5-2023

An Investigation of Seed Hardness and Seed Coat Color Values in Natto Soybean

Joshua Winter
University of Arkansas, Fayetteville

An Investigation of Seed Hardness and Seed Coat Color Values in Natto Soybean

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Crop Science

by

Joshua Winter
John Brown University
Bachelor of Science, 2020

May 2023
University of Arkansas

This thesis is approved for recommendation to the Graduate Council

Ehsan Shakiba, Ph.D.
Thesis Director

Leandro Mozzoni, Ph.D.
Committee Member

Reuben Ceballos, Ph.D.
Committee Member

Trenton Roberts, Ph.D.
Committee Member

Bo Zhang, Ph.D.
Committee Member

Ainong Shi, Ph.D.
Committee Member

ABSTRACT

Natto is a specialty fermented soyfood made from small-seeded (<10g 100 seeds^{-1}) soybean varieties. Seed hardness and seed coat color are important seed traits that determine the texture and appearance of natto and are thus valuable to breeders. Prior research has identified quantitative trait loci (QTL, hereafter) for seed hardness, but its nature as a quantitative trait heavily influenced by the environment means that it is still poorly understood. Prior research has identified the primary genetic components of seed coat color using simple visual inspection, but few studies have investigated the usefulness of more quantitative measurements, such as the color space coordinates developed by Commission Internationale de l'Eclairage.

The objectives of this research were 1) to assess seed hardness and the seed coat color components of lightness, chroma, and hue in a diverse variety of genotypes to determine suitable parents for natto breeding, 2) to evaluate the environmental influence on these traits, 3) analyze the genetic diversity of the genotypes used in this study, 4) perform genome-wide association studies (GWAS) to identify single nucleotide polymorphisms (SNPs) that may be used for marker-assisted selection (MAS) or genomic selection (GS) of seed hardness and seed coat color, and 5) to compare the effectiveness of different GWAS models for identifying SNPs associated with the traits of interest. An association panel was assembled using 168 natto accessions from the USDA soybean germplasm, 51 natto breeding lines and 49 conventional breeding lines from the University of Arkansas, and 49 natto breeding lines from Virginia Tech. All genotypes were grown in 2021 as an augmented block design with four single-replication blocks, each grown in four different locations in Arkansas. DNA was isolated from young leaf tissue of 285 lines and genotyped at the Soybean Genomics and Improvement Laboratory using the SoySNP50k platform (Illumina, Inc., San Diego, CA). Phenotypic data were collected and analyzed in JMP Pro 16, genetic diversity and population structure were calculated using GAPIT, and TASSEL

was used to perform GWAS using 32,724 SNPs. Association analyses were conducted using the general linear, mixed linear, and single marker regression models, and the significance threshold was an LOD > 3. ANOVA conducted on aggregated data showed a significant environmental effect in all locations. A student t-test indicated a sufficient genetic diversity to perform GWAS on the individual location tests.

Phenotypic analysis for identification of suitable natto breeding genotypes revealed six high-performing genotypes for seed hardness, nine for lightness and chroma, and 12 for hue. Two genotypes, PI 458281 B and PI 603713, were found to have optimal phenotype for multiple traits, and were identified as potential natto breeding parents. Genetic diversity analysis identified three distinct sub-populations within the association panel. Seed source, region of origin, variety, and level of inbreeding had no significant effect on the traits of interest. GWAS identified 11 SNPs for hardness, 15 for lightness, six for chroma, and nine for hue. One of the seed hardness SNPs, ss715579472, was colocalized to Chr 1 within 600 kbp of *Ha2*, a seed hardness QTL identified and confirmed by prior research. These results will be useful for developing new natto cultivars through marker-assisted selection.

ACKNOWLEDGEMENTS

I would like to thank Dr. Leandro Mozzoni and Dr. Ehsan Shakiba for their role as my advisors and for their guidance throughout my time in this program. I would also like to thank the members of my committee, Dr. Reuben Ceballos, Dr. Trenton Roberts, Dr. Bo Zhang, and Dr. Ainong Shi, for their assistance and guidance in helping me pull my project together in the face of many setbacks and complications.

I would like to extend a special thanks to Dr. Jeff Edwards and the Soybean Breeding Program technicians. Thank you for being so generous and accommodating during a very turbulent and difficult time in my life. I would not have finished this degree without your kindness, and for that I am incredibly grateful.

I would also like to extend my gratitude to the University of Arkansas Crop, Soil, and Environmental department, the Arkansas Soybean Promotion Board, and the staff at the USDA Soybean Genomics and Improvement Laboratory for providing the education, funding, and professional support necessary to see this project through.

DEDICATION

To my grandfather and mentor, Wendell Alumbaugh.

You taught me to look down the trail and to always ask why.

For that, I owe this accomplishment to you.

Thank you.

TABLE OF CONTENTS

CHAPTER I.....1

Literature Review.....2

Bibliography..... 10

CHAPTER II.....16

Abstract.....17

Introduction.....19

Materials and Methods.....22

Results.....26

Discussion.....31

Tables and Figures.....38

Bibliography..... 58

CHAPTER I

Literature Review

Overview of Soybean

The soybean (*Glycine max* (L.) Merr.) is a nutritionally and economically important oilseed legume. The United States is the leading soybean producer and second-leading exporter worldwide (ers.usda.gov), with a total export market valued at $27.4 billion as of 2021 (fas.usda.gov). Soybean accounts for 90% of oilseed production in the United States and is the second most widely grown crop after maize (*Zea mays* (L.)) (nass.usda.gov). In the state of Arkansas, soybean is grown across 1.3 million hectares and generates around $1.7 billion in annual revenue, making it the most valuable row crops in the state (uaex.uada.edu).

Soybean is a versatile crop, serving both as a food source for humans and livestock as well as a source of raw materials used for manufacturing. Soybeans account for approximately 70% of vegetable protein consumed worldwide (Kumar, Rani, and Chauhan, 2010; Ali, 2010). The protein content of soy is high enough quality that it can serve as the sole source of amino acids in one's diet, presuming that all minimum calorie requirements are met (Beer, et al., 1989; Rizzo and Baroni, 2018). Soybean is also valued for its oil products. Soybean oil is a colorless, odorless, flavorless oil extracted from whole soybeans. Nutritionally, soybean oil is low in cholesterol and contains vitamin E and omega-3 fatty acids (Raghuvanshi and Bishit, 2010).

Outside of food production, soybean byproducts are commonly used in manufacturing and industry. Soy protein's structure and foaming capacity enables it to easily form gel matrices in solution, lending it to a wide variety of applications in manufacturing and industrial processes. Soy protein concentrates and byproducts are used as an additive to inks, paints, adhesives, pesticides, synthetic fabrics, and gels. Soybean oil is similarly versatile; it is an effective industrial lubricant, and is manufactured into candlewax, crayons, hydraulic fluid, cushioning foam, and fiberglass. Soybean oil is also used as an additive in oil-based products such as

pesticide sprays, printer toner, and industrial solvents (Schmitz, et al., 2008; Raghuvanshi and Bishit, 2010).

Natto

Natto is a fermented soyfood originating from and popular in Japan (Zhang et al, 2008). Natto is described as possessing a soft, sticky texture and a strong, somewhat bitter aroma and taste (Schutleff and Aoyagi, 2012). The market for natto is somewhat small compared to more widely consumed soyfoods such as tofu and soymilk (Yoshikawa, et al., 2013). However, approximately 75% of soybeans used in natto production are imported from the United States, which makes growing and breeding natto soybeans a profitable endeavor for U.S.-based farmers and breeders (North Dakota Soybean Council, 2018).

Natto-type soybeans differ in several ways from other varieties grown for industrial and agricultural feed use. One of the primary distinguishing traits of natto soybeans is seed size. To be desirable for making natto, soybean seeds must possess a seed weight below 80 mg per seed, or below 10 g 100^{-1} seeds (Geater, Fehr, and Wilson, 2000). This is because small-seeded cultivars typically have other chemical seed traits that are favorable to the fermentation of natto. Small-seeded soybean varieties typically possess higher levels of free sugars than larger varieties, as well as lower protein and higher oil content (Zhang et al., 2010).

Small-seeded varieties with low sucrose and high stachyose content are ideal for natto production. The bacterium responsible for fermentation, *Bacillus subtilis var. natto* consumes both sucrose and stachyose in the fermentation process; however, the larger stachyose molecule takes longer for the bacterium to break down. If sucrose levels are too high, the bacterium may consume the sugars in the soybean too quickly, raising the internal temperature of the beans and reducing the effectiveness of fermentation (Taira, 1990).

Natto Sensory Characteristics

Traditional natto must possess certain sensory characteristics to be considered high-quality. The primary characteristics are flavor, aroma, stickiness, appearance, and texture (Wei and Chang, 2004; Yoshikawa, et al., 2013). Traditionally, these characteristics are determined by taste testing, rather than analytical laboratory testing procedures. Prior studies have made efforts to correlate these traditional characteristics to specific chemical characteristics of natto seeds, enabling more precise measurement for research and breeding purposes. The specific benchmarks for what is considered "high quality" natto is not standardized; rather, it is set by natto manufacturers based on market research and perceived consumer preference (Hosoi and Kiuchi, 2003).

Flavor and Aroma

Possessing the right flavor and aroma is highly important for natto; it should have a characteristic flavor that has variously been described as unique, sweet, astringent, earthy, and strong, while its aroma has been described as sweet, pungent, ammoniac, and strong (Wei, Wolf-Hall, and Chang, 2001; Hosoi and Kiuchi, 2003; Wei and Chang, 2004; Yoshikawa et al., 2013). There are many factors that affect the flavor and aroma of natto. One such factor, and one of the most critical, is the ammonia content of natto after fermentation (Taira, 1990; Yoshikawa, et al., 2013). Ammonia is a natural byproduct of the fermentation process, resulting primarily from the deamination of glutamate and breakdown of urea during secondary fermentation (Kada et al, 2008). If the ammonia content is too high, however, it can greatly reduce the natto's flavor and kill the *Bacillus* early in the fermentation process (Hosoi and Kuichi, 2003). The amount of ammonia produced by the fermentation has been correlated to the protein content of the soybeans used as a substrate. Fermenting seeds with higher protein content provides the bacteria with a

larger quantity of nitrogen, which results in a higher ammonia content, giving the natto an unpleasant flavor and aroma. Therefore, natto cultivars with a lower protein content are more desirable for natto production (Taria, 1990; Geater, Fehr, and Wilson, 2000; Yoshikawa, et al., 2013).

Stickiness

One of natto's signature characteristics is the presence of a viscous mucus-like substance, which causes the beans to stick string together when separated. During the fermentation process, *Bacillus subtilis* var. *natto* produces poly-γ-glutamic acid (γ-PGA), a polypeptide composed of long chains of glutamic acid, which subsequently coats the fermented beans. Prior research has identified glucose and excess L-glutamate levels in the fermentation medium as the primary determiner of γ-PGA concentration post-fermentation, which in turn is the primary determiner of natto's stickiness (Yao, et al., 2009).

Texture

To be considered desirable, natto must possess a soft, chewy texture (Wei, Wolf-Hall, and Chang, 2001). Seed hardness is the primary trait affecting natto texture due to the impact it has on the seed's water absorption and cookability (Zhang et al, 2008; Geater, Fehr, and Wilson, 2000). Seed hardness is also undesirable for non-food soybeans. Hard seeds are more durable and will retain their quality for longer periods of time; however, hard seeds are also resistant to germination, making it undesirable for most breeding programs (Zhou et al., 2010).

Seed hardness is a complex quantitative trait that is affected by numerous other factors. Seed coat permeability has been found to have a negative correlation with seed hardness (Geater, Fehr, and Wilson, 2000; Zhang et al, 2008; Orazaly et al, 2015). More permeable seed coats allow the seeds to absorb more water when soaked, which results in a softer product once

cooked. Seed coat permeability is itself affected by the seed coat thickness, the concentration of calcium (Ca) in the seed coat, and the seed coat's pigmentation (Qutob et al., 2008; Zhang et al., 2008; Yoshikawa et al, 2013). Seed hardness is also affected by seed composition. A study by Yoshikawa et al. (2013) found that protein concentration was negatively correlated with the texture of natto. These results are in line with Taira (1990), who found that harder seeds resulted in natto with a higher concentration of ammonia, which is the result of the deamination of glutamate and the breakdown of urea by urease during secondary fermentation (Kada et al, 2008). Seed hardness is heavily affected by the environment as well. Geater, Fehr, and Wilson (2000) found that seed hardness was significantly affected by the growing year and location, regardless of genotype. Prior studies suggest that rainfall, temperature, and humidity all play a role in determining seed hardness (Argel and Paton, 1999).

Several studies have investigated the genetic component of seed hardness in soybeans. Zhang et al (2008) identified two QTL for seed hardness, *Ha1* and *Ha2*, on chromosomes 19 and 1, respectively, and collectively accounted for 7.8% of phenotypic variation. A later study by Orazaly et al. (2015) confirmed the location of *Ha1* and identified another QTL for seed hardness, *Ha3*, on chromosome 16. A separate study by Hirata et al. (2014) mapped two additional QTL, *qHbs3-1* and *qHbs6-1*, to chromosomes 3 and 6, which accounted for 44.3% and 11.6% of phenotypic variation.

Appearance

It is important for natto to have an appetizing appearance. For this, natto manufacturers prefer soybeans with a light-yellow seed coat and a clear or yellow hilum, giving the seeds a uniform appearance (Hosoi and Kiuchi, 2003). Historically, seed coat has not been a commercially useful trait for soy breeders outside of being an easy visual marker of

hybridization. More recently, however, researchers have taken an interest in the antioxidant properties of the anthocyanins and proanthocyanins that make up the majority of pigments in the soybean seed coat (Yang et al, 2010).

Seed coat color is a complex genetic trait that can present five distinct phenotypes: yellow, green, brown, black, and multicolor (Song et al., 2016). The trait is primarily controlled by five loci *(I, T, R, O,* and W_1) which were identified through classical genetic study (Yang et al, 2010; Song et al, 2016). The T, R, and I loci contribute most directly to seed coat color. The R and T loci control the synthesis of anthocyanins and proanthocyanins in the seed coat, with dominant and recessive allele combinations producing specific coat color phenotypes: black (*R, T*), brown (*r, T*), imperfect black (*R, t*), and buff (*r, t*) (Sedna et al, 2012). The O and W_1 loci indirectly affect the hilum color in the presence of the other loci (Palmer et al., 2004).

The I locus controls the presence and distribution of anthocyanins and proanthocyanins in the seed coat (Todd and Vodkin, 1993). The I locus possesses four alleles (*I,* i^i, i^k, and *i*). The *I* allele inhibits the CHS gene, an important gene in the pigment production pathway, via RNA silencing, resulting in a uniformly yellow soybean seed. The i^i and i^k alleles only partially inhibits the CHS gene, resulting in a yellow seed coat but pigmented hilum, and the *i* allele does not inhibit the CHS gene at all (Sedna et al, 2012).

Efforts have been made to precisely map the loci governing seed coat color. A study by Yang et al (2010) mapped the location of the *O* and W_1 loci to chromosomes 8 and 13, respectively, and identified genetic architecture for the *I* and *T* loci on chromosomes 8 and 6, respectively. A later study by Song et al. (2016) confirmed the location of the *I*, *T*, and *R* loci and mapped a novel locus. The novel locus, *qSC1*, was mapped to chromosome 1, and was identified as controlling for the green and yellow phenotypes.

