University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Theses, Dissertations, and Student Research in Agronomy and Horticulture

Agronomy and Horticulture Department

Fall 7-2016

Distribution of Genomic Variation in the USDA Soybean Germplasm Collection and Relationship with Phenotypic Variation

Nonoy Batiller Bandillo
University of Nebraska-Lincoln

Follow this and additional works at: http://digitalcommons.unl.edu/agronhortdiss


Part of the Agricultural Science Commons, Agronomy and Crop Sciences Commons, Genomics Commons, Molecular Genetics Commons, and the Plant Breeding and Genetics Commons

Bandillo, Nonoy Batiller, "Distribution of Genomic Variation in the USDA Soybean Germplasm Collection and Relationship with Phenotypic Variation" (2016). *Theses, Dissertations, and Student Research in Agronomy and Horticulture*. 114.
http://digitalcommons.unl.edu/agronhortdiss/114

This Article is brought to you for free and open access by the Agronomy and Horticulture Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Theses, Dissertations, and Student Research in Agronomy and Horticulture by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

DISTRIBUTION OF GENOMIC VARIATION IN THE USDA SOYBEAN GERMPLASM COLLECTION AND RELATIONSHIP WITH PHENOTYPIC VARIATION

by

Nonoy B. Bandillo

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Agronomy and Horticulture

Under the Supervision of Professors Aaron J. Lorenz and George L. Graef

Lincoln, Nebraska

July, 2016

DISTRIBUTION OF GENOMIC VARIATION IN THE USDA SOYBEAN GERMPLASM COLLECTION AND RELATIONSHIP WITH PHENOTYPIC VARIATION

Nonoy B. Bandillo, Ph.D.

University of Nebraska, 2016

Advisors: Aaron J. Lorenz and George L. Graef

The USDA Soybean Germplasm Collection harbors a large stock of genetic diversity with potential to accelerate soybean cultivar development. The extent and nature of favorable alleles contained in the collection are not well known nor is the distribution of genetic variation and how it relates to phenotypic variation. The genotyping of the entire USDA Soybean Germplasm Collection marked the beginning of a systematic exploration of genetic diversity for genetic research and breeding. In this research, we conducted the first comprehensive analysis of population structure on the collection of ~14,400 soybean accessions [*Glycine max* (L.) Merr. and *G. soja* Siebold & Zucc.] that were genotyped using a 50KSNP chip. Accessions originating from Japan and Korea diverged from the Chinese accessions. The ancestry of founders of the American accessions derived mostly from two Chinese subpopulations, which reflects the composition of the American accessions as a whole. A genome-wide association study on ~12,000 accession conducted on seed protein and oil is the largest reported to date in plants and identified strong single nucleotide polymorphisms (SNPs) signals on chromosomes 20 and 15. The haplotype effects of the chromosome 20 region show a strong negative relationship between oil and protein at this locus, indicating negative pleiotropic effects or multiple closely linked loci in repulsion phase linkage. Genome-wide association mapping for ten descriptive traits

identified a total of 23 known genes and unknown genes controlling the phenotypic variants. Because some of those genes had been cloned, we were able to show that the narrow SNP signal regions had chromosomal base pair spans that, with few exceptions, bracketed the base pair region of the cloned gene coding sequences, despite variation in SNP distribution of chip SNP set. We also elucidate the genetic basis of local adaptation in soybean by exploring the natural variation available in 3,012 locally adapted landrace accessions from across the geographical range of soybean. Our approach using selection mapping and landscape genomic association methods identified important candidate genes related to drought and heat stress, and revealed important signatures of directional selection that are likely involved on geographic divergence of soybean.

