P)$ obtained from the MLM analysis, the top horizontal axis represents the physical distance

Table 5. Significance of tests for association between soybean SNP markers and IDC ratings for the 2005 soybean population

				PCA		PCA+K *		PCA+K2		
Marker	Chromosome	Genetic distance(cM)	Physical distance(Mbp)	P-value	Q-value	P-value	Q-value	P-value	Q-value	R-sq
030669_06920	3	94.69	47.16	1.70E-04	1.30E-02	1.60E-04	1.60E-02	1.70E-04	1.10E-02	0.04
057913_15004	4	10.04	12.50	7.60E-05	8.70E-03	6.10E-05	9.20E-03	7.60E-05	7.70E-03	0.06
053261_11776	5	3.449	93.73	1.18E-03	3.62E-02	1.32E-03	3.74E-02	1.50E-03	6.60E-02	0.09
039383_07310	7	39.94	7.15	2.30E-03	6.40E-02	2.40E-03	5.70E-02	1.50E-03	6.60E-02	0.18
055937_13873	7	75.42	34.44	3.00E-03	7.60E-02	1.60E-03	5.10E-02	2.00E-03	8.30E-02	0.04
050171_09440	8	47.34	9.46	1.40E-03	5.60E-02	1.20E-03	5.10E-02	1.10E-04	8.10E-03	0.10
027726_06646	8	49.33	9.49	7.00E-04	3.80E-02	3.20E-04	2.10E-02	7.20E-05	7.70E-03	0.08
057257_14650	8	49.33	9.49	7.00E-04	3.80E-02	3.20E-04	2.10E-02	7.20E-05	7.70E-03	0.08
040695_07821	19	33.22	34.92	3.60E-03	8.30E-02	3.60E-03	7.90E-02	2.20E-03	8.60E-02	0.02
020457_04632	19	35.78	35.97	1.50E-04	1.30E-02	1.60E-04	1.60E-02	1.50E-04	1.00E-02	0.04
055315_13197	19	50.3	38.72	1.90E-04	1.30E-02	1.90E-04	1.60E-02	8.80E-05	7.70E-03	0.06
042563_08305	19	50.68	38.66	1.60E-05	7.60E-03	1.50E-05	9.10E-03	2.90E-06	1.80E-03	0.06

(Mbp) and the bottom axis represent the genetic distance (cM) between the markers. Each bar represents a marker locus and the corresponding value is an indication of the significance of that marker locus. The bars facing upward represent significant associations discovered in the 2005 population and those facing downward represent significant associations discovered in the 2006 population (Figure 4). All 881 and 913 SNP markers for the 2005 and 2006 populations respectively, with their pFDR values were used to plot the needle graph. The marker loci with large P-values are more in number as depicted by the thickness of the horizontal bars for all the LGs. On applying pFDR to correct for type I error, we found evidences of strong marker-trait associations on certain chromosomes for the 2005 and the 2006 populations separately. Some significant associations detected at $p<$ 0.05 and pFDR Q< 0.1 were unique to each population, while some shared genome-wide associations on chromosomes 2, 3, 7 11, 17 and 19 were also observed in the two populations. The dots in the needle graph represent those markers which have cleared the set significance level of $p<0.05$ and pFDR Q< 0.1.

Pairwise LD (2005 and 2006 combined population)

The pairwise LD was also calculated for the combined 284 genotypes of the combined 2005 and 2006 populations. A total of 847 SNP markers were common after combining 881 and 913 SNP markers from these populations. Based on 334 SNP marker loci, 357,713 marker-pair comparisons had $r^2< 0.5$. These marker loci were used to control for population structure and estimate the kinship. Kinship matrix (K*) was generated using the shared haplotype information in the Powermarker software for the combined population

Table 6. Significant association between IDC and SNP markers detected in all the 4 out of the 5 models for the 2006 population

				Association Models Tested					
				Q			**PCA**		
Marker	Chromosome	Genetic distance (cM)	Physical distance(Mbp)	P-value	Q-value	R-sq	P-value	Q-value	R-sq
058135_15106	1	41.58	44.090	7.10E-03	8.60E-02	0.03	2.70E-03	8.60E-02	0.03
054393_12560	1	42.01	43.460	1.60E-05	3.00E-03	0.10	2.30E-06	8.30E-04	0.04
055131_13049	1	42.23	38.840	3.80E-03	6.50E-02	0.01	3.00E-03	8.60E-02	0.06
064293_18611	1	42.57	42.370	6.10E-06	2.00E-03	0.08	2.00E-06	8.30E-04	0.06
059897_16201	1	42.63	41.640	2.00E-03	5.10E-02	0.05	4.90E-04	4.70E-02	0.01
065083_19095	1	45.58	6.090	3.00E-04	2.30E-02	0.02	2.70E-03	8.60E-02	0.02
016079_02059	2	48.5	9.790	2.80E-03	5.50E-02	0.03	9.70E-04	7.10E-02	0.03
013513_00508	3	81.04	43.990	5.70E-03	7.50E-02	0.02	2.30E-03	8.60E-02	0.04
018011_02495	5	50.98	34.290	2.20E-05	3.00E-03	0.04	1.30E-04	1.80E-02	0.03
047374_12913	10	117.38	47.970	5.60E-03	7.50E-02	0.05	3.70E-03	9.10E-02	0.02
054375_12539	10	120.652	49.020	3.36E-04	2.30E-02	0.08	9.95E-06	2.36E-03	0.03
900336_00920	11	113.62	38.770	5.00E-03	7.30E-02	0.03	1.10E-03	7.10E-02	0.02
042681_08346	13	49.32	27.310	3.20E-04	2.30E-02	0.04	2.80E-03	8.60E-02	0.03
017127_02213	14	11.51	2.220	3.40E-03	6.00E-02	0.06	1.70E-03	8.60E-02	0.02
039687_07541	15	18.77	32.990	9.00E-03	9.90E-02	0.03	4.20E-03	9.70E-02	0.01
030595_06910	16	23.8	3.039	8.50E-04	3.40E-02	0.08	3.40E-03	8.60E-02	0.04
011625_00310	16	85.58	36.540	6.60E-04	3.00E-02	0.03	3.30E-03	8.60E-02	0.06
013509_00507	18	78.5	57.350	2.90E-03	5.50E-02	0.04	2.50E-04	2.90E-02	0.02
047428_12928	19	29.32	16.640	2.80E-04	2.30E-02	0.02	2.30E-03	8.60E-02	0.02
059723_16418	19	56.4	40.360	5.00E-06	2.00E-03	0.12	2.00E-05	3.50E-03	0.09
035235_07156	19	74.76	44.570	6.50E-03	8.10E-02	0.04	2.10E-03	8.60E-02	0.06
013129_01447	19	90.44	48.070	3.00E-3	5.50E-02	0.06	3.00E-03	8.60E-02	0.03
010719_00713	20	109.17	45.780	5.10E-02	7.30E-02	0.09	4.50E-03	9.40E-02	0.03

Table 6(continued)

				Association Models Tested					
				Q+K			Q+K*		
		Genetic	Physical						
Marker	Chromosome	distance (cM)	distance(Mbp)	P-value	Q-value	R-sq	P-value	Q-value	R-sq
058135_15106	1	41.58	44.090	ns	ns	ns	1.70E-03	5.20E-02	0.03
054393_12560	1	42.01	43.460	ns	Ns	ns	3.10E-07	2..20E-04	0.10
055131_13049	1	42.23	38.840	1.90E-03	4.60E-02	0.01	3.50E-03	6.40E-02	0.01
064293_18611	1	42.57	42.370	2.70E-06	1.20E-03	0.08	1.40E-06	4.80E-04	0.08
059897_16201	1	42.63	41.640	5.20E-04	2.60E-02	0.05	6.00E-04	3.90E-02	0.05
065083_19095	1	45.58	6.090	ns	ns	ns	2.60E-04	2.40E-02	0.02
016079_02059	2	48.5	9.790	1.50E-03	4.10E-02	0.03	3.40E-03	6.40E-02	0.03
013513_00508	3	81.04	43.990	3.70E-03	7.10.E-02	0.02	4.30E-03	6.90E-02	0.02
018011_02495	5	50.98	34.290	2.00E-05	3.60E-03	0.04	2.40E-05	4.40E-03	0.04
047374_12913	10	117.38	47.970	5.00E-03	8.10E-02	0.05	4.00E-03	6.60E-02	0.05
054375_12539	10	120.652	49.020	3.37E-04	1.86E-02	0.08	3.40E-04	2.36E-02	0.08
900336_00920	11	113.62	38.770	4.50E-03	7.70E-02	0.03	2.37E-04	7.10E-02	0.03
042681_08346	13	49.32	27.310	ns	ns	ns	3.40E-04	2.70E-02	0.04
017127_02213	14	11.51	2.220	ns	ns	ns	2.80E-03	6.40E-02	0.06
039687_97541	15	18,77	32,990	6.50E-03	9.40E-02	0.03	8.30E-03	9.80E02	0.03
030595_06910	16	23.8	3.039	7.20E-04	2.80E-02	0.08	8.00E-04	4.50E-02	0.08
011625_00310	16	85.58	36.549	8.70E-04	2.80E-02	0.03	9.70E-04	4.50E-02	0.03
013509_00507	18	78.5	57.350	2.50E-03	5.20E-02	0.04	3.40E-03	6.40E-02	0.04
047428_12928	19	29.32	16.640	1.60E-04	1.50E-02	0.02	9.40E-05	1.30E-02	0.02
059723_16418	19	56.4	40.360	4.20E-06	1.20E-03	0.12	4.40E-06	1.00E-03	0.12
035235_07156	19	74.76	44.570	4.70E-03	7.90E-02	0.04	5.60E-03	7.60E-02	0.04
013129_01447	19	90.44	48.070	2.30E-03	5.20E-02	0.06	2.60E-03	5.90E-02	0.06
010719_00713	20	109.17	45.780	5.40E-03	8.60E-02	0.09	3.50E-03	6.40E-02	0.09