Conventional Soybean Breeding

The soybean is a self-pollenating annual diploid with 20 chromosomes (2n = 40). Soybeans were originally domesticated in East Asia from the wild soybean (*Glycine soja*), although the precise time and geographical region are disputed (Qiu and Chang, 2010). Current archaeological findings estimate that domestication first occurred between 6000 and 9000 years ago, with evidence of domestication events occurring in China, Japan, and the Korean peninsula. Current research suggests that the earliest of these events occurred in China (Sedivy, Wu, and Hanzawa, 2017). Soybeans were introduced in North America in 1765, and coordinated breeding efforts for soybean began in the 1920s (Carter et al., 2004).

The pedigree method is the most common approach to soybean breeding. In the pedigree method, the breeder selects superior progeny derived from two carefully selected parent genotypes. The breeder makes these selections from successive generations of segregating genotypes until they achieve a nearly homozygous population, which typically occurs in the F_5 or F_6 generation. Once homozygosity is achieved, the breeder typically makes selections between multiple family lines to produce a superior cultivar. Throughout this process, the breeder keeps a record of parent-progeny relationships and notable line traits, so that they can track which parent combinations resulted in superior cultivars. These cultivars, in turn, become parents for future crosses, usually with lines possessing varied pedigrees (Allard, 1960).

Soybean Germplasm Diversity

A major concern facing breeders worldwide is the lack of genetic diversity in crop germplasms. Crop breeders rely heavily on pedigree-based breeding to produce highly uniform, high-performance lines (Allard, 1960). While this is not a problem, it can result in genetic

bottlenecks, resulting in low levels of diversity in crop germplasms. When diversity levels get too low, the risk of mass susceptibility to disease, environmental hazards, and pests rises, putting global food supplies at risk (National Research Council, 1993).

The genetic diversity of soybeans is relatively low. Hyten et al. (2006) found that the soybean has suffered from a series of genetic bottlenecks during domestication, resulting in a 50% loss of diversity between wild soybean (*Glycine soja*) and elite *Glycine max* cultivars. Carter et al. (2004) notes that, despite the limited ancestral base of domesticated soybean, the intraspecies diversity of soybeans remains rather high. However, a substantial portion of this initial diversity has been lost due to cultivar development, with the North American germplasm maintaining a consistent coefficient of parentage of 0.17-0.19 (Carter et al., 2004).

Of the over 47,000 accessions available in the soybean germplasm, it is estimated that only 1,000 of them have been used in cultivar development programs (Carter et al., 2004). Introducing genes from the 46,000 exotic accessions shows great potential for introducing genetic diversity into cultivar development programs. Several studies have demonstrated the viability of this method, successfully introgressing novel QTL for yield and other agronomic traits from exotic soybean accessions (Hegstad et al., 2019; Concibido et al., 2003).

Genome-Wide Association Studies

Genome-wide association study, or GWAS, is a statistical method for identifying quantitative trait loci (QTL) and genes by statistically associating their phenotypic expression with the presence of molecular markers. GWAS was originally developed by Hirschhorn and Daly (2005) to study the genetic basis of human disease but has been used by plant breeders to facilitate marker-assisted selection for a wide variety of agronomical and disease resistance traits. A meta-study by Shook et al. (2022) of 73 published soybean GWAS identified 59

candidate genes from these studies. Of these, 17 were for agronomic traits, including yield, maturity date, stem termination, and lodging; 19 were for seed traits, including seed weight, seed quality, and seed compositional traits; and 33 were associated with disease and pest traits, including aphid, stem canker, and soybean cyst nematode.

While GWAS is a potentially powerful technique for genetic mapping, it presents several challenges that can limit its usefulness. The most significant problem with GWAS is a tendency to produce high numbers of false positives. This is caused by using a large number of markers from closely related populations, which creates spurious relations resulting from relatedness rather than genuine association with the trait of interest (Long and Langley, 1999). To work around this issue, researchers engaging in GWAS must utilize sufficiently large and diverse populations when collecting their data and conduct linkage disequilibrium (LD) and kinship analysis in order to minimize the likelihood of these spurious associations (Brzyski, et al., 2017).

Marker-Assisted Selection

Marker-assisted selection (MAS) is a molecular breeding method that relies on molecular markers for selecting parents for breeding. Molecular markers are polymorphisms present in an organism's genome that are associated with a gene or QTL. Common molecular markers include single-nucleotide polymorphisms (SNP), simple sequence length polymorphisms (SSR), and restriction fragment length polymorphisms (RFLP), among many others. Marker-assisted selection was developed as a method of overcoming challenges that arise with conventional breeding. Many desirable traits, such as abiotic stress tolerance, are heavily affected by the growing environment, and can be difficult to breed for due to low heritability and costly phenotypic methods. Using markers to identify the specific QTL for these traits enables breeders to select parents with an optimal combination of alleles that might otherwise be masked by an

unfavorable phenotype (Francia, et al., 2005).

Marker-assisted selection has been effectively utilized by breeders to produce high-performing lines in several traits. Yield-related QTL identified by GWAS and other QTL-identifying methods have been utilized in MAS-based breeding approaches to develop high-yielding lines. [varying levels of efficacy – environmentally dependent] (Ravelombola et al., 2021; Yamaguchi et al., 2021). Seed weight, protein content, and carbohydrate content have also been the subject of MAS-based breeding approaches.

Marker-assisted selection could be a useful tool for introducing genetic diversity into a breeding program. While many exotic accessions yield fewer bushels per acre than elite inbred lines, they may still possess favorable alleles. However, introgressing such alleles into the population can present a challenge; while the introduction of genetic diversity is helpful for maintaining continual improvement in breeding programs, other factors such as unforeseen environmental and epistatic interactions may delay the expression of the superior alleles (Reyna and Sneller, 2001). Despite these challenges, there have been some studies that have effectively utilized this method for introducing exotic germplasm materials into commercial cultivars (Concibido et al., 2002; Ru and Bernando, 2019).

BIBLIOGRAPHY

Allard, R.W. 1960. *Principles of Plant Breeding*. 1st edition. Wiley Press, Hoboken, NJ. ISBN: 978-0471023098

Ali, N. 2010. Soybean Processing and Utilization. *The Soybean: Botany, Production, and Uses, Chapter 16*. Ed. Singh, G. CABI, Cambridge, MA. ISBN: 978-1-84593-644-0.

Argel, P.J. and Paton, C.J. 1999. Overcoming legume hardseededness. *Forage Seed Production: Tropical and Subtropical Species, Vol. 2*. Eds. Loch, D.S. and Ferguson, J.E. CABI, Wallingford, UK. ISBN: 978-0-85199-191-7

Beer, W.H., Murray E., Oh, S., Pederson, H.E., Wolfe, R.R., and Young, V.R. 1989. A long-term metabolic study to assess the nutritional value and immunological tolerance to two soy-protein concentrates in adult humans. *American Journal of Clinical Nutrition*. 50:997-1007. DOI: 10.1093/ajcn/50.5.997

Brzyski, D., Peterson, C.B., Sobczyk, P., Candes, E.J., Bogdan, M., and Sabatti, C. 2017. Controlling the Rate of GWAS False Discoveries. Genetics, 205(1): 61-75. DOI: 10.1534/genetics.116.193987

Carter, T.E., Nelson, R.L., Sneller, C.H., and Cui, Z. 2004. Genetic Diversity in Soybean. *Soybeans: Improvement, Production, and Uses, Chapter 8.* 3rd Edition. Boerma, H.R. and Specht, J.E. (Eds.). ASA, CSSA, and SSSA, Madison, WI.

Concibido, V., La Vallee, B., Mclaird, P., Pineda, N., Meyer, J., Hummel, L., Yang, J., Wu, K., and Delannay, X. 2003. Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. Theoretical and Applied Genetics, 106: 575-582. DOI: 10.1007/s00122-002-1071-5

Crop Production 2021 Summary. 2022. USDA, National Agricultural Statistics Services. nass.usda.gov

Francia, E., Tacconi, G., Crosatti, C., Barabaschi, D., Bulgarelli, D, Dall'Aglio, E., and Valè, G. 2005. Marker assisted selection in crop plants. Plant Cell, Tissue, and Organ Culture, 82: 317-342. DOI: 10.1007/s11240-005-2387-z

Geater, C., Fehr, W., and Wilson, L. 2000. Association of soybean seed traits with physical properties of natto. Crop Science, 40:1529-1534. DOI: 10.2135/cropsci2000.4061529x

Hegstad, J.M., Nelson, R.L., Renny-Byfield, S., Feng, L., and Chaky, J.M. 2019. Introgression of novel genetic diversity to improve soybean yield. Theoretical and Applied Genetics, 132: 2541-2552. DOI: 10.1007/s00122-019-03369-2

Hirata, K., Masuda, R., Tsubokura, Y., Yasui, T., Yamada, T., Takahashi, K., Nagaya, T., Sayama, T., Ishimoto, M., and Hajika, M. 2014. Identification of quantitative trait loci associated with boiled seed hardness in soybean. Breeding Science, 64: 362-370. DOI: 10.1270/jsbbs.64.362.

Hischhorn, J.N. and Daly, M.J. 2005. Genome-Wide Association Studies for Common Diseases and Complex Traits. Nature Reviews: Genetics, 6:95-108. DOI: 10.1038/nrgl521

Hosoi, T. and Kiuchi, K. 2003. Natto – A Food Made by Fermenting Cooked Soybeans with *Bacillus subtilis (natto)*. *Handbook of Fermented Functional Foods, Chapter 9*. Ed. Farnworth, E.R. ISBN: 978-0-20300-972-7

Hymowitz, T. 1984. Dorsett-morse soybean collection trip to East Asia: 50 year retrospective. Economic Botany, 38:378-388. October 1984. DOI: 10.1007/BF02859075

Hyten, D., Song, Q., Zhu, Y., Choi, I.Y., Nelson, R.L., Costa, J.M., Specht, J.E., Shoemaker, R.C., and Cregan, P.B. 2006. Impacts of genetic bottlenecks on soybean genome diversity. PNAS, 103(45):16666-16671. DOI: 10.1073/pnas.0604379103

Kada, S., Yabusaki, M., Kaha, T., Ashida, H., Yoshida, K. 2008. Identification of Two Major Ammonia-Releasing Reactions Involved in Secondary Natto Fermentation. Bioscience, Biotechnology, Biochemistry, 72 (7), 1869-1876. DOI: 10.1271/bbb.80129

Kim, J.M., Kim, K.H., Jung, J., Kang, B.K., Lee, J., Ha, B.K. 2020. Validation of marker-assisted selection in soybean breeding program for pod shattering resistance. Euphytica 216 (166). https://doi.org/10.1007/s10681-020-02703-w

Kumar, V. 2010. Nutritional Value of Soybean. *The Soybean: Botany, Production, and Uses, Chapter 17*. Ed. Singh, G. CABI, Cambridge, MA. ISBN: 978-1-84593-644-0

Long, A.D. and Langley, C.H. 1999. The Power of Association Studies to Detect the Contribution of Candidate Gene Loci to Variation in Complex Traits. Genome Research, 9:720-731. DOI: 10.1101/gr.9.8.720

National Research Council. 1993. Genetic Vulnerability and Crop Diversity. *Managing Global Genetic Resources: Agricultural Crop Issues and Policies.* Washington, DC: The National Academies Press. doi: 10.17226/2116.

National Academies of Sciences, Engineering, and Medicine. 1993. Managing Global Genetic Resources: Agricultural Crop Issues and Policies. Washington, DC: The National Academies Press. https://doi.org/10.17226/2116.

North Dakota Soybean Council. 2018. Association Applauds the North Dakota Soybean Council. The North Dakota Soybean Grower Magazine, p. 13.

Oil Crops Sector at a Glance. 2022. USDA Economic Research Service. ers.usda.gov

Orazaly, M., Chen, P., Zeng, A., and Zhang, B. 2015. Identification and Confirmation of Quantitative Trait Loci Associated with Seed Hardness. Crop Science. 55(2):688-694. DOI: 10.2135/cropsci2014.03.0219

Palmer R.G., Pfeiffer T.W., Buss G.R., Kilen T.C. 2004. Qualitative Genetics. *Soybeans: Improvement, Production, and Uses, Chapter 4.* 3rd Edition. Boerma, H.R. and Specht, J.E. (Eds.). ASA, CSSA, and SSSA, Madison, WI.

Panthee, D.R. Varietal Improvement in Soybean. *The Soybean: Botany, Production, and Uses, Chapter 5.* Ed. Singh, G. CABI, Cambridge, MA. ISBN: 978-1-84593-644-0

Qiu, L. and Chang, R.Z. 2010. The origin and history of soybean. *The Soybean: Botany, Production, and Uses, Chapter 1*. Ed. Singh, G. CABI, Cambridge, MA. ISBN: 978-1-84593-644-0.

Qutob, D., Ma, F., Peterson, C.A., Bernards, M.A., and Gijzen, M. 2008. Structural and permeability properties of the soybean seed coat. Botany. 86(3): 219-227. DOI: 10.1139/B08-002

Raghuvanshi, R.S., Bishit, K. 2010. Uses of Soybean: Products and Preparation. *The Soybean: Botany, Production, and Uses, Chapter 18*. Ed. Singh, G. CABI, Cambridge, MA. ISBN: 978-1-84593-644-0.

Reyna, N. and Sneller, C.H. 2001. Evaluation of Marker-Assisted Introgression of Yield QTL Alleles into Adapted Soybean. Crop Science, 41:1317-1321. DOI: 10.2135/cropsci2001.4141317x

Rizzo, G. and Baroni, L. 2018. Soy, Soy Foods, and Their Role in Vegetarian Diets. *Nutrients.* 10(1):43. January 2018. DOI: 10.3390/nu10010043

Ru, S. and Bernando, R. 2019. Predicted genetic gains from introgressing chromosome segments from exotic germplasm into an elite soybean cultivar. Theoretical and Applied Genetics. 133: 605-614. DOI: 0.1007/s00122-019-03490-2

Schmitz, J.F., Erhan, S.Z., Sharma, B.K., Johnson, L.A., and Myers, D.J. 2008. Biobased Products from Soybeans. *Soybeans: Chemistry, Production, Processing, and Utilization, Chapter 17*. Eds. Johnson, L.A., White, P.J., and Galloway, R. ACOS Press, Urbana, IL. ISBN: 978-1-893997-64-6.

Sedivy, E.J., Wu, F., and Hanzawa, Y. 2017. Soybean domestication: the origin, genetic architecture, and molecular bases. New Phytologist, 214: 539-553. DOI: 10.1111/nph.14418

Sedna, M., Kurauchi, T., Kasai, A., and Ohnishi, S. 2012. Suppressive mechanism of seed coat pigmentation in yellow soybean. Breeding Science, 61: 523-530. DOI: 10.1270/jsbbs.61.523.

Shook, J.M., Zhang, J., Jones, S.E., Singh, A., Diers, B.W., and Singh, A.K. 2022. Meta-GWAS for quantitative trait loci identification in soybean. G3, 11(7). DOI: 10.1093/g3journal/jkab117

Shurtleff, W. and Aoyagi, A. (2012). History of Natto and Its Relatives. *Soyinfo Center*, Lafayette, CA. ISBN: 978-1-928914-42-6.