ACKNOWLEDGEMENTS

I would like to sincerely thank the following persons for making this dissertation, and my sojourn at Nebraska, come to its fruition. I am indebted to my major advisor, Dr. Aaron Lorenz, who took a chance on me when I was a graduate student applicant, and he taught me well. I am indebted to Dr. George Graef for entrusting me the genomic selection project, and for being such a great advisor for more than a year. I acknowledge my committee members for taking the time to work with me and for providing extra expertise throughout this undertaking: Dr. Steve Baenziger, Dr Keenan Amundsen, and Dr. Steve Kachman. I also thank Dr. Jim Specht for sharing his expertise on soybean breeding and genetics as well as for being such a great mentor.

I couldn't have finished my work without the help of my colleagues from the Lorenz Lab (Diego, Amrit, Jon, Kadam, and Eerin) that had been really helpful on increasing my breadth of knowledge all the way from sharing papers, updating research progress, and presenting new ideas. I like to thank the UNL Soybean Breeding Team for a wonderful exposure in a soybean field. I can now cross-pollinate soybean! I also thank the UNL Filipino Student Association.

I would not be where I am today without my family. I thank my parents and my three siblings (Norbel, Niña and Nicky) for their love, support and never-ending encouragement. Special thanks to my fiancee, Fatima Amor Tenorio, for being crazy enough to date a budding plant breeder like me. She is a constant source of inspiration and encouragement during this busy time of my life.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

LIST OF TABLES

APPENDIX

Chapter 2

Chapter 3

Chapter 4

CHAPTER 1: INTRODUCTION

The C3 legume crop soybean (*Glycine max* [L.] Merr.) is the leading oil seed crop produced and consumed in the world today, and accounts for 29% of world production (Song et al., 2015). Soybean has a wide range of latitudinal adaptation in both North and South hemispheric geographical locations, which reflects its complex domestication origin and subsequent breeding history (Carter et al., 2004). The domestication of *G. max* from its wild species progenitor (*Glycine soja* Sieb. and Zucc.) occurred in China ~5,000 years ago (Carter et al., 2004; Hyten et al., 2006). Cultivation of soybean expanded from China to Korea and Japan about 2000 years ago (Kihara, 1969). It may have been introduced into North America in 1765, and into Central and South America during the first half of the last century (Hymowitz, 2004). Because of this ancient origin and the diffusion via the trading of soybean seeds, landrace adaptation to local climates and cultural practices occurred which resulted in a multitude of localized *G. max* landraces (Hyten et al., 2006). An estimated 45,000 unique Asian landraces have been collected and are maintained in G. *max* germplasm collections around the world (Carter et al., 2004). Despite this seemingly vast reservoir of genetic diversity, only 0.02% (N=80) of those landraces account for 99% of the collective parentage of North American soybean cultivars released between 1947 and 1988 (Li et al., 2002; Carter et al, 2014). Seventeen of these 80 landraces account for 86% of the collective parentage while the contribution of the remaining landraces is less than 1% (Gizlice et al., 1994; Li and Nelson, 2001; Ude et al., 2003, Carter et al., 2004). It is presumed that these genetic bottlenecks have reduced the genetic diversity of modern soybean.

Germplasm collections serve as an important source of variation for germplasm enhancement that can potentially sustains long-term genetic gain in breeding programs (Jarquin et al., 2016). In the USA, there are 30 USDA-ARS GRIN NPGS Plant Collection Sites (http://www.ars-grin.gov/npgs/sitelist.html). Each of those 30 sites exists for the collection, preservation, and evaluation of accession in major and minor plant and crop species that are of national interest. The soybean repository is located at Urbana, IL USA (https://npgsweb.ars-grin.gov/gringlobal/site.aspx?id=24), and contains accessions of the two annual species – the wild *G. soja* and the cultivated *G. max*, and accessions of each of 19 perennial Glycine species. According to data collected by the International Plant Genetic Resources Institute (IPGRI) (2001), more than 170,000 *G. max* accessions are maintained by more than 160 institutions in nearly 70 countries. The USDA Soybean Germplasm Collection (hereafter referred to as the Collection) is one of the most intensely used germplasm collections in the world, and the most intensely used in the NPGS (Nelson, 2011). Extensive soybean collecting started in the 1920s but systematic preservation did occur until the USDA Soybean Collection was established in 1949 (Carter et al., 2004). A large part of the accessions (~5,000) were collected as part of the expedition of P.H. Dorsett and W.J. Morse in Asia between 1924 and 1932 (Nelson, 2011). To date, the entire Collection contained approximately ~22,000 soybean accessions. The collection includes more than 1,100 wild soybeans from China, Korea, Japan and Russia, and more than 18,000 cultivated soybeans from China, Korea, Japan, and 84 other countries (Song et al, 2015). Most of the cultivated soybean from China, Korea, and Japan are landraces that are not the product of modern plant breeding (Carter et al., 2004).