Table 6. (continued)

				Association Models Tested		
					PCA+K*	
Marker	Chromosome	Genetic distance (cM)	Physical distance(Mbp)	P-value	Q-value	R-sq
058135_15106	1	41.58	44.090	1.30E-03	8.00E-02	0.03
054393_12560	1	42.01	43.460	2.70E-07	2.00E-04	0.04
055131_13049	1	42.23	38.840	2.40E-03	9.90E-02	0.06
064293_18611	1	42.57	42.370	8.70E-07	3.10E-04	0.06
059897_16201	1	42.63	41.640	2.60E-04	3.10E-02	0.05
065083_19095	1	45.58	6.090	2.90E-03	9.90E-02	0.02
016079_02059	2	48.5	9.790	1.00E-03	6.80E-02	0.03
013513_00508	3	81.04	43.990	2.40E-03	9.90E-02	0.04
018011_02495	5	50.98	34.290	1.00E-04	1.90E-02	0.03
047374_12913	10	117.38	47.970	ns	ns	ns
054375_12539	10	120.652	49.020	7.53E-06	1.82E-03	0.03
900336_00920	11	113.62	38.770	9.40E-04	6.80E-02	0.02
042681_08346	13	49.32	27.310	2.70E-03	9.90E-02	0.03
017127_02213	14	11.51	2.220	1.60E-03	8.50E-02	0.02
039687_97541	15	18,77	32,990	ns	ns	ns
030595_06910	16	23.8	3.039	ns	ns	ns
011625_00310	16	85.58	36.549	3.30E-03	9.90E-02	0.06
013509_00507	18	78.5	57.350	2.60E-04	3.10E-02	0.02
047428_12928	19	29.32	16.640	4.60E-04	4.10E-02	0.02
059723_16418	19	56.4	40.360	ns	ns	ns
035235_07156	19	74.76	44.570	2.30E-03	9.90E-02	0.06
013129_01447	19	90.44	48.070	3.20E-03	9.90E-02	0.03
010719_00713	20	109.17	45.780	3.40E-03	9.90E-02	0.03

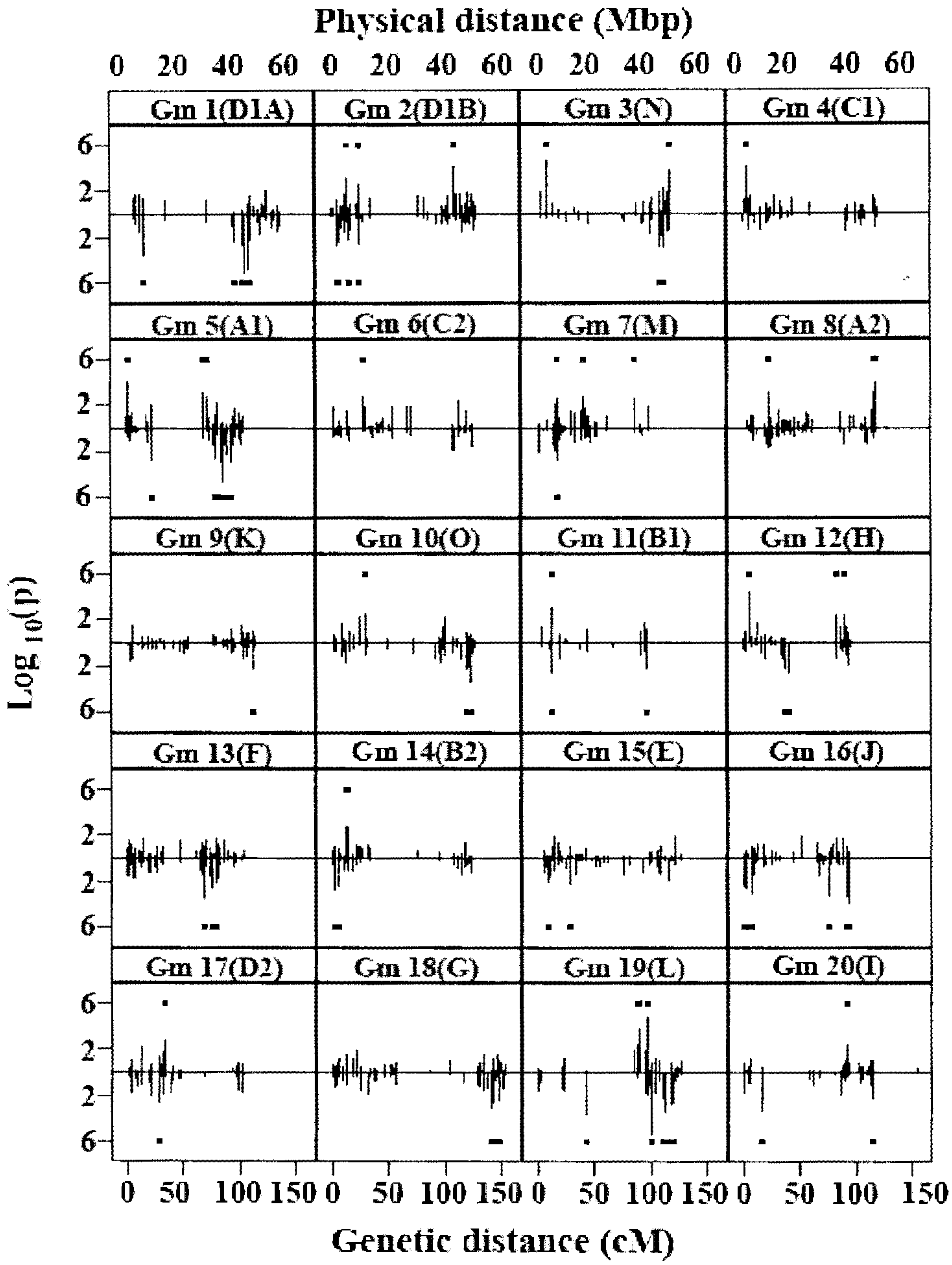


Figure 4. A needle graph depicting the multipoint marker-trait association analysis results for the two populations for all the 20 chromosomes (linkage groups). The bars facing upwards represent a marker loci significant loci in the 2005 population and those facing downwards represent significant marker loci in 2006 population with corresponding significance level of log$_{10}$(P) on the vertical axis . The dots represent the significant loci at Q<0.1.

SNP marker and IDC phenotypic associations (2005 and 2006 combined population)

For the combined population, PCA and PCA+K* analysis was implemented in SAS 9.3.1. Out of the 847 SNP markers, 17% of the markers in the PCA-model and PCA+ K* model were significantly associated with the IDC trait (p<0.05) for the average IDC visual observations at all the locations. After the correction of multiple testing using pFDR at Q<0.01, only 5%, i.e., 49 SNP markers out of the total 847 SNP markers) of the markers in the PCA model and 6%, i.e., 52 SNP markers out of the total 847 SNP markers) of the markers in the PCA+K* model revealed significant marker-trait associations. One SNP marker on 2, 6, 9, 11, 12, 15 and 16, two on 1 and 6, three on 7, 10, 13 and 17, four on 18 and 19, five on 5 and six SNP markers on 3, were significantly associated with the IDC trait in both the PCA and the PCA+K* model (Table 7). A needle graph was developed using the significant marker-trait associations discovered in just the PCA+K* model (Figure 5).

Ten SNP markers distributed on chromosomes 1, 3, 4, 5, 7, 17, 18, and 19 were highly significantly associated with the IDC trait with a pFDR <0.01 (Table 8). Chi-square test was conducted to check the difference in the mean phenotypic rating for the alleles with 5% error rate (p<0.05). Only the extreme 50 individuals are used for this. Finally, out of the 10 markers, six selected markers had a significant phenotypic mean difference for the tolerant and susceptible alleles. Further, considering each of the marker combinations as a treatment, we performed an analysis of variance and grouped them based on LS means. We found that the phenotypic mean of genotypes with all 6 tolerant

Table 7. Significance of tests for association between soybean SNP markers and IDC ratings for the combined soybean population using the PCA and PCA+K* statistical approaches