Song, J., Liu, Z., Hong, H., Ma, Y., Tian, L., Li, X., Li, Y.H., Guan, R., Guo, Y., Giu, L.J. 2016. Identification and Validation of Loci Governing Seed Coat Color by Combining Association Mapping and Bulk Segregation Analysis in Soybean. PLoS ONE, 11(7). DOI: 10.1371/journal.pone.0159064

Soybean 2021 Export Highlights. 2021 Agricultural Export Yearbook. USDA Foreign Agricultural Services. 14 April 2022.

Taira, H. 1990. Effect of cultivar, seed size, and crop year on total and free sugar contents of domestic soybeans. Nippon Shokuhin Kogyo Gakkaishi, 37:203-213.

Todd, J.J. and Vodkin, L.O. 1993. Pigmented Soybean (*Glycine max*) Seed Coats Accumulate Proanthocyanidins during Development. Plant Physiology, 102(2): 663-670. DOI: 10.1104/pp.102.2.663

Voora, V., Larrea, C., and Bermudez, S. 2020. Global Market Report: Soybeans. International Institute for Sustainable Development. October 2020.

Yamaguchi et al. (2021) high yield and lodging tolerance

Yang, K., Jeong, N., Moon, J.K., Lee, Y.H., Lee, S.H., Kim, H.M., Hwang, C.H., Back, K., Palmer, R.G., and Jeong, S.C. 2010. Genetic Analysis of Genes Controlling Natural Variation of Seed Coat and Flower Colors in Soybean. Journal of Heredity. 101(6):757-768. DOI: 10.1093/jhered/esq078

Yao, J., Xu, H., Shi, N., Cao, X., Feng, X., Li, S., and Ouyang, P. 2009. Analysis of Carbon Metabolism and Improvement in γ-Polyglutamic Acid Production from *Bacillus subtilis* NX-2. Applied Biochemistry and Biotechnology. 160:2332–2341. 2010. DOI: 10.1007/s12010-009-8798-2

Yoshikawa, Y., Chen, P., Zhang, B., Scaboo, A., and Orazaly, M. 2013. Evaluation of seed chemical quality traits and sensory properties of natto soybean. Food Chemistry, 153:186-192. 2014. DOI: 10.1016/j.foodchem.2013.12.027

Zhang, B., Chen, P., Chen, C.Y., Wang, D., Shi, A., Hou, A., and Ishibashi, T. 2008a. Quantitative Trait Loci Mapping of Seed Hardness in Soybean. Crop Science, 48: 1341-149. DOI: 10.2135/cropsci2007.10.0544

Zhang, B., Chen, P., Shi, A., Hou, A., Ishibashi, T., and Wang, D. 2008b. Putative Quantitative Trait Loci Associated with Calcium Content in Soybean Seed. Journal of Heredity. 100(2): 263-269. DOI: 10.1093/jhered/esn096

Zhang, B., Chen, P., Florez-Palacios, S.L., Shi, A., Hou, A., and Ishibashi, T. 2010. Seed quality attributes of food-grade soybeans from the U.S. and Asia. Euphytica, 173: 387-396. DOI: 10.1007/s10681-010-0126-y

Zhou, S., Sekizaki, H., Yang, Z., Satoko, S., and Pan, J. 2010. Phenolics in the Seed Coat of Wild Soybean (*Glycine soja*) and Their Significance for Seed Hardness and Seed Germination. Journal of Agricultural and Food Chemistry, 58: 10972-10978. DOI: 10.1021/jf102694k

CHAPTER II

An Investigation of Seed Hardness and Seed Coat Lightness-Chroma-Hue Values in Natto Soybeans

ABSTRACT

Natto is a specialty fermented soyfood made from small-seeded (<10g 100 seeds^{-1}) soybean varieties. Seed hardness and seed coat color are important seed traits that determine the texture and appearance of natto and are thus valuable to breeders. Prior research has identified quantitative trait loci (QTL) for seed hardness, but its nature as a quantitative trait heavily influenced by the environment means that it is still poorly understood. Seed coat color has previously been studied as a qualitative trait, wherein simple visual inspection is used to sort seeds into simple color categories. Few studies have investigated seed coat color using analytical methods, such as the color spaces method developed by the Commission Internationale de l'Eclairage.

A genome-wide association study (GWAS) panel was assembled using 168 natto accessions from the USDA soybean germplasm, 51 natto breeding lines and 49 conventional breeding lines from the University of Arkansas, and 49 natto breeding lines from Virginia Tech. Field tests were conducted in 2021 using an augmented block design consisting of four blocks with one replication per block. High-throughput sequencing was conducted by the Soybean Genomics and Improvement laboratory to identify 32,724 single-nucleotide polymorphisms (SNPs) from the young leaf tissue extracted from 284 genotypes from the association panel. A kinship matrix and principal component analysis (PCA) were calculated to account for population structure within the GWAS panel, and GAPIT was used to conduct a genetic diversity analysis. The GWAS were conducted using the general linear model (GLM), mixed linear model (MLM), and single-marker regression (SMR), with a significance threshold of LOD > 3.0.

ANOVA conducted in JMP identified a significant environmental effect on all traits of interest, Genetic diversity analysis identified three distinct sub-populations within the

association panel. GWAS identified 11 SNPs for hardness, 15 for lightness, six for chroma, and nine for hue. One of the seed hardness SNPs, ss715579472, was colocalized to Chr 1 within 600 kbp of *Ha2*, a seed hardness QTL identified and confirmed by prior research. These results will be useful for developing new natto cultivars through marker-assisted selection. Additionally, phenotypic analysis identified six genotypes in the 95th percentile for soft seeds, nine for the 95th percentile of light seeds and low chroma, and 12 genotypes in the 95th percentile for optimal hue. Two of these genotypes, PI 458281 B and PI 603713, had optimal phenotypes for multiple traits, and were identified as potential breeding parents for developing new natto breeding lines.

INTRODUCTION

Natto is a fermented food produced by fermenting whole, steamed soybeans (*Glycine max* (L.) Merr) with the bacterium *Bacillus subtilis (natto)*. Natto is primarily produced and consumed in Japan, as well as in other countries in Southeast Asia. However, the United States exports as much as 75% of soybeans used in natto production, representing a valuable niche market for growers and breeders alike (North Dakota Soybean Council, 2018).

Natto soybeans are distinct and characterized by their small seed size, weighing less than 10 g 100^{-1} seeds or 80 mg $seed^{-1}$ (Geater, Wilson, and Fehr, 2000). Soybeans with low sucrose, high stachyose, and low protein and oil are preferred because they produce a better-tasting product (Taira, 1990). To be considered high quality, natto must possess several sensory characteristics. It must have a soft texture, a uniform, pale yellow appearance, and be covered in a white mucus that binds the natto together into a sticky mass (Wei, Wolf-Hall, and Chang, 2001). Natto's sensory characteristics are affected by various seed compositional traits, including protein and oil content, seed coat characteristics, free sugar content, and amino acid composition (Geater, Wilson, and Fehr, 2000).

Seed hardness has a significant effect on the texture of natto. Seed hardness is affected by several seed compositional traits, including seed coat permeability, pigmentation, and seed composition traits for protein, oil, and minerals. Disease resistance and stress tolerance traits have also been found to play a role in seed hardness, though the exact relationship is not well elucidated (Qutob et al., 2008; Geater, Wilson, and Fehr, 2000; Orazalay et al., 2015; Yoshikawa et al., 2014).

Natto appearance is determined by seed coat color. Natto manufacturers prefer soybeans with a light-yellow seed coat and a clear or yellow hilum, giving the seeds a uniform appearance

(Wei, 2001). Early genetic research identified five genetic loci (*I, R, T, W_1*, and *O*) that control seed coat color (Palmer et al., 2004). The *I, T,* and *R* loci have the most significant effect on seed pigmentation. The *I* locus controls the distribution of pigments in the seed coat by way of RNA silencing of chalcone synthesis genes. In contrast, the T and R loci interact to determine the specific pigment combination present. These five loci were later discovered to be involved in anthocyanin biosynthesis pathways, which affect flower color as well as seed coat color (Yang et al., 2010). A later study by Song et al. (2016) identified a novel controlling locus and confirmed three previously reported loci for seed coat color, with each locus controlling at least one trait pair of seed coat colors.

Using a more analytical method of classifying seed coat color may allow for a deeper understanding of the genetic architecture for the soybean seed coat. Prior studies have assessed seed color using a simple visual inspection; however, there may be other factors underlying seed coat color that a simple visual inspection cannot account for. In other legumes, prior studies have used Commission Internationale de l'Eclairage (CIE) color spaces describe seed and pod coloration in the common bean and snap bean, respectively (Hossian et al, 2011; Myers et al, 2019). CIE color spaces break down color into several components: Lightness (L*), which measures the white-black ratio of a color from 1-100; a^* and b^*, which are coordinates plotted on a grid with a red-green and yellow-blue axis, respectively; Chroma (C^*), which is the intensity of a color defined as $\sqrt{a^2 + b^2}$; and Hue (h^*), which the arc and angle of a^* and b^* expressed as degrees on a circle. Studies in other legumes suggest that CIE color spaces can not only give deeper insights into the genetics of seed coat color, but that it can be used to infer information about seed topography and similar traits (Hossian et al, 2011).

A major issue facing soybean breeders is the narrow genetic diversity of cultivated lines. While intensive selection and pedigree-based breeding have led to the development of high-yielding elite cultivars, it results in genetic uniformity that can leave crop populations susceptible to disease, pests, and environmental hazards. In addition, 47,000 unique accessions are available for soybeans, yet only 1,000 of these have been utilized in breeding programs. Integrating the exotic accessions that have thus far been ignored is a potential method for re-introducing genetic diversity into modern soy cultivars (Carter et al., 2004).

Marker-assisted selection (MAS) is a technique that utilizes molecular markers to select parents that can be used for creating new elite cultivars. Single-nucleotide polymorphisms (SNPs) are one of the most abundant markers available in an organism's genome. In soybean, the development of high-density SNP sets has made it much easier to develop high-resolution genetic maps and more precisely identify QTL and candidate genes (Song et al., 2013).

Genome-wide association study (GWAS) is a technique that statistically associates SNPs with phenotypic traits to identify significant markers and potential QTL that govern a trait of interest. Genome-wide association studies have been used in soybean research as a tool to facilitate MAS and GS. Genome-wide association studies have identified candidate genes for a wide variety of agronomic, seed compositional, and disease- and stress-related traits (Shook et al., 2022).

While prior research has investigated the genetic components of seed hardness and seed coat color, these studies were conducted primarily on recombinant inbred lines derived from elite cultivars, and those which used a selection of exotic accessions used only a small number of them (Hirata et al., 2014; Orazaly et al., 2015; Song et al., 2016). Few studies have investigated

the genetic components of seed hardness and seed coat color with a large population of exotic material.

The purpose of this study was to assess a diverse variety of soybean genotypes for superior seed hardness and seed coat color components by assessing and evaluating the performance of plant introductions (PI) against the performance of recombinant inbred lines (RILs); identifying potential breeding parents for hardness, lightness, chroma, and hue; analyzing the genetic diversity of both the RIL and PI populations; perform GWAS to identify SNPs that may be used for MAS or GS of seed hardness and seed coat color; and to compare the effectiveness of different GWAS models for identifying SNPs for these traits.

MATERIALS AND METHODS

Plant Materials and Experimental Design

An association mapping population of 317 genotypes was assembled from 168 plant introduction accessions from the USDA GRIN database (ars-grin.gov), 100 recombinant inbred lines from the University of Arkansas soybean germplasm, and 49 recombinant inbred lines provided by the Virginia Tech soybean germplasm.

All plant materials were grown in an augmented complete randomized block design (Federer and Raghavarao, 1975). There were four blocks, with one replication per block. Seeds were planted in irrigated, 3.05 m single-row plots with 1.5 m alleys. All materials were planted in late spring 2021 and harvested in late fall 2021.

There were four environments present in this study, one for each block. The first environment was the Milo J. Shult Agricultural Research and Extension Center, located in Fayetteville, AR. The field used for this block had a silt loam soil texture, consisting primarily of Pickwick soil series (aaes.uada.edu). The second environment was the Vegetable Research

Station in Alma, AR. The primary soil texture is fine silt loam, and the primary soil series are the Roxanna and Dardanelle series (aaes.uada.edu). The third environment is the Pine Tree Research Station located in Colt, AR. The soils at this location are silt loam, are generally alkaline, and the primary soil series are Calloway, Calhoun, Henry, and Loring (aaes.uada.edu). The fourth environment for this study was the Rice Research and Extension Center, located in Stuttgart, AR. The soils at this station are silt loam and are almost entirely from the DeWitt soil series (aaes.uada.edu).

Phenotypic Analysis

Seed preparation was conducted in accordance with protocols used in several other natto studies (Wei, Wolf-Hall, and Chang, 2001; Wei and Chang, 2004; Zhang, et al., 2008). Ten g of unbroken, uniform seeds were weighed and cleaned. The seeds were allowed to soak overnight in 150 mL of deionized water. Seeds were drained after soaking and stone seeds were removed. The seeds were then steamed in the autoclave for 30 minutes at 121°C and allowed to cool before hardness and color analysis.

Seed hardness was measured as maximum Newtons of force at 5mm of shearing. Seed hardness was assessed using a TX.XT Plus Connect Texture Analyzer (Texture Technology Corporation, Sterling, VA) with a single blade probe according to existing laboratory protocols (UARK Soybean Breeding Laboratory). The probe was set at a resting position 5mm above the testing surface and sheared through the seeds at a speed of 1 mm/s, touching the testing surface before returning to start position. This procedure was repeated three times for each genotype and location combination, and the texture analyzer's software package averaged the recorded data.

Seed coat color was assessed by measuring the tristimulus color values of each test. Tristimulus color values are a system of measurement that detects the components of color like

the human eye. Tristimulus color is generally divided into three components: lightness, chroma, and hue (CIE, 2011). Lightness, or luminance, is a unitless measurement of an object's perceived brightness on a scale of 0 (completely black) to 100 (completely white). Chroma is a unitless measurement of color intensity relative to an object's lightness with no upper limit. Hue is a measurement of how closely an object's color resembles any of four defined colors mapped to the quadrants of a grid where the x and y axes are represented as *a* and *b*: red (+*a*), yellow (+*b*), green (-*a*), and blue (-*b*). In the L*C*h* model used for this study, hue is defined as the tangent in degrees of an object's *a* and *b* measurements, where red is valued at 0 and increasing counter-clockwise (**Fig. 1**). Seeds from each test were placed in an off-white receptacle and measured once using a CR-10 Tristimulus Colorimeter (Konica Minolta, Ramsey, NJ).

Genotypic Analysis

Young leaf tissue samples were collected from seedlings grown in the Rosen Center greenhouses in Fayetteville, AR. Each sample consisted of a dime-sized composite taken from eight individual plants in the V2 stage. Leaf tissue was then lyophilized and subjected to a modified CTAB DNA extraction as described by Doyle and Doyle (1990) and assessed by NanoDrop for purity and concentration. Extracted DNA was sent to Dr. Qijian Song at the Soybean Genomics and Improvement Laboratory (USDA, Beltsville, MD), where it was subject to genotype-by-sequencing using the SoySNP50k iSelect BeadChip platform (Illumina, Inc., San Diego, CA) (Song et al., 2013). Markers with <5% minor allele frequency were filtered out using TASSEL 5.2.86, bringing the total SNPs used for the GWAS to 32,724. Missing data values were imputed in TASSEL using Linkage Disequilibrium-*k* Nearest Neighbor imputation (LD-kNNi) as described by Money et al. (2015).