Curators of germplasm collections are charged with evaluating existent and newly acquired accessions to provide data on traits of interest to breeders and other researchers. Over time, a substantial amount of phenotypic data has been collected in nearly all germplasm collections. Relative to the annual *Glycine* accessions, nearly all have been phenotypically characterized for ten "primary" descriptor traits by the curator and collaborators. The word primary is used as an adjective here because these key traits are also used by soybean breeders to characterize new created breeding lines (*vis-à-vis* existing cultivars) when such lines are submitted for agronomic evaluation in the Northern and Southern Uniform trials. The ten primary traits are Maturity Group, Stem Termination Type, Flower Color, Pubescence Color, Form, and Density, Pod Color, Seed Coat Luster and Color, and Hilum Color. At least two and often several phenotypes are listed as categories for each trait. Thus, a primary trait description of a (hypothetical) soybean accession might be **IV D P T N E Br D Y Bl**. The ten codes shown here correspond to one of multiple coded phenotypic categories possible for each of the ten foregoing traits that are successively ordered here. The phenotypic category names and codes for each trait can be found at https://npgsweb.ars-grin.gov/gringlobal/descriptors.aspx.

Aside from the ten descriptor traits, quantitative phenotypic data are also routinely collected by the USDA researchers and collaborators on the collection. Traits included in phenotypic evaluations are flowering date, maturity date, lodging, height, stem termination, shattering, seed quality, seed mottling, seed weight and yield (Bernard et al., 1998). Phenotypic evaluations are conducted periodically to characterize newly acquired accessions. Accessions originally classified as maturity group 000-I are mostly evaluated in Minnesota. Maturity groups I – IV are predominantly evaluated in Urbana while

accessions belonging to MGs V – IX are evaluated in Stoneville, Mississippi (https://npgsweb.ars-grin.gov/gringlobal/method.aspx?id=11002).

Past soybean inheritance studies involving both qualitative and quantitative traits have led to the assignment of gene symbols to the two (or three) alleles at each of the one or more loci that were inferred to govern the trait. Palmer et al. (2004) listed 251 soybean classical gene loci, and also noted that 72 of these were members of 20 classical linkage groups (CLGs). Based on molecular marker genotyping of bi-parental mapping populations in which some of those 72 classical genes were segregating (e.g., Shoemaker and Specht, 1995), the linkage of those markers with some of those genes has led to the assignment of 19 of those 20 CLGs to molecular linkage groups (MLGs) that were labeled A1 to O (Cregan et al., 1999). Still, the majority of the classical gene loci have yet to be genetically mapped. Moreover, genetically mapped classical gene loci have low resolution centi-Morgan (cM) positions, except for a few cloned genes that now have a chromosomal base pair (bp) position. Establishing a genetic map position for all (or even a majority of the) 251 classical gene loci would be an laborious and expensive endeavor using a bi-parental mapping population approach, because (1) many such populations would need to be created from parents with contrasting classical gene phenotypes, (2) all progeny in each population would have to be genotyped with a suitably dense array of molecular markers, and (3) each population would have to consist of a very large number of F_2 progeny to ensure that sufficient numbers of classical gene – marker recombinants would be available for the desired mapping resolution. Most soybean geneticists would consider such a project impractical.