		Genetic	Physical	PCA		PCA+K *		
Marker	Chromosome	distance (cM)	distance(Mbp)	P-value	Q-value	P-value	Q-value	R-sq
058135_15106	1	41.58	44.09	1.90E-04	1.32E-02	1.22E-04	8.84E-03	0.026
054393_12560	1	42.007	43.46	1.95E-05	2.70E-03	1.62E-05	2.70E-03	0.032
056237_14178	2	30.668	5.46	4.64E-03	7.36E-02	3.57E-03	7.70E-02	0.005
028539_05944	3	74.162	43.49	2.12E-03	4.90E-02	2.07E-03	5.60E-02	0.13
048557_10665	3	79.051	43.81	1.40E-03	3.87E-02	1.12E-03	3.62E-02	0.029
044603_08734	3	85.822	45.01	4.12E-05	4.57E-03	2.08E-05	2.70E-03	0.137
060109_16388	3	86.907	45.39	2.82E-04	1.42E-02	2.34E-04	1.36E-02	0.188
016535_02085	3	88.225	45.42	9.21E-04	3.07E-02	5.07E-04	1.94E-02	0.171
030669_06920	3	94.686	47.16	1.58E-03	3.98E-02	4.00E-03	7.70E-02	0.016
044521_08714	4	33.106	6.38	9.40E-04	3.07E-02	1.12E-04	8.84E-03	0.013
907035_01038	4	53.274	37.36	1.75E-03	4.21E-02	2.26E-03	5.87E-02	0.009
053261_11776	5	3.449	0.94	1.18E-03	3.62E-02	1.32E-03	3.74E-02	0.006
052043_11321	5	21.708	8.91	2.66E-03	5.32E-02	1.26E-03	3.74E-02	0.076
058785_15434	5	27.639	6.85	8.09E-03	9.81E-02	4.91E-03	8.49E-02	0.009
054163_12369	5	29.91	27.21	8.05E-05	7.43E-03	1.36E-04	8.86E-03	0.033
042331_08243	5	42.038	32.36	8.49E-03	9.81E-02	5.36E-03	8.49E-02	0.004
014557_01578	6	67.988	11.82	8.45E-03	9.81E-02	3.95E-03	7.70E-02	0.015
031395_07087	7	38.468	6.54	2.43E-04	1.42E-02	9.55E-05	8.84E-03	0.059
900461_00929	7	39.074	7.30	4.13E-03	7.15E-02	2.62E-03	6.09E-02	0.127
039383_07310	7	39.939	7.15	5.59E-04	2.21E-02	2.97E-04	1.41E-02	0.155
042049_08162	9	12.142	2.01	6.87E-04	2.54E-02	1.20E-03	3.72E-02	0.032
055653_13572	10	34.209	4.81	4.40E-03	7.22E-02	5.17E-03	8.49E-02	0.026
019105_03305	10	50.102	11.77	3.08E-04	1.42E-02	5.00E-04	1.94E-02	0.033
029491_06207	10	82.103	40.22	2.69E-03	5.32E-02	4.03E-03	7.70E-02	0.024

Table 7(continued)

Marker	Chromosome	Genetic distance (cM)	Physical distance(Mbp)	PCA		PCA+K *		
				P-value	Q-value	P-value	Q-value	R-sq
041167_07925	11	76.21	17.42	2.67E-02	5.32E-02	1.09E-03	3.62E02	0.01
007732_00002	12	61.15	13.60	4.43E-02	7.22E-02	4.32E-03	8.01E-02	0.008
058031_15072	13	31.361	7.70	3.43E-04	1.46E-02	4.26E-04	1.85E-02	0.013
043173_08548	13	33.672	8.26	3.08E-03	5.89E-02	3.75E-03	7.70E-02	0.02
055499_13329	13	61.354	31.47	6.34E-03	9.01E-02	7.84E-03	9.80E-02	0.114
038977_07417	15	33.181	7.03	7.18E-03	9.56E-02	6.54E-03	9.65E-02	0.007
059379_15786	16	47.279	26.40	5.63E-03	8.44E-02	7.47E-03	9.70E-02	0.011
031827_07220	17	8.164	1.79	3.45E-03	6.16E-02	2.36E-03	5.89E-02	0.011
050543_09730	17	64.319	12.97	2.77E-04	1.42E-02	4.15E-05	4.50E-03	0.057
017059_02191	17	64.734	13.06	1.25E-03	3.65E-02	2.50E-04	1.36E-02	0.027
047504_12947	18	55.603	22.54	8.36E-03	9.81E-02	7.18E-03	9.70E-02	0.016
013509_00507	18	78.501	57.35	1.59E-05	2.70E-03	1.66E-05	2.70E-03	0.067
059017_15576	18	78.894	57.38	5.17E-03	7.96E-02	7.72E-03	9.80E-02	0.055
054089_12331	18	79.119	57.29	2.67E-03	5.32E-02	5.64E-03	8.73E-02	0.011
042665_08342	19	46.098	38.00	3.41E-03	6.16E-02	5.12E-03	8.49E-02	0.065
055315_13197	19	50.297	38.72	1.59E-06	8.80E-04	3.70E-06	2.40E-03	0.061
042563_08305	19	50.678	38.66	5.73E-06	1.59E-03	1.18E-05	2.70E-03	0.066
059723_16418	19	56.404	40.360	1.74E-04	1.32E-02	3.05E-04	1.41E-02	0.133

All the 847 SNP markers in the combined population were used to plot the needle graph taking into consideration the significance level of P<0.05 and pFDR Q<0.1.

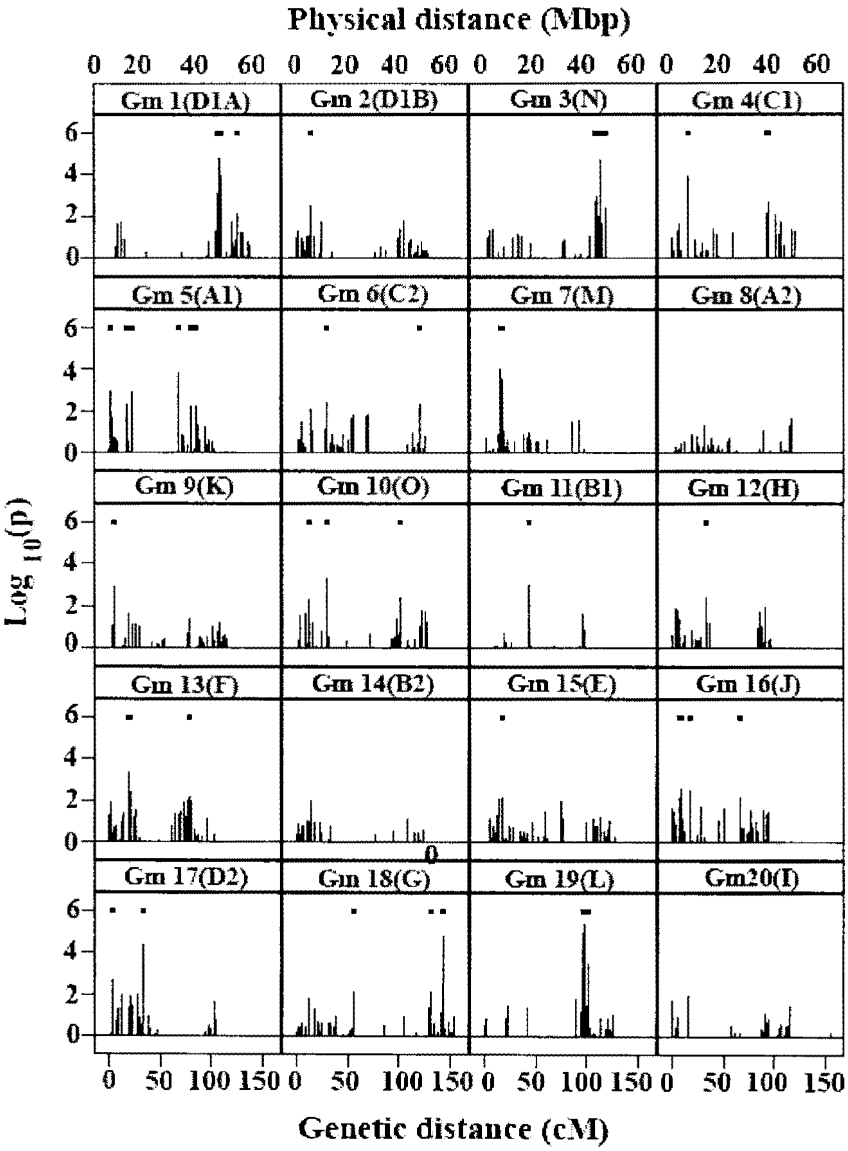


Figure 5. A needle graph depicting the multipoint marker-trait association analysis results for the combined population for all the 20 chromosomes (linkage groups). The bars represent significant marker loci present in the both the population with corresponding significance level of $\log_{10}(P)$ on the vertical axis . The dots represent the significant loci at Q <0.01.

Table 8. Ten significant marker loci with phenotypic mean difference for tolerant and susceptible alleles

Marker	Chromosome	Genetic distance(cM)	Physical distance(Mbp)	Phenotypic Mean Chi- sq
058135_15106	1	41.58	44.09	ns
054393_12560	1	42.01	43.46	***
044603_08734	3	85.82	45.01	***
044521_08714	4	33.11	6.38	ns
054163_12369	5	29.91	27.21	ns
031395_07087	7	38.47	6.54	***
050543_09730	17	64.32	12.97	***
013509_00507	18	78.50	57.35	ns
055315_13197	19	50.30	38.72	***
042563_08305	19	50.68	38.66	***

Only markers with pFDR <0.01 were accepted

alleles is significantly not different from the genotypes with only two tolerant alleles at 44603_08734 and 050543_09730, and 050543_09730 and 042563_08305 (Table 9).

Table 9. IDC mean scores for six SNP markers genotypic classes

SNP marker						
044603_08734	031396_07087	050543_09730	054393_12560	042563_08305	055315_13197	IDC rating
Tx	Gx	Ax	Gx	Ax	Cx	1.80
Ay	Ay	A	Ay	A	C	2.18
T	A	A	A	A	C	2.28
T	G	A	A	A	C	2.30
A	G	A	A	Ty	C	2.35
T	A	A	A	T	Ay	2.38
A	A	A	A	A	A	2.40
T	G	A	G	T	A	2.47
T	A	A	A	A	A	2.48
T	A	A	G	A	C	2.53
A	A	A	G	A	C	2.60

However, three markers effectively identified IDC tolerant and susceptible genotypes (Table 10). In the present study we also found candidate genes located within the previously identified QTL regions (Table 11).