Data Analysis

Phenotypic data were analyzed in GAPIT and JMP Pro 16. GAPIT 3 (Lipka et al., 2012) was used to analyze the distribution of phenotypic data across all locations. In JMP (SAS Institute, Cary, NC), one-way ANOVA was calculated for all four traits to examine the overall effect of genotype and environment on trait means. Analyses of variance were used to compare the data from all four locations as well as the data for each pair of locations to determine the environmental and genotypic effects. Variations in trait means between genotypes were analyzed using the student's t-test. These data were then used to select the 15 best performing genotypes for each trait. For seed hardness and chroma, the genotypes that had the lowest values for the traits of interest were considered superior, because softer and duller seeds are preferred for natto production (Yoshikawa, et al., 2014). For lightness and hue, genotypes that had the highest values for the traits of interest were considered superior, because lighter, yellower beans are preferred for natto production (Yoshikawa, et al., 2014). GAPIT 3 was also used to conduct principal component analysis (PCA) and genetic diversity analysis. Principle component analysis (PCA) was conducted by setting PCA = 2 to 10 and NJ tree = 2 to 10. Phylogenic trees were drawn using neighbor-joining (NJ) method.

Association Analysis

TASSEL 5 was used to conduct GWAS using 32,724 SNPs (Bradbury et al., 2007). Population structure was accounted for using the PCA data from GAPIT as well as the kinship (K) matrix, which was calculated in TASSEL 5. Genome-wide association studies were conducted using 3 models: general linear model (GLM), mixed linear model (MLM), and single-marker regression (SMR). Four GWAS were conducted for each trait using the data from each individual location. For all GWAS, the significance threshold was determined to be LOD > 3.0.

RESULTS

Phenotypic Analysis

The mean seed hardness across 317 genotypes was 29.7 Newtons (N) and showed a slightly negative distribution, with more values on the lower extremes than in the higher ones (**Fig 2a**). For the individual genotypes, mean hardness ranged from 18.89 N in PI424529 to 42.46 N in PI594683A, for a 23.57 N difference. The standard deviation was 3.82 N, the variance was 14.58 N, and the CV was 12.86 N, meaning that 66.7% of genotypes will have a mean hardness between 25.862 N and 33.501 N, and 95% of hardness values will fall between 22.047 N and 37.32 N. One-way ANOVA of hardness showed no significant genotypic effect across all four locations ($p = 0.9499$), while there was a significant environmental effect ($p < 0.001$). Conducting ANOVA on each pair of growing locations yielded similar results, where genotype consistently demonstrated no significant effect on seed hardness and growing environment had a significant effect. A student's t test revealed significant differences between genotypes, particularly between the genotypes with the highest and lowest hardness values. The 95^{th} percentile of genotypes for hardness can be found in **Table 1**.

The mean lightness across 317 genotypes was 50.48 and demonstrated a normal distribution (**Fig 2b**). The mean lightness ranged from 38.07 in R17-839 to 59.98 in V18-1624, for a difference of 21.91. The standard deviation was 3.56, the CV was 7.05, and the variance was 12.69, showing that 66.7% of genotypes had a mean lightness between 46.93 and 54.05, and 95% of genotypes had a mean lightness between 43.37 and 57.61. One-way ANOVA showed no significant genotype effect on lightness values across locations ($p < 0.5988$), while the environment had a significant effect ($p < 0.0001$). Conducting ANOVA on each pair of growing locations yielded similar results, where genotype consistently demonstrated no significant effect

on seed lightness and growing environment had a significant effect. A student's t-test revealed significant differences between genotypes, particularly between the genotypes with the highest and lowest lightness values. The 95th percentile of genotypes for lightness is shown in **Table 2**.

The mean chroma for all genotypes was 31.08 and had a normal distribution (**Fig 2c**). Chroma ranged from 21.57 in R17-839 to 37.29 in R18C-13379, a difference of 15.72. The SD was 2.29, the CV was 7.36, and the variance was 5.23, showing that 66.7% of genotypes had values between 28.79 and 33.37, while 95% of genotypes had a chroma between 26.5 and 35.66. One-way ANOVA revealed that there was no significant genotype effect on chroma values across locations ($p < 0.5482$), while the environment had a significant effect on chroma ($p < 0.0001$). ANOVA conducted on each pair of growing locations found that, in all but one case, genotype consistently demonstrated no significant effect on chroma, and that growing environment had a significant effect. The exception to this is the ANOVA comparing Kibler and Pinetree, which found that genotype had a small but significant ($p < 0.16$) effect on chroma, and environment showed a highly significant ($p <0.0001$). A student's t-test revealed significant differences between genotypes, particularly between the genotypes with the highest and lowest chroma values. **Table 3** shows the 95th percentile of genotypes for chroma.

The mean hue from all genotypes was 82.15° and demonstrated a slightly negative distribution (**Fig 2d**). The mean hue ranged from 75.40° in R14-7075 to 90.17° in R15-1587, a difference of 14.77°. Standard deviation was 2.8°, CV was 3.41°, and variance was 7.86°, meaning that 66.7% of genotypes possessed a hue between 79.35° and 84.95°, while 95% of genotypes were between 76.55° and 87.75°. One-way ANOVA showed no significant genotype effect on hue values across locations ($p < 0.4429$), while the environment had a significant effect on hue ($p < 0.0001$). ANOVA conducted on each pair of growing locations found that, in all but

one case, genotype consistently demonstrated no significant effect on hue, and growing environment had a significant effect. The exception to this is the ANOVA comparing Kibler and Fayetteville, which found that Genotype had a small but significant ($p < 0.0303$) effect on hue, and environment had no significant effect ($p < 0.2234$). Compared to the other ANOVA results, the overall variation between these locations was much lower, but still significant ($p < 0.0309$). A student's t-test revealed significant differences between genotypes, particularly between the genotypes with the highest and lowest hue values. The 95th percentile of genotypes for hue can be found on **Table 4**.

A correlation analysis showed that there was no significant correlation between the seed hardness and seed coat color traits examined in this study. There is a significant positive correlation between seed lightness and chroma ($r = 0.74$), but no significant correlation between hue and lightness or chroma.

Population Analysis and Genetic Diversity

Based on PCA and phylogenetic analysis, the GWAS panel of 284 genotypes was divided into three sub-populations, labeled A, B, and C (**Fig. 4**). All genotypes were arranged into a phylogenetic tree using the neighbor-joining (NJ) method, which was drawn in GAPIT 3. Sub-population A (**Fig 4a**) contained all the sequenced genotypes obtained from the University of Arkansas and Virginia Tech breeding programs and 18 of the accessions from the USDA, for a total of 137 genotypes. Of the accessions in sub-population A, 10 were cultivars developed in breeding programs. The remaining USDA accessions were split between sub-populations B and C. Sub-population B, containing 100 genotypes, is divided into two major groups – Group B1 and B2 (**Fig 4b**). Of the 54 genotypes in Group B1, 40 originated from South Korea, 10 from Japan, two from the United States, and two from mainland China. Of the 46 genotypes in Group

B2, 30 originated from mainland China, 9 from Vietnam, two from Taiwan, two from Japan, one from the United States, and one from Turkey. Sub-population C contained 47 genotypes, all of which were from South Korea.

Genome-Wide Association Study

Genome-wide association studies were conducted on the test of each growing location for seed hardness, lightness, chroma, and hue. Each GWAS was conducted in TASSEL 5 using the SMR, GLM, and MLM models. To visualize the results, quantile-quantile (QQ, hereafter) and Manhattan plots were generated in TASSEL.

Seed Hardness: 11 SNPs were significantly (LOD > 3) associated with seed hardness, located on chromosomes (Chr hereafter) 1, 3, 9, 10, 11, 17, and 18. Chromosome 9 contains three SNPs, Chr 1 and 18 contains two SNPs, and the remainder each contain one SNP (**Table 6**). Of these, three SNPs were identified in the Stuttgart test, two were identified in Pinetree, two were identified in Kibler, and four were identified in Fayetteville. The GLM (**Table 6**; **Figs. 4a** and **5a**) identified all 11 SNPs, while MLM (**Table 6**; **Figs. 4b** and **5b**) excluded the SNPs identified on Chr 9, leaving eight. The SMR model (**Table 6**; **Figs. 4c** and **5c.**) showed only three of these SNPs to be significant. None of the GWAS models identified any SNPs that were significant across multiple locations.

Lightness: Genome-wide association study identified 15 significant SNPs associated with seed lightness (**Table 7**), located on Chr 2, 3, 4, 5, 7, 13, 16, 17, 18, and 19. Chromosome 3 contains three SNPs, Chr 2, 7, and 17 contain two SNPs each, and the remainder contain one SNP each. Of these SNPs, three were identified from the Stuttgart test, two from Pinetree, six from Kibler, and three from Fayetteville. The GLM (**Table 7**; **Figs. 6a** and **7a**) found 13 SNPs to be significant and excluded ss715585247 (Chr 3) and ss715597558 (Chr 7). The MLM (**Table 7**;

Figs. 6b and **7b**) found 10 SNPs to be significant. The MLM excluded the same SNPs as GLM, as well as ss715590367 (Chr 5), ss715598746 (Chr 7), and ss715627430 (Chr 17). The SMR model identified 11 significant SNPs associated with seed lightness (**Table 7**; **Figs. 6c** and **7c**). Contrary to GLM and MLM, SMR did find ss715585247 (Chr 3) and ss715597558 (Chr 7) to be significant, while excluding ss715590367 (Chr 5), ss715598746 (Chr 7), ss715627430 and ss715630513 (Chr 17). None of the GWAS models identified any significant SNPs across multiple locations.

Chroma: Genome-wide association study identified a total of six significant SNPs for seed chroma, located on Chr 1, 4, 5, 7, 10, and 11 (**Table 8**). Of these, four SNPs were identified in the Stuttgart test and two in the Fayetteville test. The GLM (**Table 8**; **Figs. 8a** and **9a**) identified all six of these SNPs, while MLM (**Table 8**; **Figs. 8b** and **9b**) excluded ss715610854 (Chr 11). The SMR model (**Table 8**; **Figs. 8c** and **9c**) found only four of these SNPs to be significant and excluded ss715607569 (Chr 10) and ss715610854 (Chr 11). No significant SNPs were identified for seed chroma in the Kibler and Pinetree tests, and no SNPs were found to be significant for more than one test.

Hue: The GLM (**Table 9**; **Figs. 10a** and **11a**) found nine SNPs significantly associated with seed hue, located on Chr 1, 7, 17 (one SNP), 8 (two SNPs), 18, and 19 (three SNPs). Of these, eight SNPs were identified in the Stuttgart test, two were identified in the Kibler test, and the remaining SNP was identified in the Fayetteville test. The MLM (**Table 9**; **Figs. 10b** and **11b**) identified 10 significant SNPs, only excluding ss715596450 (Chr 7). The SMR model (**Table 9**; **Figs. 10c** and **11c**) identified six significant SNPs, excluding ss715627726 (Chr 17), ss715630585 (Chr 18) and ss715636003 (Chr 19). None of the models found any SNPs to be significant in Pinetree, nor across multiple locations.

DISCUSSION

Natto manufacturers prefer softer seeds; incorporating exotic and rarely used genotypes into existing breeding populations can be an effective tool for producing softer, better seeds for natto production (Zhang et al., 2008a). The distribution of seed hardness (**Fig 2a**) showed that the hardness of the seeds in this study skewed to the softer side of the mean. This implies that several genotypes from this collection may be suitable for developing new natto lines. The average seed hardness from this study ranged from 15 N g^{-1} to 35 N g^{-1}. These results show that the collection in this study is slightly harder in general than other collections used in similar studies. Zhang et al (2008a), for example, found that the average hardness of small-seeded soybean varieties has an approximate range of 12 N g^{-1} to 26 N g^{-1}.

The ANOVA test demonstrated that seed hardness is significantly affected by the environment. A primary factor affecting seed hardness is calcium content. Prior research has identified a strong positive correlation between hard-seededness and calcium content in the seed coat, particularly in small-seeded soybean varieties (Zhang et al., 2008b; Orazaly et al., 2018). Soil sample data from the Stuttgart environment showed soil calcium levels of 923 ppm, which is within the typical range seed at this station (J. McCoy, personal communication, 2023). Similarly, soil data from the Pine Tree Research Station showed soil calcium levels were tested at 1,576 ppm, which is also within the average range for that location (J. Hedge, personal communication, 2023). Examining the phenotypic data in JMP shows that there is no significant difference between the seed hardness values of genotypes grown in these locations. It is difficult to conclusively determine the full impact of soil calcium on the results of this study without additional data. The effect of soil calcium on seed hardness could be a promising avenue to examine in future studies.

A study by Argel and Paton (1999) reported that seed hardness is also affected by environmental moisture, particularly humidity and rainfall during the growing season. Weather data from the four growing locations shows that the 2021 growing season had less precipitation than normal (Southern Regional Climate Center, 2021). However, all of the field trials were irrigated throughout the growing season; it is therefore unlikely that water availability was a major factor in determining seed hardness for this study.

Natto manufacturers prefer seeds with light yellow seed coats and yellow hila. All genotypes used in this study were yellow seeded, with a range of average hue values between 75 and 90°. The benchmark for a "pure" yellow hue is 90°, while 75° is yellow-orange (**Fig. 1**). Historically, seed coat color is described in qualitative terms, but some studies have been conducted on various legume species to investigate and classify seed coat color in terms of CIE color components. One study by Hossian, et al. (2011) examined the genetic architecture of seed coat color components in chickpeas (*Cicer arietinum* L.) and found that analyzing the color components can allow researchers to infer other qualities of the seed coat, such as seed coat thickness and topography. This study also served to identify the possibility of subjective errors when performing a visual inspection to place seeds into non-quantitative color categories. A GWAS study by Myers et al (2019), examined the pathways involved in anthocyanin synthesis in the pods of snap beans (*Phaseolus vulgaris* L.), using L*a*b* coordinates to phenotype pod color. The GWAS successfully identified six significant quantitative trait nucleotides (QTN) for *L*a*b** coordinates. The data from both of these studies suggest that while L*C*h* can be a useful metric for analyzing seed coat color, L*a*b* may be more practically useful due to its relative simplicity.

Several lines were identified as potential breeding parents for multiple natto quality traits. PI 458281 B (KAS 580-7), a genotype from South Korea, collected from Jeollanam-do, was in the 95th percentile for soft seeds, low chroma, and optimal yellow hue. This makes PI 458281 B a good candidate for natto production and a parent for breeding new natto cultivars. PI 603713 (ZDD12392) is a genotype collected from an unknown location in China that outperformed 95% of all genotypes for low chroma and 90% of genotypes for soft seeds. R17-839, a conventional breeding line sourced from the University of Arkansas germplasm, possessed the lowest chroma of all genotypes included in this study, and the second-best hue. That being said, this line is not from a small-seeded variety, which is an undesirable characteristic for natto production.

It is worth noting that the origin of a given genotype had no significant effect on the examined natto characteristics. Attempting to map the best and worst genotypes onto the phylogenetic tree produced by the genetic similarity analysis showed no significant correlation between sub-population and trait values. Furthermore, there was no significant difference in natto trait values between the conventional and small-seeded soybean genotypes.