Genome-wide association study (GWAS) offers a powerful strategy for elucidating the genetic architecture of qualitative and quantitative traits (Huang et al., 2011; Wang et al., 2012; Romay et al., 2015). Compared to traditional linkage mapping, GWA analysis provides much greater mapping resolution and evaluates greater allelic diversity simultaneously (Myles et al., 2009; Yu and Buckler, 2006; Zhu et al., 2008). Harnessing the genetic variation contained in crop germplasm collections for mapping QTL through GWA has been found successfully conducted in several crop species, including barley (Hordeum vulgare L.) (Munoz-Amatriain et al., 2014), maize (*Zea mays* L.) (Romay et al., 2013), rice (*Oryza sativa* L.) (Huang et al., 2010), and wheat (*Triticum aestivum* L.) (Cavanagh et al., 2013). To date, only a small fraction of the soybean germplasm repositories around the world has been explored through GWAS and such studies have typically included relatively small (i.e., <1000) numbers of accessions in any given population (Hwang et al.,2014; Sonah et al., 2015; Vaughn et al., 2014).

Next-generation sequencing technologies have enabled sequencing of a large number of accessions at relatively low cost providing opportunities to perform large-scale GWAS as well as inspect the genomic regions selected in the history of crop improvement. Some present examples of wide-scale genotypic characterization of the germplasm collections include the genotyping by sequencing of the CIMMYT maize collection (Hearne et al., 2015) and the sequencing of 3,000 rice genomes (Li et al., 2014). More recently, the entire USDA Soybean Germplasm Collection has been genotyped with 50K SNPs (Song et al., 2015) through a collective effort of the USDA Agricultural Research Service Soybean Genomics Group (Song et al., 2015). The availability of such information

creates a tremendous resource for dissecting genotype-phenotype relationship and understanding the distribution of genomic variation in the Collection.

In this study, we leverage the fantastic genomic resources available in the USDA Soybean Germplasm Collection to meet the following objectives: 1) Perform a comprehensive population structure analysis on the entire collection, 2) Demonstrate mapping resolution of increased genetic diversity for detecting genetic variant, 3) Uncover the genetic architecture and important genes underlying economically important traits, 4) Determine the frequency and distribution of alleles/haplotypes governing economically important traits, and 5) Provide insights on the genetic basis of local adaptation. The results reported herein provide a fuller understanding of the distribution of genetic variation contained within the collection and how it relates to phenotypic variation for economically important traits.

2 REFERENCES

Bernard, R.L., C.R. Cremeens, R.L. Cooper, F.L. Collins, O.A. Krober, K.L. Athow, F.A. Laviolette, C.J. Coble, and R.L. Nelson. 1988. Evaluation of the USDA Soybean germplasm collection: Maturity groups 000 to IV (FC 01.547-PI 266.807). U.S. Department of Agriculture, Tech. Bull. 1844.

Carter, T. E., R. L. Nelson, C. H. Sneller, and Z. Cui. 2004. Genetic Diversity in Soybean. Soybeans: Improvement, Production, and uses (American Society of Agronomy Monograph Series): 303-416.

Cavanagh, C.R., S. Chao, S. Wang, B.E. Huang, S. Stephen, S. Kiani, K. Forrest, C. Saintenac, G.L. Brown-Guedira, A. Akhunova, D. See, G. Bai, M. Pumphrey, L. Tomar, D. Wong, S. Kong, M. Reynolds, M.L. da Silva, H. Bockelman, L. Talbert, J.A. Anderson, S. Dreisigacker, S. Baenziger, A. Carter, V. Korzun, P.L. Morrell, J. Dubcovsky, M.K. Morell, M.E. Sorrells, M.J. Hayden, and E. Akhunov. 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc. Natl. Acad. Sci. U. S. A. 110:8057-8062.

Cregan P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, et al. 1999. An integrated genetic linkage map of the soybean. Crop Sci. 39:1464-1490.