Table 10. Selection of soybean genotypes based on three marker loci selection

044603_08734	042563_08305	050543_09730	Mean
Tx	Ax	Ax	2.34
Tx	Tx	Ax	2.59
Ay	Ax	Ax	2.75
Ay	Tx	Ax	2.77
Tx	Ax	Gy	2.85
Tx	Tx	Gy	2.98
Ay	Ax	Gy	3.13
Ay	Tx	Gy	3.13

Table 11. Six marker loci in the combined population mapped near known Fe metabolism gene

Marker	Chromosome	Marker Physical distance (Mbp) gene	Known Iron metabolism	Gene Physical distance (Mbp)
054393_12560	1	43.5	ATIREG1	43.2
044603_08734	3	45	NAS	45.3
031395_07087	7	6.5	FRO	6.5
050543_09730	17	13	FRO	14.2
042563_08305	19	38.7	NAS-like	38.2
055315_13197	19	38.7	NAS-like	38.2

DISCUSSION

Association mapping has been extensively used in human genetics; however, lately it has also gained importance in the plant genetics. In this study, association mapping was used to identify QTL associated with IDC in soybean using SNP markers. The study material for this research was comprised of two independent soybean populations. A set of 1536 widely-distributed SNP markers from the Soy Linkage Panel 1.0 was selected. The idea behind using two independent populations was to recheck whether significant marker-trait associations discovered in the first population could be reproduced in the second population. The data from both years was also the combined data for the 2005 and the 2006 populations. These lines, when grown in different environments, helped to neutralize the effects of environmental variations and thereby increased the heritability of each individual QTL.

Marker assisted selection, a tool employed by breeding programs, requires markers which are a good representative of the genetic variations present in a wide variety of soybean germplasm and not just in segregating populations (Malosetti et al. 2007). SNP markers were chosen for QTL mapping and association analysis because they are the most abundant markers and have a low mutation rate. A significant advantage of using these markers is that the availability of high throughput and highly automated technologies facilitate the genotyping of a few to millions of SNPs in a few to millions of genotypes over a short period of time. The Golden Gate assay (Illumina Inc., San Diego, CA) can perform genotyping of up to 1,536 SNPs in 192 DNA samples within three days with the desired accuracy despite the recent diplodized tetraploid event in soybean (Hyten et al.

2008). These markers have been frequently used for construction of genetic maps to fine map and clone agronomically important traits in several crop species (Rostoks et al. 2005)

Some previous studies have shown that soybean has a relatively low SNP frequency compared to other cultivated crop species (Zhu et al. 2003; Hyten et al. 2006). However, Choi et al. (2007) discovered 5500 SNPs in 2032 gene transcripts and mapped at least one SNP from 1141 gene transcripts and created version 3 of the soybean integrated linkage map. In the present study, we observed widely distributed SNPs throughout the genome with the genetic spacing between any two markers to be less than 10 cM in both the populations. The size of the 2005 and the 2006 populations, i.e.,143 and 141 lines respectively, seems appropriate to estimate multi-locus LD using co-dominant SNP markers. According to a study by Li et al. (2007), a sample size of 30 would be adequate to estimate LD in a population using co-dominant markers regardless of the level of heterozygosity. Among the 1536 SNP markers chosen for the GoldenGate assay, 1265 were polymorphic in the two populations. It is possible that the monomorphic SNP markers might have been identical by descent between the genotypes studied. According to a simulated data study by Tabangin et al. (2009), SNPs with minor allelic frequency (MAF) of 1% or 5% tend to increase false positives. In general, genome-wide association studies remove SNPs with MAF<10% (Tabangin et al. 2009). In the present study, from the initial set of 1265 polymorphic SNP markers, we discarded SNP marker with MAF<10% to avoid false positives. However, we also acknowledge that removal of SNPs with low MAF may hamper the ability to detect rare variants with a significant effect. But given the complex nature of IDC tolerance, a single major factor is not expected. The average PIC value for markers with MAF>10% was 0.3045 and 0.2969 for the 2005 and the 2006 populations respectively. The PIC value is dependent on two factors: first, the variability at the marker locus and second, the transferability of alleles among germplasms. Since in our study, the SNPs were derived from EST which tend to be more conserved compared to random regions of the genome, hence, lower PIC values were expected.

In our study, some genotypes had missing SNP data and these were imputed in fastPHASE. In genome wide association studies, it is a little uncertain as to what proportions of missing SNP are captured by the genotyped SNP. The missing data is imputed in fastPHASE with the assumption that the imputed SNPs might cover the variations which would otherwise remain undetected. Jannink et al. (2009) performed marker imputation studies in barley and concluded that fastPHASE accurately imputed nearly 80% of the markers correctly more than 95% of the time.

In association mapping, one has to understand the structure of LD in a population. LD helps to determine the marker density required to effectively assay the common variants. Hyten et al. (2007) reported that a wild outcrossing ancestor of soybean, *G. soja,* showed lower levels of LD than self-fertilizing Asian *G. max* landraces. North America (N.Am.) cultivars developed from Asian *G. max* landraces and the elite cultivars developed from N. Am. Cultivars showed high levels of LD persisting from 90 to 574 Kb. In general, LD decreases rapidly in outcrossing plant species as compared to selfing plant species. The different soybean breeding programs analyzed here exploit the narrow-based germplasm belonging to the maturity groups 00, 0 and 1. This results in a narrow genetic diversity, i.e., a limited number of allelic combinations on chromosomes. The populations analyzed in the current study belong to this narrow-based germplasm and as expected, LD persisted to a longer physical and genetic distance. LD declined to an $r^2<0.05$ within 7

Mbp and 20 cM in the 2005 population, and within 5 Mbp and 12 cM in the 2006 population.

In association mapping studies, spurious association is a common problem arising due to poorly understood population structure and ancestral relationships. In our study, ancestry informative markers (AIM) were selected to infer population structure and ancestral relationship and all markers with high intrachromosomal LD with an $r^2>0.5$ were removed. Only 312 and 356 AIMs markers with r2<0.5 in the 2005 and the 2006 populations respectively, with high allele frequency differences among the ancestral populations were retained to estimate population structure and ancestral kinships. According to Barnholtz-Sloan et al. (2008), approximately 50 to 100 AIMs can determine an individual's ancestry.

The mixed linear model (MLM) approach which considers the population structure (Q) as a fixed effect and relatedness (K) as the variance-covariance structure of the random effect is generally used in genome wide association studies (Wang et al. 2008). In the present study, the naive model which does not take into account population structure and kinship was discarded because of high incidence of spurious associations (40.80% and 35.30% of the values were under the 5% significance level for the 2005 and 2006 populations, respectively). Yu et al. (2006), introduced the mixed-model approach for studying association mapping in allogamous species such as human and maize and concluded that the Q+K model resulted in a better approximation of the expected *p*-value with respect to the cumulative distribution of *p*-values, followed by K model, the Q model and lastly, the naive model. Price et al. (2006) and Balding (2006) observed that the principal components analysis was computationally less intensive then estimating structure using the Bayesian method. The software program STRUCTURE used for association mapping is developed for unrelated genotypes that belong to populations in Hardy-Weinberg equilibrium (Pritchard et al. 2000). However, for germplasm sets of most species these assumptions might not be met, and thus STRUCTURE demands careful interpretation (Camus-Kulandaivelu et al. 2007). Zhao et al. (2007) demonstrated the ability of an alternative kinship matrix (K*), estimated as the (is it correct) fraction of shared fragment haplotypes, to capture the underlying structure in *A. thaliana* and compared it with other models such as Q, PCA, Q+K, Q+K*, PCA+K and PCA+K* using a P-P plot. The K* model performed the best, while the PCA model performed better than the Q model. However, a mixed-model approach combining PCA and Q with kinship estimates performed similarly. The study suggested that the alternative kinship was successful in reducing the type I error in the same way as the MLM approach used by Yu et al (2006). Casa et al. (2008), with the help of a P-P plot, concluded that Q+K and Q+K* were better mixed model approaches with 4.8% of the *p*-values under the 5% threshold as compared to other models like K or K* with 5.1% of the *p*-values under the 5% threshold. Since the most favorable model approach for conducting a genome-wide association is under question, we decided to test different general linear models (GLM) and mixed linear model (MLM) in this study. We observed that for the 2005 population, the PCA, PCA+K and PCA+K* models and for the 2006 population, the Q, PCA, Q+K, Q+K* and PCA+K* models showed uniform distribution of the observed *P*-value with respect to the cumulative distribution of *P*-values. As explained by Stich et al. (2008), the K-model approach alone was not an appropriate approach to discover marker-trait association and the P+K model approach is a promising alternative to Q+K model approach.

In this study we have used genome wide AM to evaluate the genetic basis of IDC tolerance in soybean. Previous QTL studies using bi-parental populations have shown two inheritance patterns for IDC efficiency: first, a major gene with several modifier genes and second, the polygenic inheritance. Lin et al. (2000) studied the inheritance of iron in two intraspecific populations. Using the 2-year combined visual score data in the Pride x A15 population, polygenic inheritance was postulated and QTL for visual score exhibiting phenotypic variations ranging from 7.7 to 10.8% were mapped on chromosomes 3, 14, and 18. These together explained a total of 21.5% of the phenotypic variation. QTL for chlorophyll concentration controlling 34.8% of the total phenotypic variation were mapped on chromosomes 12, 14, and 20. Similarly, in Anoka x A7 population in which major gene and modifying gene inheritance was studied, chromosome 3 accounted for 72.7% and chromosome 5 accounted for 35.2% of the visual score variations. Together, these loci controlled 73.5% of the total variations for the visual scores. QTL for chlorophyll concentration mapped on chromosome 3 and 20 together explained 80.7% of the phenotypic variations. However, the markers flanking the QTL were population specific. This limits their efficiency for marker-assisted selection. In our study of the 2005 population, we used the PCA GLM, PCA+K MLM and PCA+K* MLM approaches, and observed that that 31, 28 and 17 markers, in this order, were associated with IDC. These markers were widely distributed on chromosomes 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 14, 17 and 19.