Comparing the performance of the models used in these GWAS, SMR, GLM, and MLM, shows mixed results. In terms of accurately identifying SNPs, all three models were comparable. All three produced similar R^2 values for the identified SNPs, ranging from approximately 0.04 to 0.09. The MLM and GLM models identified more SNPs for each trait than SMR. Additionally, most of the SNPs that were identified by SMR were also identified by MLM and GLM, which implies that SMR was less powerful than the other models. This is because SMR is a simpler GWAS model than GLM and MLM. Single marker regression is an additive model that treats each marker as a fixed effect without taking population structure into account and assumes that a marker will only affect a trait if it is in LD with an unknown QTL. Single marker regression also

assumes that the markers are evenly spaced throughout a given chromosome and that the MAF of every marker is above 0.2. In practice, however, the spacing and MAF of markers varies widely, which reduces the power of this model (Hayes, Gondro, and van der Werf, 2013). In contrast, GLM and MLM specifically account for the effect of population structure. The models are similar to one another, with one of the primary differences being that GLM clusters all genotypes in one group, while MLM clusters them into more groups depending on the level of compression (Zhang et al, 2010). This is likely why the GLM and MLM results are similar.

The GWAS identified 11 SNPs significantly associated with seed hardness in this study. These SNPs were identified on Chr 1, 3, 9, 10, 11, 17, 18 with Chr 9 containing the most significant SNPs. The QQ plots for seed hardness (**Fig. 4**) and genomic inflation factor λ (**Table 5**) show that the genomic inflation was between 0.91 and 0.15 for GLM, 0.91 and 1.13 for MLM, 0.85 and 1.18 for SMR, all within acceptable levels. Generally, QQ plots and λ are used to compare the distribution of the test statistic against the expected null distribution (Williams et al, 2021). For λ, values less than 1.2 are considered acceptable for GWAS (Doherty et al, 2018). The Manhattan plots for seed hardness (**Fig. 5**) suggest two significant SNPs on Chr 6 and 20 as well, but these were not consistently significant in individual tests. Some of the significant SNPs identified in this study are consistent with markers found in previous studies. Previous studies collectively identified 13 QTL for seed hardness using SSR as the genetic marker. Zhang et al (2008a) identified two QTL on Chr 1 and 19, while Orzalay et al (2015) identified 12 QTL on Chr 1, 7, 11, 16, 18, and 20, and Hirata et al (2014) identified 2 QTL on Chr 3 and 6. One of these QTL *Ha2* (Zhang et al, 2008a; Orazaly et al, 2015), was colocalized within 600 kbp of SNP ss715579472 on Chr 1. This suggests that ss715579472 is a viable marker for MAS. A GWAS study by Zhang et al (2018) found 57 SNPs significantly associated with seed hardness

on Chr 2, 5- 9, 11, 12, 14, 15, and 17-20. None of these were found to be within 1 Mbp of any SNPs identified in this study. Further research will be needed to confirm this study's SNPs and identify the QTL and candidate genes associated with them.

Genome-wide association studies conducted for seed coat color in this research identified 14 SNPs significantly associated with seed lightness, six with chroma, and 11 with hue. The QQ plots for lightness (**Fig. 6**) show a slight inflation of observed LOD values compared to the expected, particularly for the SMR model. The λ for lightness (**Table 5**) ranged from 0.91 to 0.97 for GLM, 0.9 to 1.08 in MLM, and 0.7 to 1.01 for SMR, indicating that genomic inflation is within acceptable limits. The QQ plots for chroma (**Fig. 9**) suggest that acceptable levels of genetic inflation, supported by the λ for chroma (**Table 5**), which had a range of 0.85 to 1.08 for GLM, 0.89 to 1.08 in MLM, and 0.85 to 1.0 in SMR. For Hue, the QQ plots (**Fig. 10**) indicate acceptable levels of genomic inflation. The λ for hue (**Table 5**) ranged from 0.87 to 1.11 in GLM, 0.94 to 1.12 in MLM, and 1.02 to 1.21 in SMR. As of this writing, there is little prior research to compare the lightness and chroma GWAS to, as these traits have not been thoroughly investigated in a breeding context. For hue, however, the body of research is larger, as it matches closely with the *qualitative* seed coat color trait that has been investigated in prior studies. The ANOVA results from this study suggest that hue was primarily affected by the environment, while genotype was insignificant. This is in direct conflict with the established body of literature. Classical geneticists determined that 3 primary loci governed seed coat color, with further studies identifying the *I* allele on Chr 8 to be responsible for the yellow seed coat via RNA suppression of pigment production (Bernard and Weiss, 1973; Palmer et al, 2004; Sedna et al., 2011). A more recent GWAS study by Song et al (2016) further confirmed the location of the I locus on Chr 8 and identified 3 other loci that govern seed coat color, located on Chr 1, 6, and 9. The individual

GWAS conducted for this study identified 3 SNPs that are supported by the literature on Chr 1 (1 SNP) and 8 (2 SNPs). None of these SNPs were significant in more than one location, however; the SNPs on Chr 1 were only identified in the Fayetteville test, while the SNPs on Chr 8 were only found in the Kibler test. Further study will be needed to confirm these SNPs and identify the QTL and candidate genes associated with them.

This study was conducted using an augmented block design. The augmented block design has historically been used to incorporate exotic breeding material, such as the PIs used in this GWAS panel, into existing populations while keeping heterogeneity under control (Federer and Raghavarao, 1975). In traditional plant breeding schemes, this design allows the breeder to estimate block effects with fewer replications, which is useful for assessing exotic materials in multiple environments without sacrificing too many resources on accessions that have a higher risk of poor performance. When applied to genetics studies such as GWAS, however, this study found that the design's applicability is more limited. The experiment was designed to collect data from four different blocks in a single growing year. Each block was grown in a different location in the state of Arkansas, and only a single replication was grown at each location. All of the growing locations had different soil series, different levels of soil moisture, and required different field management practices. Taken all together, these factors created an inflated environmental effect that made it difficult to calculate heritability and obfuscated the genotypic effects on the traits of interest. There was enough diversity within each location that GWAS could still be conducted on each individual location, however. For future studies examining these traits, a design with 2 different environments, 2 or more replications, and grown over a period of at least 2 years would be a more effective approach.

This study found that environment had a significant impact on seed hardness, lightness, chroma, and hue. Genome-wide association studies identified 11 SNPs significantly associated with seed hardness, 15 associated with seed lightness, six associated with seed chroma, and nine associated with seed hue. This study also identified several genotypes that could potentially be used as parents when breeding natto soybeans for these traits. Future studies are needed to confirm the significance of the identified SNPs and the parental fitness of the top-performing genotypes. Natto is a valuable soybean commodity with an annual market value of $1.39 billion as of January 2022 (Technavio Research, 2022). Through extensive phenotypic and molecular evaluation of traits desirable for natto production, this study provides valuable information for developing new natto cultivars via marker assisted selection.

TABLES AND FIGURES

Name	Mean Hardness	Origin
PI 424529	18.88948	South Korea
PI 567501	21.17049	China
PI 408169C	21.37614	South Korea
PI 458281B	21.5845	South Korea
PI 507088	21.755	Japan
PI 538024	21.93925	United States

Table 1. Genotypes in the 95th percentile for optimal seed hardness.

Name	Mean Lightness	Origin
V18-1624	59.98376	Virginia Tech
R15-4813	59.43372	University of Arkansas
PI 407805D	58.65	South Korea
PI 542053	58.60039	United States
R19-4124b	58.13372	University of Arkansas
R19-41580	58.125	University of Arkansas
PI 398976	57.80418	South Korea
R14-7075	57.76706	University of Arkansas
PI 594680	57.7	China

Table 2. Genotypes in the 95th percentile for optimal seed lightness.

Name	Mean Chroma	Origin
R17-839*	21.56936	University of Arkansas
PI 518756	22.5	Brazil
R18-14572*	25.1	University of Arkansas
MFS-561	25.725	University of Arkansas
PI 399118	25.95	South Korea
PI 603713	26.075	China
PI 408081	26.15	South Korea
PI 398479	26.275	South Korea
PI 458281 B	26.4	South Korea

Table 3. Genotypes in the 95th percentile for optimal seed chroma. The asterisk (*) indicates the genotype is from a non-natto variety.

Name	Mean Hue	Origin
R15-1587*	90.168967	University of Arkansas
R17-839*	89.802301	University of Arkansas
R17-3144	89.590935	University of Arkansas
PI 458281 B	89.5	South Korea
PI 574476 A	88.86435	China
R18C-1450*	88.702301	University of Arkansas
PI 606412	88.265991	Vietnam
PI 398374	88.225	South Korea
R14-7368	88.225	University of Arkansas
R17-3328	88.015991	University of Arkansas
PI 567734	87.909065	China
PI 398478	87.85	South Korea

Table 4. Genotypes in the 95th percentile for optimal seed hue. The asterisk (*) indicates the genotype is from a non-natto variety.

Trait	Location	Genomic Inflation Factor (λ)		
		GLM	MLM	SMR
Hardness	FAY	1.14184	1.13878	1.18478
	KIB	0.91826	0.91694	0.96442
	PTR	0.9398	0.9823	0.8592
	STU	1.0779	1.07478	1.06726
Lightness	FAY	0.9224	0.92136	0.87682
	KIB	0.9861	1.08306	0.70218
	PTR	0.9738	0.973	1.0153
	STU	0.91032	0.90896	1.00484
Chroma	FAY	1.00214	1.05384	0.88386
	KIB	1.01508	1.05194	1.00074
	PTR	1.08332	1.08204	0.85404
	STU	0.86318	0.8945	0.94266
Hue	FAY	1.11228	1.12916	1.20556
	KIB	1.0972	1.09516	1.1978
	PTR	1.09926	1.09708	1.14112
	STU	0.86958	0.94914	1.01908

Table 5. Genomic inflation values (λ) of the general linear model (GLM), mixed linear model (MLM), and single-marker regression (SMR) GWAS conducted for seed hardness, lightness, chroma, and hue. Genomic inflation factor is defined as the median observed p-value divided by the median expected p-value.

SNP	Chr	Position	Alleles	Location	MAF	GLM		MLM		SMR	
						LOD	R^2	LOD	R^2	LOD	R^2
ss715579472	1	4387206	T:C	STU	0.1045627	4.85	0.07998	4.50	0.07998	NS	N/A
ss715580529	1	55129174	T:C	PTR	0.391635	4.68	0.07821	4.31	0.07819	4.75	0.07955
ss715585240	3	3460824	G:T	KIB	0.21673	3.52	0.05952	3.34	0.05952	NS	N/A
ss715604885	9	48411573	A:G	FAY	0.148289	3.11	0.06752	NS	N/A	NS	N/A
ss715604903	9	48612289	C:T	FAY	0.148289	3.11	0.06263	NS	N/A	NS	N/A
ss715604904	9	48615436	G:A	FAY	0.148289	3.11	0.06752	NS	N/A	NS	N/A
ss715606322	10	3526450	A:G	STU	0.0684411	3.36	0.05601	3.19	0.05601	3.99	0.06694
ss715610431	11	3764301	C:T	STU	0.4087452	3.71	0.06168	3.50	0.06168	NS	N/A
ss715627912	17	4189610	G:A	KIB	0.1159696	3.49	0.05909	3.31	0.05909	NS	N/A
ss715632582	18	696206	T:G	FAY	0.4638783	3.25	0.07061	3.06	0.07061	3.37	0.0725
ss715631457	18	49191102	C:T	PTR	0.4410646	3.91	0.06581	3.14	0.05631	NS	N/A

Table 6. List of significant single-nucleotide polymorphisms (SNPs) associated with seed hardness.

SNP	Chr	Position	Alleles	Location	MAF	GLM		MLM		SMR	
						LOD	R^2	LOD	R^2	LOD	R^2
ss715581676	2	2784160	C:A	STU	0.2509506	4.23	0.07283	3.96	0.07283	3.22	0.05538
ss715580957	2	10903396	G:A	PTR	0.2661597	3.46	0.05884	3.29	0.05884	3.38	0.05725
ss715585247	3	344529	G:A	KIB	0.4562738	NS	N/A	NS	N/A	3.87	0.0639
ss715586476	3	44170599	G:A	STU	0.4771863	5.63	0.09589	5.16	0.09589	4.74	0.08056
ss715586479	3	44177812	A:G	STU	0.4771863	5.63	0.09589	5.16	0.09589	4.74	0.08056
ss715588940	4	51272076	G:A	FAY	0.3498099	3.41	0.06896	3.21	0.06896	3.29	0.06655
ss715590367	5	28417535	T:C	KIB	0.3193916	3.70	0.06014	NS	N/A	NS	N/A
ss715598746	7	8085704	A:G	KIB	0.3688213	3.22	0.05266	NS	N/A	NS	N/A
ss715597558	7	37177741	G:T	FAY	0.1273764	NS	N/A	NS	N/A	3.14	0.0636
ss715614830	13	29630754	G:A	PTR	0.2053232	3.20	0.05476	3.05	0.05476	3.60	0.06113
ss715625250	16	5887013	A:G	KIB	0.2091255	3.90	0.06335	3.12	0.0546	4.20	0.06921
ss715627430	17	38237259	T:C	KIB	0.2509506	3.65	0.05937	NS	N/A	NS	N/A
ss715627440	17	38274690	G:A	KIB	0.2661597	4.12	0.06674	3.13	0.05483	NS	N/A
ss715630513	18	4191148	A:G	FAY	0.0855513	3.27	0.06627	3.09	0.06627	3.61	0.0727
ss715635189	19	42347705	G:A	KIB	0.1749049	3.41	0.05568	3.34	0.05866	3.42	0.0566

Table 7. List of significant single-nucleotide polymorphisms (SNPs) associated with seed lightness.

SNP	Chr	Position	Alleles	Location	MAF	GLM		MLM		SMR	
						LOD	R^2	LOD	R^2	LOD	R^2
ss715579518	1	45706109	A:C	STU	0.1197719	4.27	0.07251	3.95	0.07236	4.78	0.08114
ss715588606	4	48398087	A:C	FAY	0.0836502	3.54	0.07034	3.29	0.07101	3.92	0.07877
ss715590472	5	30377957	G:A	STU	0.1673004	3.25	0.05563	3.17	0.05769	4.09	0.06991
ss715598886	7	8763811	T:G	FAY	0.0684411	4.27	0.07018	4.04	0.07207	4.66	0.0784
ss715607569	10	46434446	A:G	STU	0.4296578	3.77	0.06433	3.05	0.05532	NS	N/A
ss715610854	11	5303401	C:T	STU	0.3555133	3.79	0.06472	NS	N/A	NS	N/A

Table 8. List of significant single-nucleotide polymorphisms (SNPs) associated with seed chroma.

SNP	Chr	Position	Alleles	Location	MAF	GLM		MLM		SMR	
						LOD	R^2	LOD	R^2	LOD	R^2
ss715579862	1	49394831	A:C	FAY	0.2756654	4.39	0.08862	3.80	0.08307	4.13	0.08291
ss715596450	7	15729459	T:C	STU	0.1140684	3.85	0.06673	NS	N/A	3.75	0.06474
ss715600210	8	19343981	T:C	KIB	0.1444867	4.36	0.0724	4.08	0.0724	3.51	0.05812
ss715600410	8	19968349	A:C	KIB	0.1292776	3.80	0.06345	3.59	0.06345	3.09	0.05137
ss715627726	17	40272132	C:T	STU	0.1882129	3.59	0.06224	3.34	0.0614	NS	N/A
ss715630585	18	40431935	T:C	STU	0.0608365	3.43	0.04794	3.00	0.04288	NS	N/A
ss715636003	19	50217107	T:G	STU	0.1235741	3.48	0.06053	3.00	0.05507	NS	N/A
ss715636004	19	50228894	C:T	STU	0.2148289	4.42	0.07631	3.53	0.06512	3.83	0.06616
ss715636012	19	50293736	G:A	STU	0.1768061	4.64	0.07986	3.83	0.07085	4.28	0.07348

Table 5. List of significant SNPs associated with seed hue.