Similarly, for the 2006 population, 64 markers were associated with IDC using the Q GLM approach, 35 with the PCA GLM approach, 25 with the PCA+K* MLM approach, 40 with the Q+K MLM approach and 63 markers were associated with IDC using the

Q+K* MLM approach. The markers were distributed on chromosomes 1, 2, 3, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20. We have identified novel IDC related QTL which were previously not detected on certain chromosomes. The identification of all these new IDC related QTL on several different chromosomes gives an indication of the plethora of candidate genes involved in IDC response.

Since none of the QTL identified were common to both the populations, their application in the marker-assisted selection in breeding for IDC trait would have been questionable. Hence, we decided to combine the genotypic and the phenotypic data of the 2005 and the 2006 populations to search for common markers. Using the PCA GLM approach, we observed that 49 markers were significantly associated with IDC; the corresponding figure for the PCA+K* MLM approach was 52. These markers were distributed on chromosomes 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 15, 16, 17, 18 and 19.

In any QTL analysis, the usefulness of a marker is determined by its ability to distinguish between the resistant and susceptible genotypes in an environment independent fashion. Another characteristic of a marker is its reproducibility, i.e., the marker associated with a trait in one population should be associated with the same trait in a second population. On the basis of a 1 year data, Charlson et al. (2003) observed that SSR marker Satt481, mapped on chromosome 19, was associated with IDC resistant genotypes in 27% instances. Charlson et al. (2005) further confirmed the ability of the Satt481 (accounting for 12% of the total phenotypic variation) to effectively screen for IDC efficient genotypes. Other significant IDC-related QTL like Satt211 mapped on chromosome 5, and Satt104 mapped on chromosome 20 did not improve the population mean scores when selection was done based on homozygous lines for the tolerant allele. In the present study, the probable reasons for not detecting the previously identified significant associations could have been the fixation of alleles contributing to the phenotypic variations in the population studied and the population specific nature of the analyzed QTL. To overcome these shortcomings, Wang et al. (2008) utilized the AM approach, an alternative for IDC related QTL discovery, using two advanced breeding lines provided by different private and public breeding programs. In their study, the populations analyzed represented the breeding material for the north central US, and the markers identified might have broad applicability for the same region. They observed that Satt144 mapped on chromosome 13 and Satt239 mapped on chromosome 20 were significantly associated with the IDC trait in the 2002 population, and the results were reproducible in the 2003 population. These markers showed significant association with the IDC trait in all the different genetic analyses like single factor analysis, Q GLM, K MLM, Q+K MLM.

We tested different models to discover significant marker trait associations and detected different marker/trait pairs. We decided to work with markers which were present in all the three selected models for the 2005 population and in four out of the five selected models for the 2006 population. For the combined population, we selected only the common markers between the two models (PCA and PCA+K*) studied. Forty-three significant marker-trait associations were found in common between the studied PCA and PCA+K* approaches. The individual QTL detected in one population might have poor penetrance or expression for the iron deficiency chlorosis and hence might not have been detected in another population. Also, since IDC is a quantitative trait, the phenotypic observations are sensitive to the field environment. As the number of field trials and the environments increased, the probability of detecting IDC related QTL in the combined population also increased.

O'Rourke et al. (2009) took advantage of the microarray technology to study the differential expression of IDC related genes in two near isogenic lines of soybean and utilized the whole genome sequence assembly information of soybean to genetically position the identified genes. Clusters of differentially expressed genes were seen throughout the genome and a few of them were located within the previously iron QTL regions. Gene clusters were identified on chromosomes 2, 5, 6, 7, 9, 12 and 13. An extension of our association mapping study would be to search for candidate genes located within the identified QTL regions in the present study and then look for candidate polymorphisms.

REFERENCES

1. Abdurakhmonov, I. Y., and A. Abdukarimov. 2008. Application of association mapping to understanding the genetic diversity of plant germplasm resources. *J. Plant Genomics* 2008 (id 574927): 1-18.
2. Agrama, H. A., G. C. Eizenga, and W. Yan. 2007. Association mapping of yield and its components in rice cultivars. *Mol Breed.* 19:341-356.
3. Aranzana M. J., S. Kim, K. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. Tang, C. Toomajian,B Traw, H. Jheng, J Bergelson, C Dean, P. Marjoram, and M. Nordborg. 2005. Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* 1: e60.
4. Arumuganathan, K., and E. D. Earle. 1991. Estimation of nuclear DNA content of plants by flow cytometry. *Plant Molecular Biology Reporter.* 9: 229-241.
5. Ash, M., J. Livezey, and E. Dohlman. 2006. Soybean Backgrounder, Electronic Outlook Report from the Economic Research Service, USDA, OCS-2006-01.
6. Audebert, A. 2006. Iron partitioning as a mechanism for iron toxicity tolerance in lowland rice. In *Iron toxicity in rice-based systems in West Africa.* eds. Audbert, A., Narteh, L.T., Kiepe, P., Millar, D., Beks B. Cotonou: WARDA [Africa Rice center]. 34-46.
7. Balding, D. J. 2006. A tutorial on statistical methods for population association studies. *Nat. Rev. Gen.* 7: 781-791.

8. Barker, A. V., and D. J. Pilbeam, eds. 2007. Handbook of plant nutrition. Vol 117 Edition 1:335-337 New York, Philadelphia, Oxford, Melbourne, Stockholm, Beijing, New Delhi, Johannesburg, Singapore and Tokyo: Taylor & Francis

9. Barnholtz-Sloan, J. S., B. McEvoy, M. D. Shriver, and T. R. Rebbeck. 2008. Ancestry estimation and correction for population stratification in molecular epidemiologic association studies. *Cancer Epidemiol Biomarkers.* 3: 471-477.

10. Bashir, K., H. Inoue, S. Nagasaka, M. Takahashi, H. Nakanishi,and S. Mori. 2006. Cloning and characterization of deoxymugineic acid synthase genes from graminaceous plants. *Journal of Biological Chemistry* 281:32395–32402.

11. Bauer, P., H. Q. Ling, and M. L. Guerinot. 2007. FIT, the FER-LIKE IRON DEFICIENCY INDUCEDTRANSCRIPTION FACTOR in Arabidopsis. *Plant Physiol. Biochem.* 45:260–261.

12. Becana, M., J. F. Moran, and I. Iturbe-Ormaetxe. 1998. Iron dependent oxygen free radical generation in plant subjected to environmental stress: toxicity and antioxidant protection. *Plant and Soil* 201: 137-147.

13. Becker, M., and F. Asch. 2005. Iron Toxicity – Conditions and management concepts. *J. PlantNutr. Soil Sci.* 168: 558-573.

14. Beckmann, J. S., and M. Soller. 1986. Restriction fragment length polymorphisms and genetic improvement of agricultural species. *Euphytica* 35: 111-124.

15. Bernardo. R. 2008. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Sci.* 48: 1649-1664.

16. Briat, J.F., C. Curie, and F. Gaymard. 2006. Ferritins and iron accumulation in plant tissues. In: Iron Nutrition in Planta and rhizospheric microorganisms. eds. Barton L, Abadia, J. Berlin: Sprinder. 345-361.

17. Briat, J.F., and S. Lobreaux. 1998. Iron storage and ferritin in plants. *Metal Ions Biol. Sys.* 35:563–583.

18. Brown, J. C. 1956. Iron Chlorosis. *Ann review of plant physiology*. 7:171-190.

19. Brown, J. C., R. S. Holmes, and L. O. Tiffin. 1958. Iron chlorosis in soybeans as related to the genotype of the rootstock. Soil Sci. 86: 75-82.

20. Brown, J. C., R. S. Holmes, and L. O. Tiffin. 1959a. Hypothesis concerning iron chlorosis. *Soil Sci. Soc. Am. Proc.* 23: 231-234.

21. Brown, J. C., L. O. Tiffin, R. S. Holmes, A. W. Specht, and J. W. Resnicky. 1959b. Internal inactivation of iron in soybean as affected by root medium. *Soil Sci.* 87: 89-94.

22. Brown, J. C.1978. Mechanism of iron uptake by plants. *Plant Cell Environ.* 1: 249-257.

23. Brown, J. C., R. L. Chaney, and J. E. Ambler. 1971. A new mutant inefficient in the transport of iron. *Physiologica Plantarum* 25: 48-53.

24. Bughio, N., H. Yamaguchi, N. K. Nishizawa, H. Nakanishi, and S. Mori. 2002. Clonning an iron-regualted metal transporter from rice. *Journal of Experimental Botany* 53: 1677-1682.

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

26. Camus-Kulandaivelu, L., J. B. Veyrieras, B. Gouesnard, A. Charcosset, and D. Manicacci. 2007. Evaluating the reliabilityof structure outputs in case of relatedness between individuals. *Crop Sci.* 47: 887–892.

27. Casa, A. M., G. Pressoir, P. J. Brown, S. E. Mitchell, W. L. Rooney, M. R. Tuinstra, C.D. Franks, and S. Kresovich. 2008 Community resources and strategies for association mapping in sorghum. *Crop Sci.* 48: 30–40.

28. Charlson, D. V., S. R. Cianzio, and R. C. Shoemaker. 2003. Associating SSR markers with soybean resistance to iron deficiency chlorosis. *Journal of Plant Nutr.* 26: 2267-2276.

29. Charlson, D. V., T. B. Bailey, S. R. Cianzio, and R.C. Shoemaker. 2005. Molecular marker Satt481 is associated with iron deficiency chlorosis resistance in a soybean breeding population. *Crop Sci.* 45: 2394-2399.

30. Chen, Y., and P. Barak. 1982. Iron nutrition in calcareous soil. *Adv Agron.* 35:217-240.

31. Choi, I. Y., D. L. Hyten, L. K. Matukumalli, Q.J. 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. 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.

32. Cianzio, S. R., W.R. Fehr, and I. C. Anderson. 1979. Genotypic evaluation for iron deficiency chlorosis in soybean for visual scored and chlorophyll concentration. *J. Plant Nutr.* 19: 644-646.

33. Cianzaio, S. R., and Fehr, W. R. 1980. Genetic control of iron deficiency chlorosis in soybeans. *Iowa State Journal of Research* 54: 367-375.

34. Cianzio, S. R., and Fehr, W. R. 1982. Variation in the inheritance of resistance to iron deficiency chlorosis in soybeans. *Crop Sci.* 22: 433-434.

35. Cianzio, S. R. 1991. Recent advances in breeding for improving iron utilization by plants. Plant and Soil. 130: 63-68.

36. Clemens, S., M. G. Palmgren, and U. Krämer. 2002. A long way ahead: understanding and engineering plant metal accumulation. *Trends Plant Sci* 7: 309–315.

37. Cockram, J., J. White, F.J. Leigh, V. J. Lea, E. Chiapparino, D.A. Laurie, I. J. Mackay, W. Powell, and D.M. O'Sullivan. 2008. Association mapping of partitioning loci in barley. *BMC Genetics* 9: 16.

38. Collard, B. C. Y., M. Z. Z. Jahuffer, J. B. Brouwer, and E. C. K. 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.

39. Connolly, E.L., and M. L. Guerinot. 1998. In: Plasma membrane redox system and their role in biological stress and disease. eds. H. Asard, A.Berczi, and R.J. Caubergs. Kluwer Academic Drodrecht, 179-192.

40. Coulombe, B., R. Chaney, and W. Wiebold. 1984. Bicarbonate directly induces iron chlorosis in susceptible soybean cultivars. *Soil Sci. Soc. Am. J.* 48:1297-1301.

41. 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.

42. Curie, C., J. M. Alonso, M. Le Jean, J. R. Ecker, and J. F. Briat. 2000. Involvement of NRAMP1 from Arabidopsis thaliana in iron tranport. *Biochem J.* 347:749-755.

43. da Silveira, V. C., A. P de Oliveira, R. A. Sperotto, L. S. Espindola, L. Amaral, J. F.Dias, J. B. da Cunha. And J. P. Fett. 2007. Influence of iron on mineral status of two rice (*Oryza sativa* L.) cultivars. *Brazilian Journal of Plant Physiology* 19: 127-139.

44. Dancis, A., D. S. Yuan, D. Haile, C. Askwith, D. Eide, C. Moehle, J. Kaplan, and R.D. Klausner. 1994. Molecular characterization of a copper transport protein in S. cerevisiae: an unexpected role for copper in iron transport. *Cell* 76: 393-402.

45. DeBoer, A.H., H.B.A, Prius, and P.E. Zanstra.1983. Biphasic composition of transroot electrical potential in roots of Plantago species: Involvement of spacially separated electrogenic pumps. *Planta* 157: 259-266.

46. Diers, B. W., S. R. Cianzio, and R.C. Shoemaker. 1992. Possible identification of quantitative trait loci affecting iron efficiency in soybean. *J.Plant Nutr.* 15: 2127-2136.

47. Diers, B.W., B. K. Voss, and W. R. Fehr.1991. Moving-mean analysis of field tests for iron efficiency of soybean. *Crop Sci.*31:54-56.

48. Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure from small quantities of fresh leaf tissues. *Phytochem Bull.* 19:11-15

49. Doyle, J. J. and J. L. Doyle. 1990. Isolation of plant DNA from fresh tissue. *Focus* 12: 13-15.

50. Duy, D., G. Wanner, A. R. Medua, N.V. Wiren, J. Sol, and Philippar. 2007. PIC1, an ancient Permease in Arabidopsis Chloroplasts Mediates Iron Transport. The Plant Cell 19: 986-1006.

51. Echardt, U., A. M. Marques, and T. J. Buckhout. 2001. Two iron-regulated cation transporters from tomato complement metal uptake-deficient yeast mutants. *Plant Mol. Biol.* 45: 437-448.

52. Eide, D., M. Broderius, J. Fett, and M. L. Guerinot. 1996. A novel iron regulated metal transporter from plants identified by functional expression in yeast. *Proc Natl Acad Sci.* U S A. 93: 5624-5628.

53. Falconer, D. S. and T. F. C. Mackay. 1996. Introduction to quantitative genetics. Fourth Edition. Longman, New York. Falk, CT (1992).

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

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

56. Franzen, D.W., and J. L. Richardson. 2000. Soil factors affecting iron chlorosis of soybean in the Red River Valley of North Dakota and Minnesota. *J.Plant Nutr.* 23: 67-78.

57. Froehlich, D. M., and W. R. Fehr. 1981. Agronomic performance of soybeans with differing levels of iron deficiency chlorosis on calcereous soils. *Crop Sci.* 21: 438-441.

58. Frudakis, T.N. 2008. Molecular photofitting: predicting ancestry and phenotype using DNA. 57-143. Elsevier: London

59. Gardner, J. C., and T. L. Payne. 2003. A soybean biotechnology outlook. *AgBioForum* 6(1&2). *www.agbioforum.org/v6n12/v6n12a01-gardner.htm*

60. Garris, A. J., S. R. Mccouch, and S. Kresovich. 2003. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (Oryza sativa L.). *Genetics* 165: 759–769.

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

62. Gerendas, J., and U. Schurr. 1999. Physiochemical aspects of ion relations and pH regulation in plants – a quantitative approach. *J. Exp Bot.* 50: 1101-1114

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

64. Grimm, D.R., D. Denesh, J. Mudge, N.D. Young, and P.B. Cregan. 1999. Assessment of single nucleotide polymorphisms (SNPs) in soybean. *Plant Animal Genome* VII: 140.

65. Grusak, M.A., and S. Pezeshgi. Shoot-to-root signal transmission regulates root Fe(III) reductase activity in the dgl mutant of Pea. *Plant Physiol.* 1996. 110:329-334.

66. Guerinot, M. L., Y. Yi. 1994. Iron –nutritious, noxious, and not readily available. *Plant Physiol.* 104: 815-820.

67. Hamblim, M. T., M. G. Salas Fernandez, A. M. Casa, S. E. Mitchell, A. H. Paterson, and S. Kresovich. 2005. Equilibrium processes cannot explain high levels of short-

and medium-range linkage disequilibrium in the domesticated grass sorghum bicolor. *Genetics* 171: 1247–1256.

68. Hardy, Q. J., and X. Vekemans. 2002. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Eco. Res.* 2: 618-620.

69. Hansen, N. C., V. D. Jolley, S. L. Naeve, and R. J. Goos. 2004. Iron deficiency of soybean in the North Central U.S. and associated soil properties. *Soil Sci. Plant Nutr.* 50: 983-987.

70. Harrison P. M., and P. Arosio. 1996. The ferritins: molecular properties, iron storage function, and cellular regulation. *Biochemica et Biophysica Acta* 1275: 161-203.

71. Hell, R. U., and U. W. Stephan. 2003. Iron uptake and homeostasis in plants. *Planta* 216: 541-551.

72. Helms, T. C., B. D. Nelson, and R. J. Goos. 2005. Registration of LaMoure' Soybean. *Crop Sci.* 45: 410.

73. Helms, T.C, B. D. Nelson, and R. J. Goos. 2009. Registration of 'Ashtabula' Soybean. *Journal of Plant Registration* 3: 253-255.

74. Henriques, R., J. Jasik, M. Klein, E. Martinoia, U. Feller, J. Schell, M.S. Pais, and C. Koncz. 2002. Knock-out of Arabidopsis metal transporter gene IRT1 results in iron deficiency accompanied by cell differentiation defects. *Plant Mol. Biol.* 50: 587-597

75. Higuchi K, K. Kanazawa, N. K. Nishizawa, and S. Mori. 1996. The role of nicotianamine synthase in response to Fe nutrition status in Gramineae. *Plant and Soil* 178. 171-177.

76. Hill W. S, and B. S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations. *Theor Popul Biol.* 33:54–78.

77. Horlock, C. M., D. S. Teakle., and R. M. Jones. 1997. Natural infection of the native pasture legume, Glycine latifolia, by alfalfa mosaic virus in Queensland. *Australasian Plant Pathol.* 26: 115-116.

78. Hyde, B. B., A. J. Hodge, A. Kahn, and M. L. Birnstiel. 1963. Studies of phytoferritin: I. Identification and localization. *J. of Ultrastructure Res.* 9:248-258.

79. Hymowitz, T. 2003. Historical roots of the soybean in North America. Urbana, IL: National Soybean Research Laboratory. Avaliable on the World Wide Web: *http://www.nsrl.uiuc.edu/aboutsoy/*

80. Hymowitz, T. 1970. On the domestication of the soybean. *Econ Bot.* 24: 408-421.

81. 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. Impact of genetic bottlenecks on soybean genome diversity. *PNAS.* 103: 16666-16671.

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

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

84. Ishimaru, Y., M. Suzuki, T. Tsukamoto, K. Suzuki, M. Nakazono, T. Kobayashi, Y. Wada, S. Watanabe, S. Matsuhashi, M. Takahashi, H. Nakanishi, S. Mori, and N.K.

Nishizawa. 2006. Rice plants take up iron as an Fe^{3+}-phytosiderophores and as Fe^{2+}. *Plant J.* 43: 335-346.

85. Jakoby, M., H. Y. Wang, W. Reidt, B. Weisshaar, and P. Baur. 2004. FRU (BHLH029) is required for induction of iron mobilization genes in Arabidopsis thaliana. FEBS Lett. 577: 528-534.