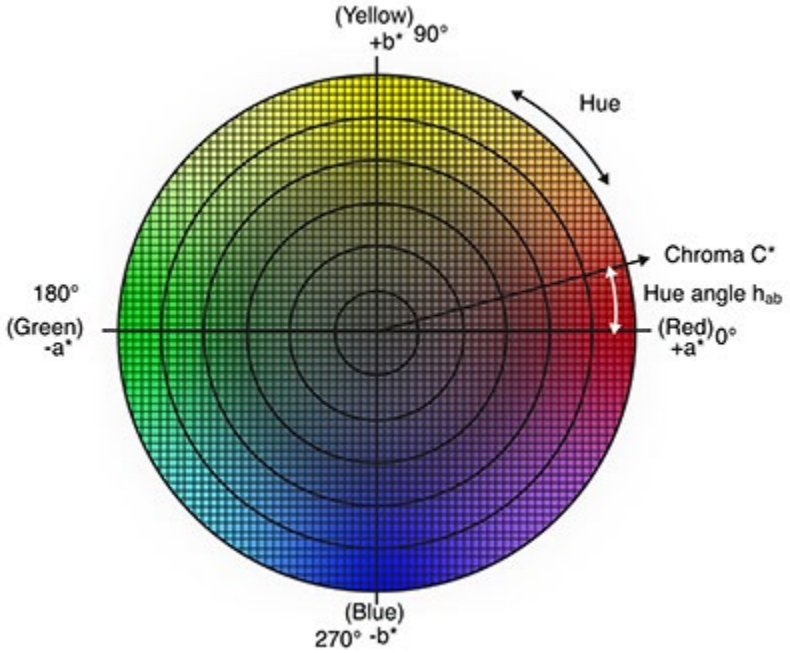

Figure 1. Color spaces model developed by the Commission Internationale de l'Eclairage (CIE). The *a* and *b* values represent the color of the sample. Chroma is defined as the tangent of *a* and *b*. Hue is defined as the angle formed from Chroma. 90° is considered the benchmark for perfect yellow hue.

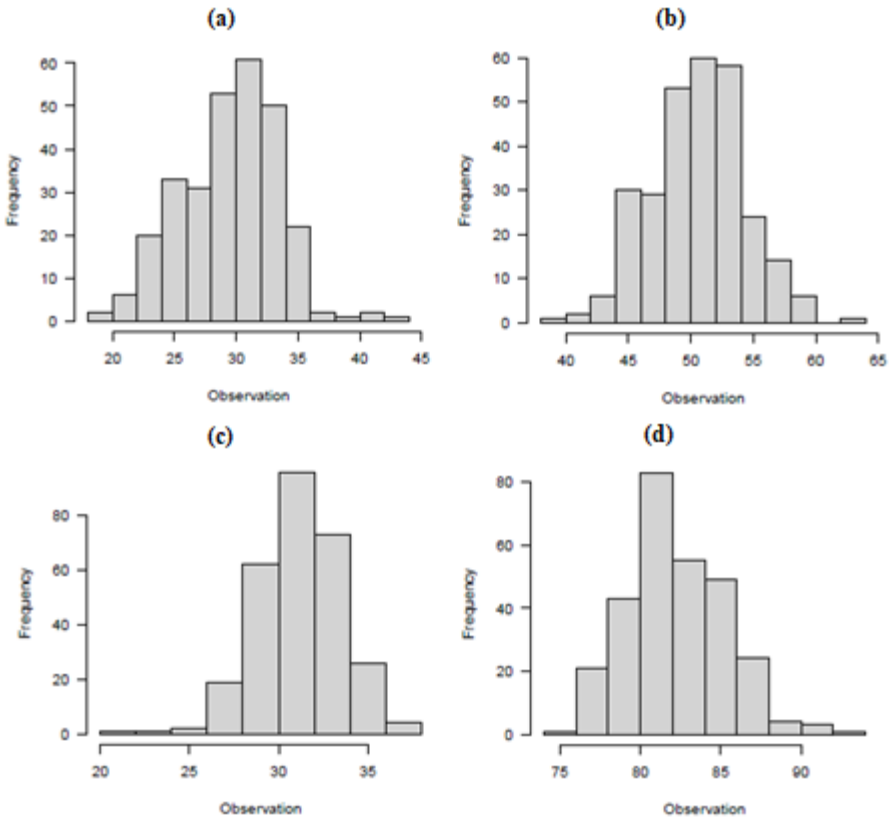

Figure 2. Distribution histograms for seed hardness (a), seed lightness (b), seed chroma (c), and seed hue (d). Hardness and hue were skewed slightly to the right, while chroma skewed slightly to the left, and lightness had a normalized distribution.

Figure 3. Phylogenetic tree of the genome-wide association panel. The phylogenetic tree is divided into three distinct sub-populations; subpopulation A (red), subpopulation B (green), and subpopulation C (blue).

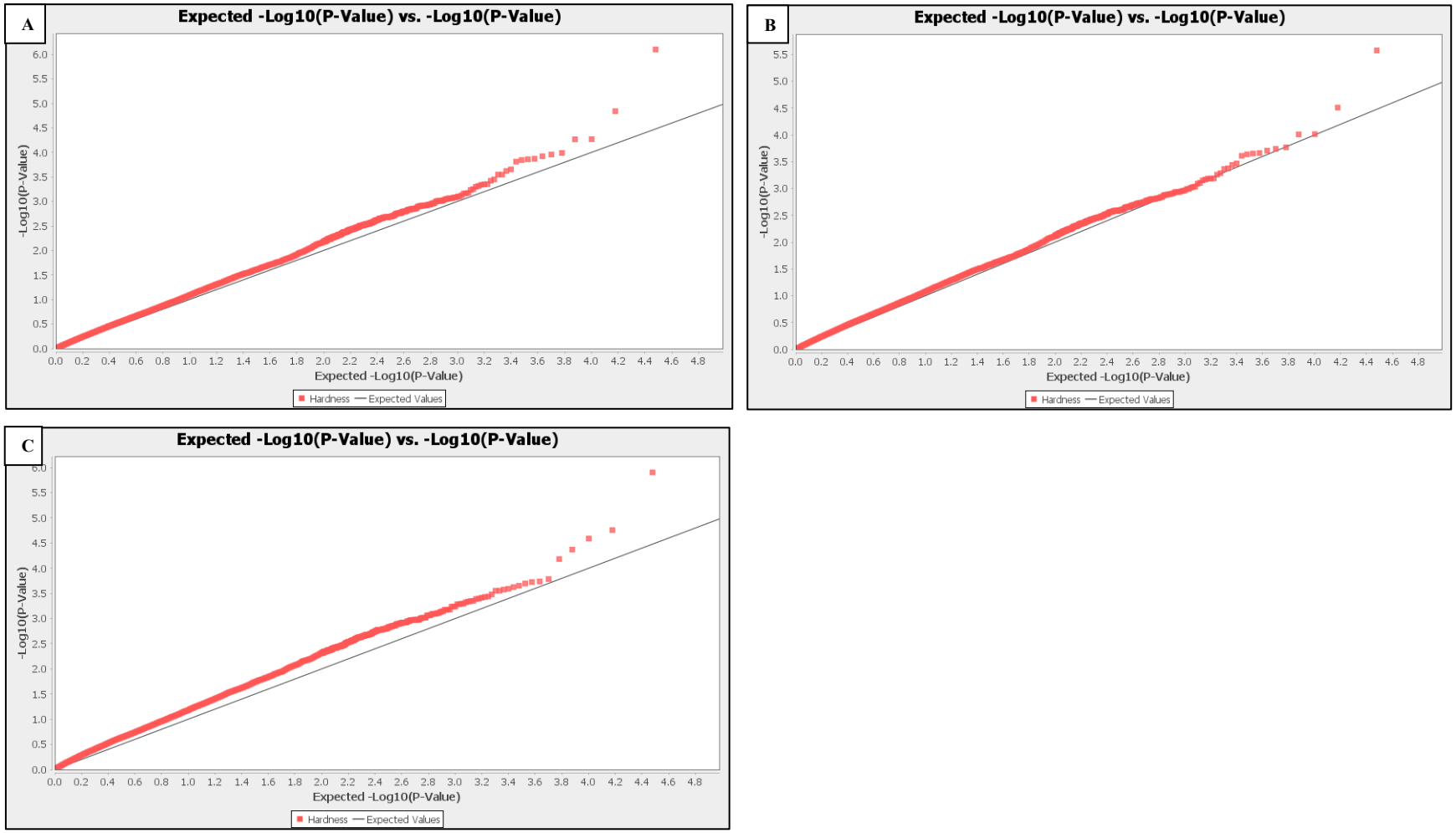

Fig. 4. Quantile-Quantile plots for seed hardness genome-wide association studies (GWAS). The plots correspond to the following models: A, general linear model (GLM); B, mixed linear model (MLM); C, single-marker regression (SMR).

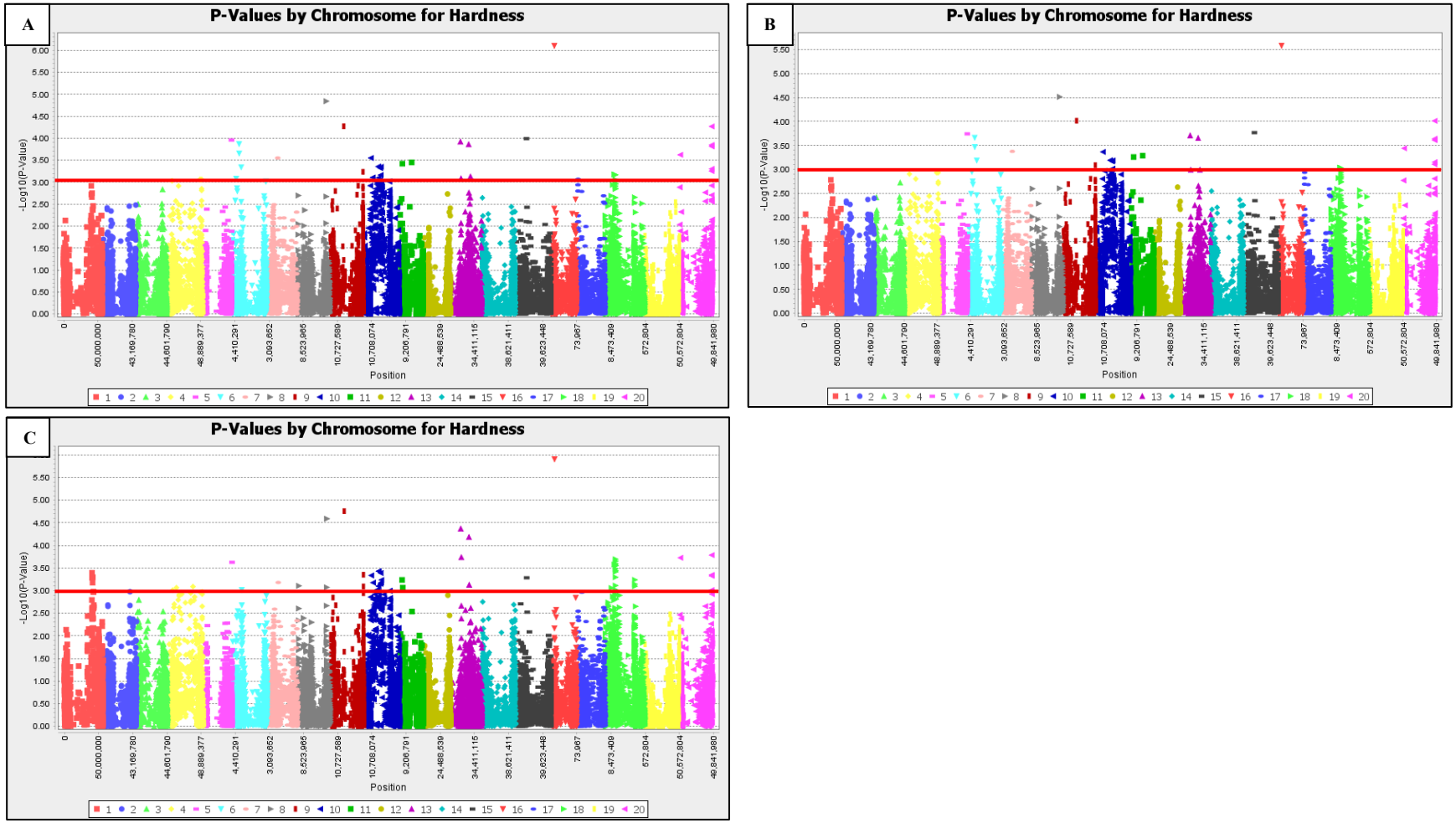

Fig. 5. Manhattan plots for seed hardness genome-wide association studies (GWAS). The plots correspond to the following models: A, general linear model (GLM); B, mixed linear model (MLM); C, single-marker regression (SMR). The red line indicates the significance threshold (LOD > 3.0).

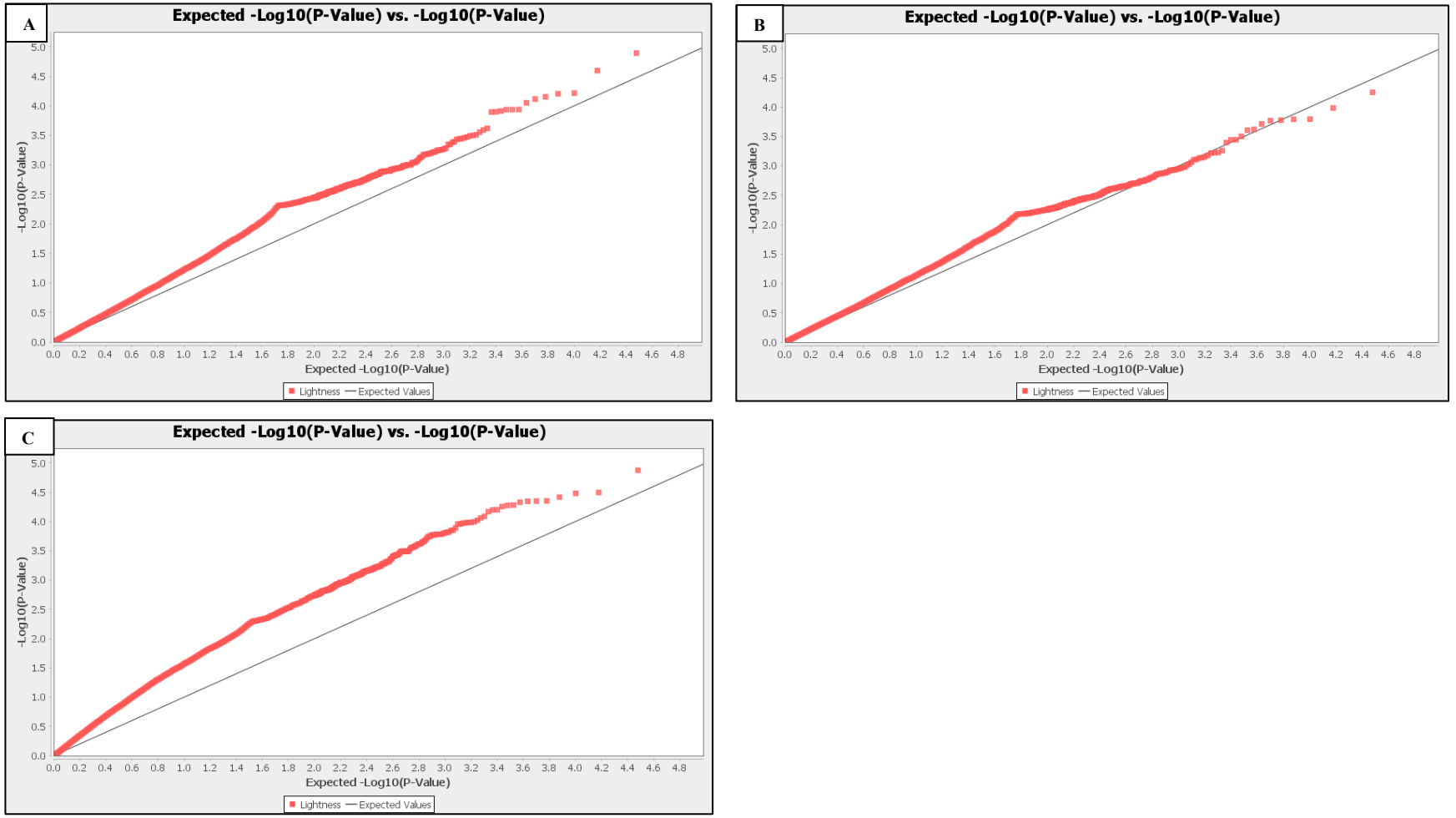

Fig. 6. Quantile-Quantile plots for seed lightness genome-wide association studies (GWAS). The plots correspond to the following models: A, general linear model (GLM); B, mixed linear model (MLM); C, single-marker regression (SMR).

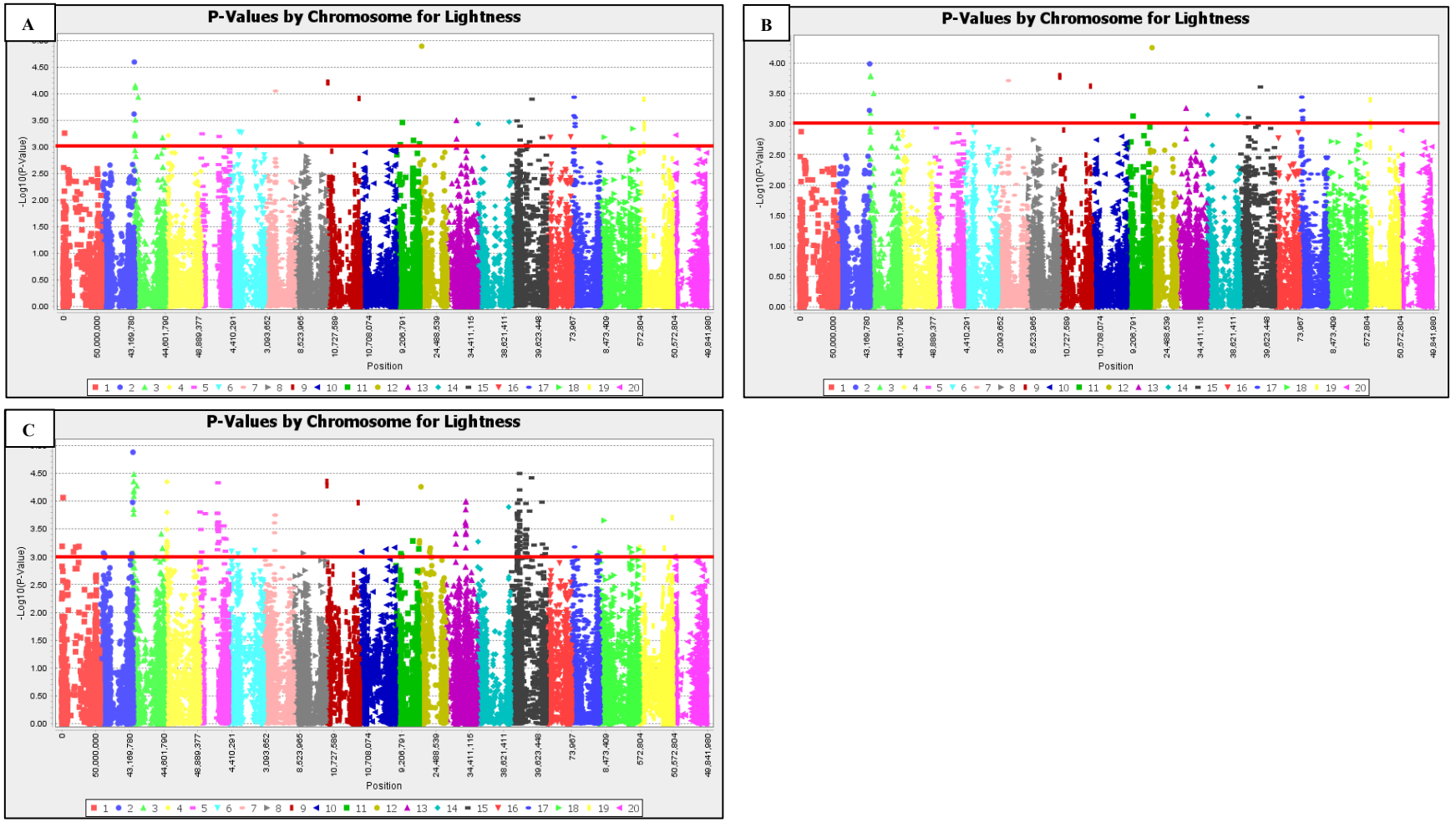

Fig. 7. Three Manhattan plots for seed lightness genome-wide association studies (GWAS). The plots correspond to the following models: A, general linear model (GLM); B, mixed linear model (MLM); C, single-marker regression (SMR). The red line indicates the significance threshold (LOD > 3.0).

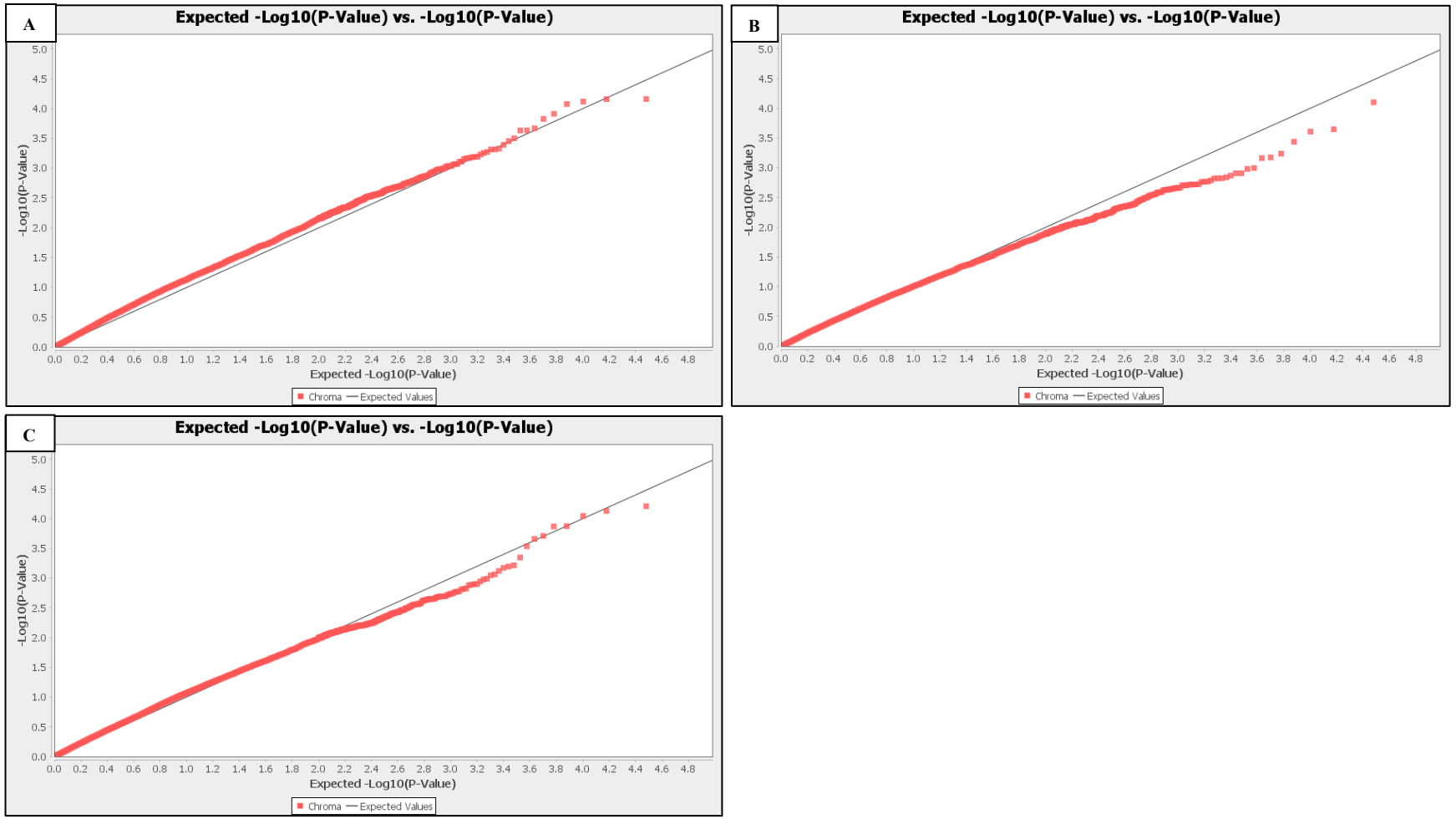

Fig. 8. Three Quantile-Quantile plots for seed chroma genome-wide association studies (GWAS). The plots correspond to the following models: A, general linear model (GLM); B, mixed linear model (MLM); C, single-marker regression (SMR).

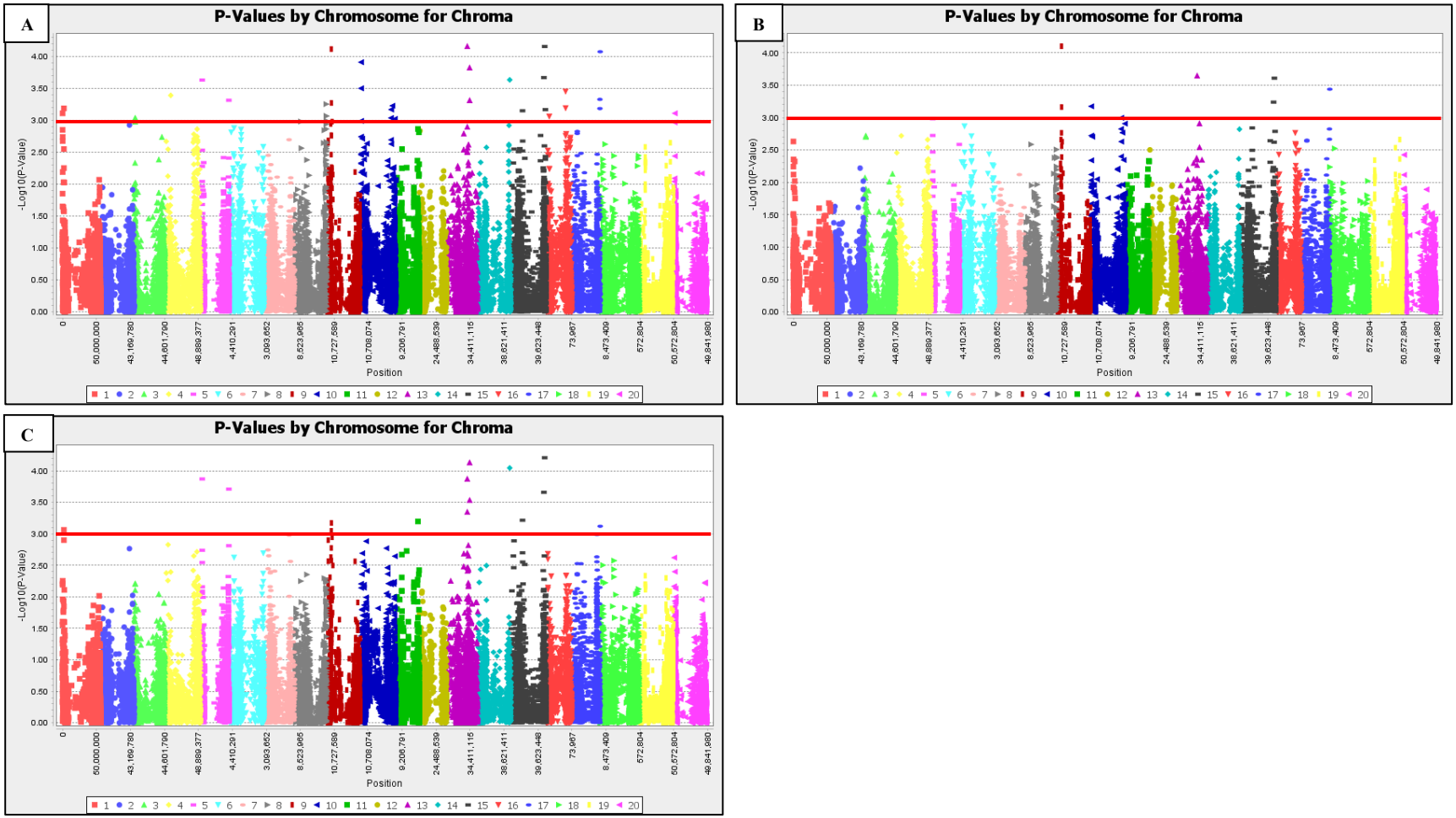

Fig. 9. Three Manhattan plots for seed chroma genome-wide association studies (GWAS). The plots correspond to the following models: A, general linear model (GLM); B, mixed linear model (MLM); C, single-marker regression (SMR). The red line indicates the significance threshold (LOD > 3.0).