86. Jannink, J. L., H. Iwata, P.R. Bhat, S. Chao, P. Wenzl, and G. J. Muehlbauer. 2009. Marker imputation in barley association studies. *The Plant Genome* 2:11-22.

87. Jessen, H. I., M. B. Dragonuk, R. W. Hintz, and W. R. Fehr.1988a. Alternative to breeding strategies for the improvement of iron efficiency in soybean. *J.Plant Nutr.* 11: 717-726.

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

89. Kaufmann, M.J., and J.W. Gerdemann. 1958. Root and stem rot of soybean caused by Phytophthora sojae n. sp. *Phytopathology* 48:201-208.

90. 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.

91. Kenworthy, W. J. 1989. Potential genetic contribution of wild relatives to soybean improvement. In Proceedings of the 4th World Soybean Research Conference, held at Buenos Aires, Argentina, 5-9 March 1989. Edited by A.J. Pascale. Assoc. Argentina de la Soja, Buernos Aires. Pp. 883-888.

92. Keshun, L. 1997. Soybean: Chemistry, Technology and Utilization. New York: Chapman & Hall. Pages 25-28.

93. Köhler, B, and K. Raschke. 2000. The delivery of salts to the xylem: three types of anion conductance in the plasmalemma of the xylem parenchyma of roots of Barley. *Plant Physiol.* 122: 243–254.

94. Kwok, P.Y. 2001. Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet*. 2:235-258.

95. Le Jean, M., A. Schikora, S. Mari, J. F. Briat, and C. Curie. 2005. A loss-of-function mutation in AtYSL1 reveals its role in iron and nicotianamine seed loading. *Plant J.* 44: 769-782.

96. Li, L., X. Cheng, and H.Q. Ling. 2004. Isolation and characterization of Fe (III)-chelate reductase gene LeFRO1 in tomato. *Plant Mo. Bio.* 54: 125-136.

97. Li, Y., Y. Li, W. Song, H. Kun, Z. Wang, W. Hou, Y. Zeng, and R. Wu. 2007. Estimation of Multilocus Linkage Disequilibria in Diploid Populations With Dominant Markers. *Genetics* 176: 1811-1821.

98. Lin, S. F., S. Cianzio, and R. Shoemaker. 1997. Mapping genetic loci for iron deficiency chlorosis in soybean. *Mol. Breed.* 3: 219-229.

99. Lin, S. F., R. Shoemaker, S. Cianzio, and D. Grant. 2000. Molecular characterization of iron deficiency chlorosis in soybean. *J. Plant Nutr*. 23. 1929-1939.

100. Ling, H. Q., G. Koch, H. Baumlein, and M. W. Ganal. 1999. Map-based cloning of chloronerva, a gene involved in iron uptake of higher plants encoding nicotianamine synthase. *Proc Natl Acad Sci.* USA 96: 7098–7103.

101. Liu, K., and S. V. Muse. 2005 PowerMarker: integrated analysis environment for genetic marker data. *Bioinformatics* 21: 2128–2129.

102. Loiselle, B. A., V. L. Sork, J. Nason, and C. Graham. 1995. Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae). *Am. J. Bot.* 82: 1420-1425.

103. Ma, J. F., and K. Nomoto. 1996. Effective regulation of iron acquisition in graminaceous plants. The role of mugenic acids as phytosiderophores. *Physiologia Plantaru* 97: 609-617.

104. Mackay, T. F. C. 2001. The genetic architecture of quantitative traits. *Annu.. Rev. Genet.* 35: 350-339.

105. Mackay, I., and W. Powell. 2007. Method for linkage disequilibrium mapping in crops. *Trends Plant Sci.* 12: 57-63.

106. Malosetti, M., C. G. van der Linden, B. Vosman, and F. A. van Eeuwijk. 2007. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. *Genetics* 175: 879-889.

107. Marschner, H. 1995. Mineral nutrition in higher plants. 2nd ed. Academic Press London, England.

108. Marschner, H., V. Romheld, and M. Kissel. 1986. Different strategies in higher plants in mobilization and uptake of iron. *Journal of Plant Nutrition* 9:3–7

109. Marentes, E., B. J. Shelp, R. A. Vanderpool, and G. A. Spiers. 1997. Academic Press. Retranslocation of boron in broccoli and lupin during early reproductive growth. *Physiologia Plantarum* 100: 389–99.

110. Maser, P., S. Thomine, J. I. Schroeder, J. M. Ward, K. Hirschi, H. Sze, I. N. Talke, A. Amtmann, F. J. M. Matthuis, D. Sanders, J. R. Harper, J. Tchieu, M. Gribskov, M. W. Persans, D. E. Salt, S. A. Kim, and M. L. Guerinot. 2001. Phylogenetic relationships within cation transporter families of Arabidopsis. *Plant Physiol.*126: 1646-1667.

111. Maughan, P. J, M. A. Saghai Maroof, and G. R. Buss.1995. Microsatellite and amplified sequence length polymorphisms in cultivated and wild soybean. *Genome* 38: 715-723.

112. McKie, A. T., P. Marciani, A. Rolfs, K. Brennan, K. Wehr, D. Barrow, S. Mirret, A. Bomford, T. J. Peters, F. Farzaneh, M. A. Hediger, M. W. Hentze and R .J. Simpson. 2000. A novel duodenal Iron-Regulated Transporter. IREG1, implicated in the basolateral transfer of iron to the circulation. *Mol. Cell.* 5: 299-309.

113. Messina, M. J. 1999. Legumes and soybean: overview of their nutritional profiles and health effects. *American Journal of Clinical Nutrition* 70: 439S-450S.

114. Mengel, K. 1994. Iron availability in plant tissues- iron chlorosis on calcareous soils. *Plant and Soil* 165: 275-283.

115. Michael, K. L., L. C. Taylor, S. L. Schultz, and D. R. Walt. 1998. Randomly ordered addressable high-density optical sensor arrays. *Anal. Chem.* 70: 1242-1248.

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

117. Morgante, M., A. Rafalski, P. Biddle, S. Tingey, and A. M. Olivieri. 1994. Genetic mapping and variability of seven soybean simple sequence repeat loci. *Genome* 37: 763-769

118. Morrell, P. L., D. M. Toleno, K. E. Lundy, and M. T. Clegg. 2005. Low levels of linkage disequilibrium in wild barley (Hordeum vulagare ssp. spontaneum) despite high rates of self-fertilzation. *Proc. Natl. Acad. Sci.* U S A 102: 2242-2447.

119. Mori, S. 1999. Iron acquisition by plants. *Curr. Opin. Plant Biol.* 2: 250-253.

120. Morse, W. J. 1918. "The Soy Bean: its Culture and Uses." *USDA Farmers' Bulletin*, No. 973. Washington, D.C.: U.S. Department of Agriculture,

121. Morse, W. J., J. L. Cartter and L. F. Williams. 1949. Soybeans: Culture and varieties. U.S. Dep. Agric. Farmers' Bull No. 1520: 1-38.

122. Mukherjee, I., N. H. Campbell, J. S. Ash, and E. L. Connolly. 2006. Expression profiling of the Arabidopsis ferric chelate reductase (FRO) gene family reveals dfferential regulation by iron and copper. *Planta* 223: 1178-1190.

123. Nikolic, M., and V. Romheld. 1999. Mechanism of Fe uptake by the leaf symplast: Is Fe inactivation in leaf a cause of Fe deficiency chlorosis? *Plant Soil* 215:229-237

124. Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory, J. Hagenblad, M. Kreitman, J. N. Maloof, T. Noyes, P. J. Oefner, E. A. Stahl, and D. Weigel. 2002. The extent of linkage disequilibrium in Arabidopsis thaliana. *Nat. Genet.* 30: 190-193

125. Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajlan, H. Zheng, E Bakker, P. Calabrese, J. Gladstone, R. Goyal, M. Jakobsson, S. Kim, Y. Morozov, B. Padhukasahasram, V. Plagnol, N.A. Rosenberg, C. Shah, J. D. Wall, J. Wang, K. Zhao, T. Kalbfleisch, V Schulz, M. Kreitman, and J. Bergelson. 2005. The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol.* 3: e196

126. Ohashi, H. 1982. Glycine max (L.) Merr. subsp. soja (Sieb and Zucc.). Journal of Japanese Botany57:30.

127. Onken, A. B., and H. J. Walker. 1966. An investigation of iron chlorosis in grain sorghum grown on a calcerous soil of the High Plains of Texas. Texas Agric. Exp. Stn. MP-823.

128. Orf, J. H., and R. L. Denny. 2004. Registration of 'MN0302' Soybean. Crop Science.44: 692-693

129. O'Rourke J. A., R. T. Nelson, G. David, J. Schmutz, J. Grimwood, S. Cannon, C. P. Vance, M. A. Graham, and R. C. Shoemaker.2009. Integrating microarray analysis and the soybean genome to understand the soybeans iron deficiency response. BMC Genomics. 10: 376-292.

130. Patterson, N., A. L. Price, and D. Reich. 2006. Population Structure and Eigenanalysis. PLoS Genetics. 2: 2074–2093.

131. Pioneer International Inc. 2009. News release. http://www.pioneer.com/web/site/portal/menuitem.e5962e5b31b28c754a624a62d10093a0/

132. Powell, W., M. Morgante, C. Andre, M. Hanafey, J. Vogel, S. Tingey and A. Rafalski. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol. Breeding 2: 225-238.

133. Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904–909

134. Pritchard, J. K, and N. A. Rosenberg.1999. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 65: 220-28.

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

136. Pyhajarvi, T., M. R. García-Gil, T. Knurr, M. Mikkonen, W. Wachowiak, and O. Savolainen. 2007. Demographic history has influenced nucleotide diversity in European Pinus sylvestris populations. Genetics177: 1713–1724.

137. Rafalski, A., M. Morgante. 2004. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. 20: 103-111.

138. Rakshit, S., A. Rakshit, H. Matsumura, Y. Takahashi, Y. Hasegawa, A. Ito, T. Ishii, N. T. Miyashita, R. Terauchi. 2007. Large-scale DNA polymorphism study of Oryza sativa and O. rufipogon reveals the origin and divergence of Asian rice. Theor. Appl. Genet. 114: 731–743.

139. Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt et al. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98: 11479–11484.

140. Riggs, R. D., S. Wang, R. J. Singh, and T. Hymowitz. 1998. Possible transfer of resistance to Heterodera Glycines from Glycine tomentella to Glycine max. J. Nematol. 30: 547-552.

141. Risch, N. J. 2000. Searching for genetic dertermination in the new millennium. Nature 405: 847-855.

142. Robinson, N. J., C. M. Procter, E. L. Connolly, and M. L. Guerinot. 1999. A ferric chelate reductase for iron uptake from soils. Nature. 397: 694-697.

143. Rongwen, J., M. S. Akkaya, A. A. Bhagwat, U. Lavi, and P.B. Cregan. 1995. The use of microsatellite DNA markers for soybean genotype identification. Theor. Appl. Genet. 90:43-48.

144. Rooney, W. L, and C. W. Smith. 2000. Techniques for developing new cultivars. 329–347 in Sorghum, edited by C. W. Smith and R. A. Frederiksen. John Wiley & Sons, New York.

145. Rostoks, N., S. Mudie, L. Cardle, J. Russell, L. Ramsay, A. Booth,J. Svensson, S. Wanamaker, H. Walia, E. Rodriguez, P. Hedley, H. Liu, J. Morris, T. Close, D. Marshall, and R. Waugh. 2005. Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol. Genet. Genomics.274:515–527.

146. SAS, 9.1.3 2009 SAS. Stastical Analysis Software for windows, 9.1.3 Edition Cary, NC. USA.

147. Scheet, P., and M. Stephens.2006. A fast and flexible statistical model for large-scale Population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet.78:629–644

148. Schmidt, W. 1999. Mechanism and regulation of reduction based iron uptake in plants. New Phytol. 141: 1-26.

149. Schmid, K. J., T. R. Sorensen, R. Stracke, O. Torjek, T. Altmann et al., 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. Genome Res. 13: 1250-1257

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

151. Singh, B. B., S. C. Gupta, and B. D. Singh. 1974. Sources of field resistance to rust and yellow mosaic diseases of soybean. Indian J. Genet. Plant Breed. 34: 400-404.

152. Singh, R. J., and T. Hymowitz. 1988. The genomic relationship between Glycine max (L.) Merr. and G. soja Sieb. And Zucc. As revealed by pachytene chromosome analysis. Theoretical and applied genetics. 76: 705-711.

153. Smith, B. N. 1984. Iron in higher plants: storage and metabolic role. Journal of Plant Nutrition 7: 759-66

154. Smith, C.W. 1995. Crop production: evolution, history and technology. Wiley, New York 352-359.

155. Stich, B., J. Mohring, H. P. Piepho, M. Heckenberger, E. S. Buckler, and A. E. Melchinger .2008. Comparison of mixed-model approaches for association mapping.Genetics. 178:1745–1754

156. Stephan, U. W., I. Schmidke, V. W. Stephan, and G. Scholz. 1996. The nicotianamine molecule is made-to-measure for complexation of metal micronutrients in plants. Biometals 9: 84-90

157. Stephan, U. W. 2002. Intra- and intercellular iron trafficking and subcellular compartmentation within roots. Plant and soil. 241: 19-25.

158. Stracke, S., D. Perovic, N. Stein, T. Thiel and A. Graner. 2003. Linkage disequilibrium in barley. 11th Molecular Markers Symposium of the GPZ, http://meetings.ipkgatersleben.de/ moma2003/index.php.

159. Tabangin, M. E., G. W. Jessica and J. M. Lisa. 2009. The effect of minor allele frequency on the likelihood of obtaining false positives. BMC. 3: S7-S41.

160. Takagi, S. 1976. Naturally occurring iron-chelating compounds in oat- and rice root washing. I. Activity measurement and preliminary characterization. Soil Sci. Plant Nutr. 22: 423-433.

161. Takahashi, M. T., H. Nakanishi, S. Kawaski, N. K. Nishizawa, and S. Mori. 2001. Enhanceed tolerance of rice to low iron availability in alkaline soils using barley nicotiananmine aminotransferase genes. Nat Biotrchnol. 19: 466-469.

162. Takahashi, M., Y. Terada, I. Nakai, Y. Nakanishi, E. Yoshimura, S. Mori, and N. K. Nishizawa.2003.Role of nicotianamine in the intracellular delivery of metals and plant reproductive development. The Plant Cell15: 1263–1280.

163. Tanksley, S. D., N. D. Young, A. H. Paterson and M. W. Bonierbale. 1989. RFLP Mapping in Plant Breeding: New Tools for an Old Science. *Nature Biotechnology* 7: 257-264.

164. TASSEL 2009 User manual, trait analysis by association, evolution and linkage. *www.maizegenetics.net/tassel*

165. Tautz, D. and M. Renz. 1984 Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res.* 12: 4127-4138.

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

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

168. Thomine, S., R. Wang, J. M. Ward, N. M. Crawford, and J. I. Schroeder. 2000. Cadmium and iron transport by members of plant transporter gene family in Arabidopsis with homology to NRAMP genes. *Proc. Nat. Acad Sci.* USA 97: 4991-4996.

169. Tiffin, L. O. 1966. Iron translocation: plant culture, exudates sampling, iron citrate analysis. *Plant Physiol*. 45: 280-283

170. Tommasini,L.,T. Schnurbusch, D.Fossati, F. Mascher, and B. Keller. 2007. Association mapping of Stagnospora nodorum blotch resistance in modern European winter wheat varieties. *Theor Appl Genet*. 115: 697-708

171. Tsuchihashi, Z., and N. C. Dracopoli. 2002. Progress in high throughout SNP genotyping methods. *The Pharmacogenomics J.* 2: 103-110.

172. Tylka, G. 2001. SCN responsible for yellow soybean fields. Integrated Crop Manage. IC-486 (21). Iowa State Univ Press, Ames

173. Vasconceles, M., H. Eckert, V. Arahana, G. Graef, M. A. Grusak, and T. Clemente. 2006. Molecular and phenotypic characterization of transgenic soybean expressing the Arabidopsis ferric chelate reductase gene, FRO2. *Planata* 225: 1116-1128.

174. Vert, G., J. F. Briat, C. Curie. 2001. Arabidopsis IRT2 gene encodes a root periphery iron transporter. *Plant J.* 26: 181-189.

175. Vert, G., N. Grotz, F. Dedaldechamp, F. Gaymard, M. L. Guerinot, J. Briat, and C. Curie. 2002. IRT1, an Arabidopsis transporter essential for iron uptake from the soil and for plant growth. *The Plant Cell* 14: 1223-1233.

176. Vignal, A., D. Milan, M. S. Cristobal and A. Eggen. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* 34: 275-305.

177. Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. V. Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M.Kuiper, and M. Zabeau. 1995. AFLP: a new technique for DNA fingerprinting. *Nucl. Acids Res.* 23: 4407-4414.

178. Waldo, G.S., E. Wright, Z. H. Whang, J. F. Briat, E. C. Theil, and D. E. Sayers.1995. Formation of theferritin iron mineral occurs in plastids. *Plant Physiol.* 109:797-802

179. Walter, S.O., and R. S. Aldrich. 1970. Modern soybean production. *The Farm Quaterly.* Publ., Cincinnati,OH.

180. Wang, J., P. E. McClean, R. Lee, J. Goos, and T. Helms. 2008. Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. Merr.) advanced breeding lines. *Theor. Appl. Genet.* 116: 777–787

181. Waters, B. M., D. G. Blevins, and D. J. Eide. 2002. Characterization of FRO1, a pea-ferric-chelate reductase involved in root acquisition. *Plant Physiol.* 129: 85-94.

182. Waters, B. M., H. Chu, R. J. Didonato, L. A. Roberts, R. B. Eisley, B. Lahner, D. E. Salt, and E. L. Walker. 2006. Mutations in Arabidopsis Yellow-Stripe-Like1 and Yellow Stripe-Like3 reveal their roles in metal ion homeostasis and loading of metal ions in seeds. *Plant Physiol.* 141: 1446–1458

183. Weber, A. L., Q. Zhao, M. D. McMullen, and J. F. Doebley. 2009. Using association mapping in teosinte to investigate the function of maize selection-candidate genes. *PLoS ONE* 4: e8227.

184. Weiss, M. G. 1943. Inheritance and physiology of efficiency in iron utilization in soybeans. *Genetics* 28: 253

185. Williams J. G. K., A. R. Kubelik, K. J. Livak, J. A. Rafalski, and S. V. Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acid Res.* 18: 6531-6535

186. Windish, L. G. 1981. The Soybean Pioneers, Trailblazers, Crusaders and Missionaries. M&D Printing, Henry Illinois.

187. Wright, S. 1951. The genetic structure of populations. *Annals of Eugenics* 15: 323-354.

188. Yuan, Y. X., J. Zhang, D. W. Wang, and H. Q. Ling. 2005. AtbHLH29 of Arabidopsis thaliana is a finctional ortholog of tomato FER involved in controlling iron acquisition in strategy I plants. *Cell Research* 15: 613-621.

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

190. Zhao, K., M. J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and N. Magnus. 2007. An Arabidopsis example of association mapping in structured samples. *PLoS Genet.* 3: 71–82.

191. 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 polymorphism in soybean. *Genetics* 163:1123–1134.