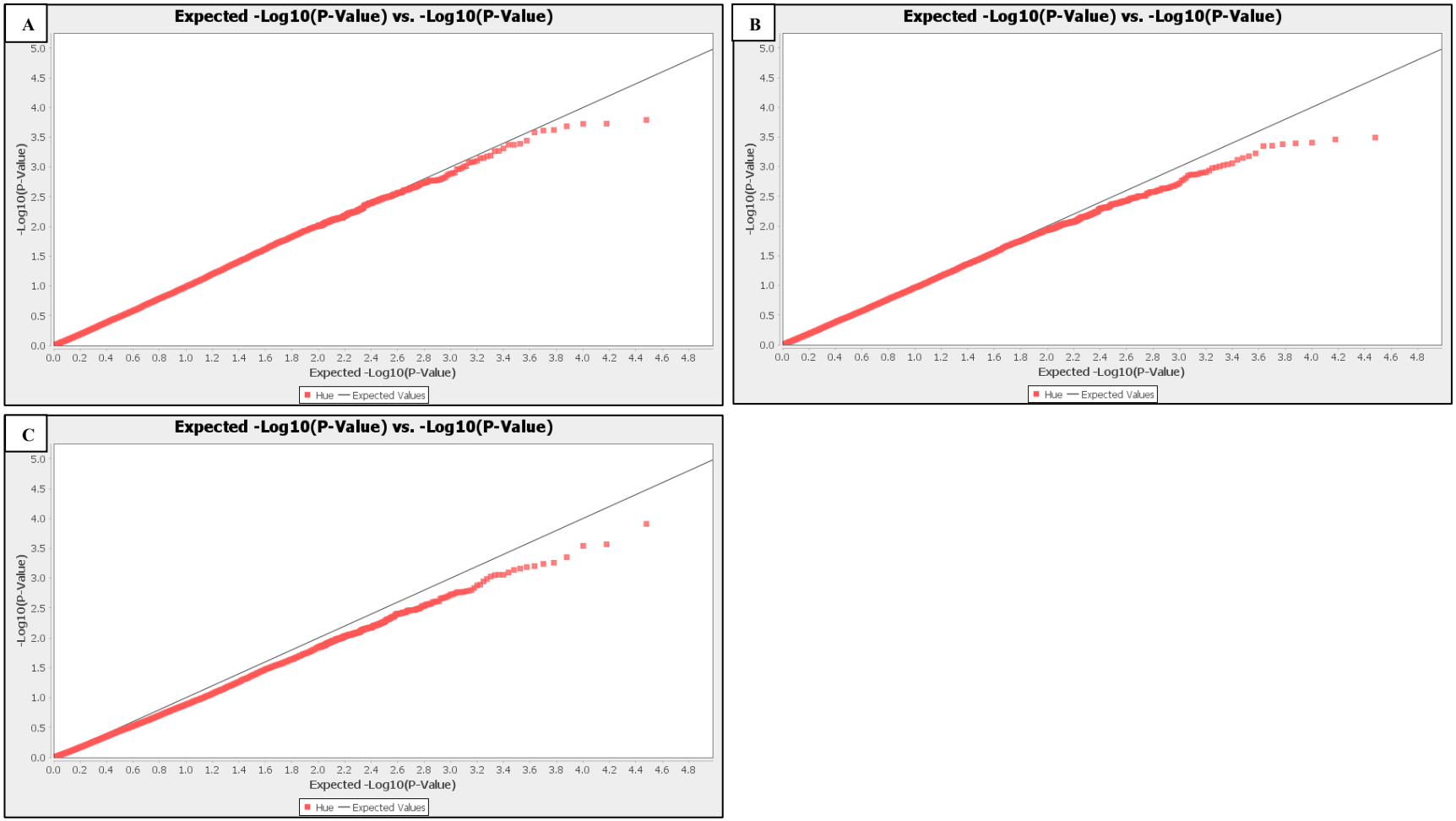

Fig. 10. Three Quantile-Quantile plots for seed hue genome-wide association studies (GWAS). The plots correspond to the following models: A, general linear model (GLM); B, mixed linear model (MLM); C, single-marker regression (SMR).

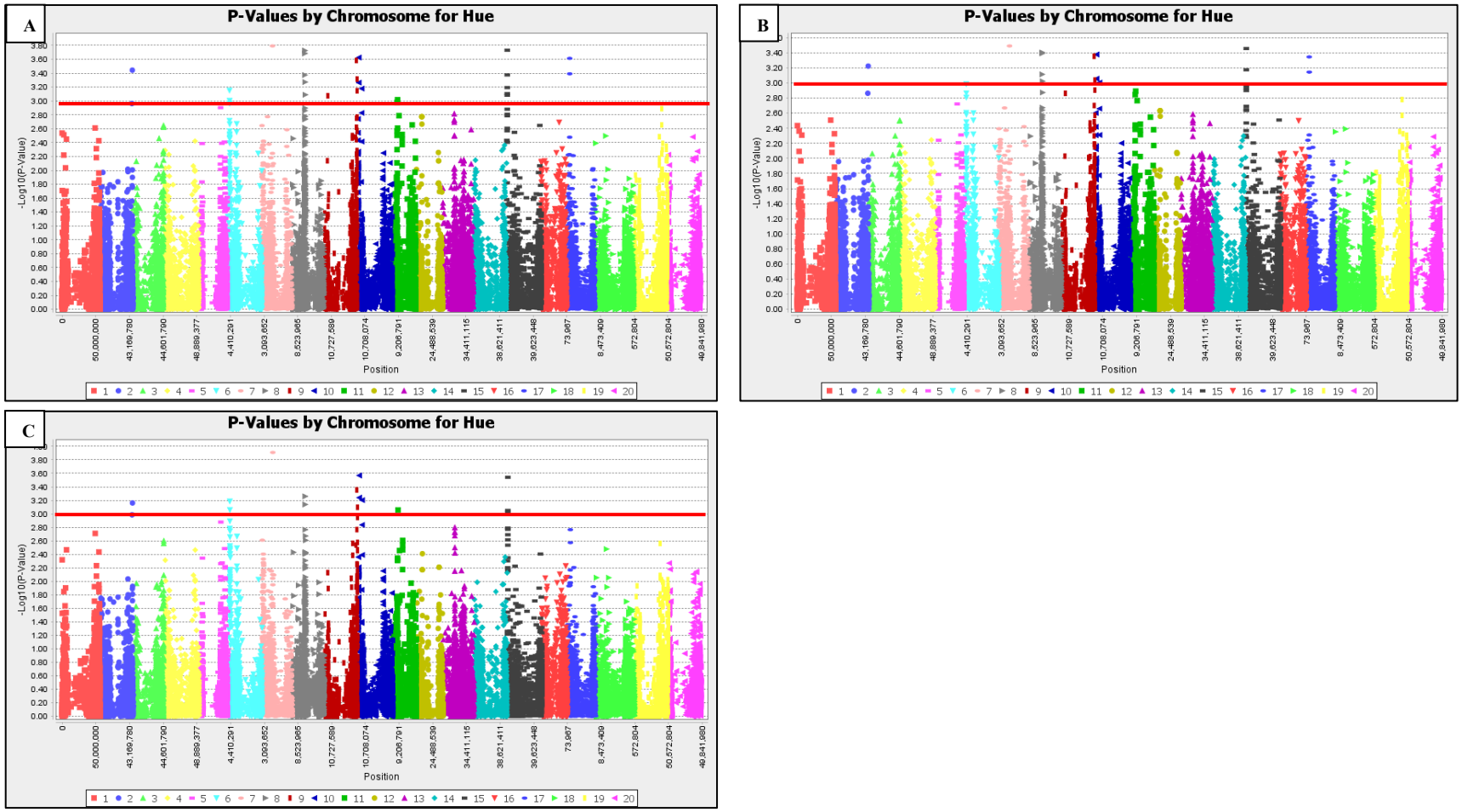

Fig. 11. Three Manhattan plots for seed hue genome-wide association studies (GWAS). The plots correspond to the following models: A, general linear model (GLM); B, mixed linear model (MLM); C, single-marker regression (SMR). The red line indicates the significance threshold (LOD > 3.0).

BIBLIOGRAPHY

Argel, P.J. and Paton, C.J. 1999. Overcoming legume hardseededness. *Forage Seed Production: Tropical and Subtropical Species, Vol. 2*. Eds. Loch, D.S. and Ferguson, J.E. CABI, Wallingford, UK. ISBN: 978-0-85199-191-7

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics, 23(19): 2633-5. DOI: 10.1093/bioinformatics/btm308

Technavio Research. 2022. Natto market size to grow by USD 1.39 Bn, driven by growing consumer awareness regarding health benefits of natto. Published by Cision PR Newswire. Web. Accessed 21 March 2023. https://www.prnewswire.com/news-releases/natto-market-size-to-grow-by-usd-1-39-bn--driven-by-growing-consumer-awareness-regarding-health-benefits-of-natto--technavio-301460556.html

Dhaubhadel, S., Gijzen, M., Moy, P., Farhangkhoee, M. 2006. Transcriptome Analysis Reveals Critical Role of CHS7 and CHS8 Genes for Isoflavonoid Synthesis in Soybean Seeds. Plant Physiology, 143 (1): 326-338. DOI: 10.1104/pp.106.086306

Doherty, A., Smith-Byrne, K., Ferreira, T., Holmes, M.V., Holmes, C., Pulit, S.L., and Lindgren, C.M. 2018. GWAS identifies 14 loci for device-measured physical activity and sleep duration. Nature Communications, 9, 5257. DOI: 10.1038/s41467-018-07743-4

Doyle, J. and Doyle, J.L. 1990. Isolation of Plant DNA from Fresh Tissue. Focus, 12, 13-15.

Gao, R., Han, T., Xun, H., Zeng, X., Li, P., Li, Y., Wang, Y., Shao, Y., Cheng, X., Feng, X., Zhao, J., Wang, L., Gao, X. 2021. MYB transcription factors GmMYBA2 and GmMYBR function in a feedback loop to control pigmentation of seed coat in soybean. Journal of Experimental Botany, 72(12): 4401-4418. DOI: 10.1093/jxb/erab152

Geater, C., Fehr, W., and Wilson, L. 2000. Association of soybean seed traits with physical properties of natto. Crop Science, 40:1529-1534. DOI: 10.2135/cropsci2000.4061529x

Gillman, J.D., Tetlow, A., Lee, J.D., Shannon, J.G., Bilyeu, K. 2011. Loss-of-function mutations affecting a specific Glycine max R2R3 MYB transcription factor result in brown hilum and brown seed coats. BMC Plant Biology, 11: 155. DOI: 10.1186/1471-2229-11-155

Hayes, Ben. 2013. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). *Genome-Wide Association Studies and Genomic Prediction, Chapter 6.* Methods in Molecular Biology, vol 1019. Eds. Gondro, C., and van der Werf, J. DOI: 10.1007/978-1-62703-447-0_6

Hirata, K., Masuda, R., Tsubokura, Y., Yasui, T., Yamada, T., Takahashi, K., Nagaya, T., Sayama, T., Ishimoto, M., and Hajika, M. 2014. Identification of quantitative trait loci associated with boiled seed hardness in soybean. Breeding Science, 64: 362-370. DOI: 10.1270/jsbbs.64.362.

Hossian, S., Panozzo, J.F., Pitock, C., Ford, R. 2011. Quantitative trait loci analysis of seed coat color components for selective breeding in chickpea (*Cicer arietinum,* L.). Plant Science, 91: 49-55. DOI: 10.4141/CJPS10112

Huang, M., Liu, X., Zhau, Y., Summers, R.M., Zhang, Z. 2019. BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience. 8(2): giy154. DOI: 10.1093/gigascience/giy154

Lipka, A.E., Tian, F., Wang, Q., Pieffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., and Zhang, Z. 2012. GAPIT: Genome association and prediction integrated tool. Bioinformatics, 28(18): 2397-2399. DOI: 10.1093/bioinformatics/bts444.

Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. PLoS Genet 12(2): e1005767. DOI: 10.1371/journal.pgen.1005767

McLellan, M.R., Lind, L.R., and Kime, R.W. 1995. Hue Angle Determinations and Statistical Analysis for Multiquadrant Hunter L,a,b Data. Journal of Food Quality, 18: 235-240. DOI: 10.1111/j.1745-4557.1995.tb00377.x

Mullin, W.J. and Wu, W. 2001. Study of Soybean Seed Coat Components and Their Relationship to Water Absorption. Journal of Agriculture and Food Chemistry, 49(11): 5331-5335. DOI: 10.1021/jf010303s

Myers, J.R., Wallace, L.T., Moghaddam, S.M., Kleintop, A.E., Echeverria, D., Thompson, H.J., Brick, M.A., Lee, R., and McClean, P.E. 2019. Improving the Health Benefits of Snap Bean: Genome-Wide Association Studies of Total Phenolic Content. *Nutrients*, 11(10), 2509. DOI: 10.3390/nu11102509

North Dakota Soybean Council. 2018. 'Association Applauds the North Dakota Soybean Council.' *The North Dakota Soybean Grower Magazine*, p. 13. August 2018.

Orazaly, M., Chen, P., Zeng, A., and Zhang, B. 2015. Identification and Confirmation of Quantitative Trait Loci Associated with Seed Hardness. Crop Science. 55(2):688-694. DOI: 10.2135/cropsci2014.03.0219

Palmer R.G., Pfeiffer T.W., Buss G.R., Kilen T.C. 2004. Qualitative Genetics. *Soybeans: Improvement, Production, and Uses.* 3rd ed. ASA, CSSA, and SSSA, Madison, WI. p. 137–214.

Perez P, de los Campos G (2014). "Genome-Wide Regression and Prediction with the BGLR Statistical Package." Genetics, 198(2), 483–495.

Qutob, D., Ma, F., Peterson, C.A., Bernards, M.A., and Gijzen, M. 2008. Structural and permeability properties of the soybean seed coat. Botany. 86(3): 219-227. DOI: 10.1139/B08-002

Shou, S., Meyer, C.J, Ma, F., Peterson, C.A., Bernards, M.A. 2007. The outermost cuticle of soybean seeds: chemical composition and function during imbibition. Journal of Experimental Botany, 58(5): 1071-1082. DOI: 10.1093/jxb/erl268

Song, J., Liu, Z., Hong, H., Ma, Y., Tian, L., Li, X., Li, Y.H., Guan, R., Guo, Y., and Qiu, L.J. 2016. Identification and Validation of Loci Governing Seed Coat Color by Combining Association Mapping and Bulk Segregation. PLoS ONE. 11(7). DOI: 10.1371/journal.pone.0159064

Southern Regional Climate Center. 2021. Texas A&M University. Accessed 6 March 2023.

Taira, H. 1990. Effect of cultivar, seed size, and crop year on total and free sugar contents of domestic soybeans. Nippon Shokuhin Kogyo Gakkaishi, 37:203-213.

Takahashi, R., Yamagishi, N., Yoshikawa, N. 2012. A MYB Transcription Factor Controls Flower Color in Soybean. Journal of Heredity, 104(1): 149-153. DOI: 10.1093/jhered/ess081

Wei, Q. and Chang, S.K.C. 2004. Characteristics of Fermented Natto Products as Affected by Soybean Cultivars. Journal of Food Processing and Preservation, 28: 251-273. DOI: 10.1111/j.1745-4549.2004.23047.x

Williams, C.J., Li Z., Harvey N., et al. 2021. Genome-wide association study of response to interval and continuous exercise training: the Predict-HIIT study. Journal of Biomedical Science, 28, 37. DOI: 10.1186/s12929-021-00733-7

Yang, K., Jeong, N., Moon, J.K., Lee, Y.H., Lee, S.H., Kim, H.M., Hwang, C.H., Back, K., Palmer, R.G., and Jeong, S.C. 2010. Genetic Analysis of Genes Controlling Natural Variation of Seed Coat and Flower Colors in Soybean. Journal of Heredity. 101(6):757-768. DOI: 10.1093/jhered/esq078

Yoshikawa, Y., Chen, P., Zhang, B., Scaboo, A., and Orazaly, M. 2013. Evaluation of seed chemical quality traits and sensory properties of natto soybean. Food Chemistry, 153:186-192. 2014. DOI: 10.1016/j.foodchem.2013.12.027

Zhang, B., Chen, P., Chen, C.Y., Wang, D., Shi, A., Hou, A., and Ishibashi, T. 2008a. Quantitative Trait Loci Mapping of Seed Hardness in Soybean. Crop Science, 48: 1341-149. DOI: 10.2135/cropsci2007.10.0544

Zhang, B., Chen, P., Shi, A., Hou, A., Ishibashi, T., and Wang, D. 2008b. Putative Quantitative Trait Loci Associated with Calcium Content in Soybean Seed. Journal of Heredity. 100(2): 263-269. DOI: 10.1093/jhered/esn096