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

Hearne, S., J. Franco, J. Chen, C. P. Sansaloni, C. D. Petroli et al., 2015. Genome wide assessment of maize genebank diversity; synthesis of next generation technologies and GIS based approaches. Proceedings of Plant and Animal Genome XXIII, San Diego, CA.

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, C. Li, C. Zhu, T. Lu, Z. Zhang, M. Li, D. Fan, Y. Guo, A. Wang, L. Wang, L. Deng, W. Li, Y. Lu, Q. Weng, K. Liu, T. Huang, T. Zhou, Y. Jing, W. Li, Z. Lin, E.S. Buckler, Q. Qian, Q.F. Zhang, J. Li, and B. Han. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961-967.

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

Hymowitz, T. 2004. Speciation and cytogenetics: improvement, production, and uses. In: Boerma HR, Specht JE, editors. ASA, CSSA, ASSA. Madison, Wisconsin. pp 97-136.

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

Kihara, H. 1969. History of biology and other sciences in Japan in *Glycine max* retrospect. Proc. XII Intern. Cong. Genetics. 3:49-70.

Li J.-Y., Wang J., Zeigler R. S. 2014. The 3,000 rice genomes project: new opportunities and challenges for future rice research. Giga science 3: 8.

Li, Z.L., and R.L. Nelson. 2001. Genetic diversity among soybean accessions from three countries measured by RAPDs. Crop Sci. 41:1337-1347.

Myles, S., J. Peiffer, P.J. Brown, E.S. Ersoz, Z. Zhang, D.E. Costich, and E.S. Buckler. 2009. Association mapping: Critical considerations shift from genotyping to experimental design. Plant Cell 21:2194-2202.

Nelson R. L. 2011. Managing self-pollinated germplasm collections to maximize utilization. Plant Genet Resour. 9: 123-133.

Palmer R.G., T.W. Pfeiffer, TW, G.R. Buss, and T.C. Kilen. 2004. Qualitative genetics. In: H.R. Boerma and J.E. Specht, editors. Soybeans: improvement, production, and uses. 3rd ed. Agron. Monogr. 16. ASA, CSSA, and SSSA, Madison, WI. p. 137-233.

Romay, M.C., M.J. Millard, J.C. Glaubitz, J.A. Peiffer, K.L. Swarts, T.M. Casstevens, R.J. Elshire, C.B. Acharya, S.E. Mitchell, S.A. Flint-Garcia, M.D. McMullen, J.B. Holland, E.S. Buckler, and C.A. Gardner. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14:R55-2013-14-6-r55.

Sonah, H., L. O'Donoughue, E. Cober, I. Rajcan, and F. Belzile. 2015. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant. Biotechnol. J. 13:211-221.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Ficus, R.L. Nelson, and P.B. Cregan. 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3 5:1999-2006.

Ude, G.N., W.J. Kenworthy, J.M. Costa, P.B. Cregan, and J. Alvernaz. 2003. Genetic diversity of soybean cultivars from China, Japan, North America, and North American ancestral lines determined by amplified fragment length polymorphism. Crop Sci. 43:1858-1867.

Vaughn, J.N., R.L. Nelson, Q. Song, P.B. Cregan, and Z. Li. 2014. The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. G3-Genes Genomes Genetics 4:2283-2294.

Wang, M., N. Jiang, T. Jia, L. Leach, J. Cockram, R. Waugh, et al. 2012. Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theor. Appl. Genet. 124:233-246.

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

Zhu, C., M. Gore, E.S. Buckler, and J. Yu. 2008. Status and prospects of association mapping in plants. Plant Genome 1:5-20.

CHAPTER 2: A POPULATION STRUCTURE AND GENOME-WIDE ASSOCIATION ANALYSIS ON THE USDA SOYBEAN GERMPLASM COLLECTION

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

LICENSE AGREEMENT:

Confirmation Number: 11601826
Order Date: 10/28/2016

Nonoy Bandillo
nbandillo@huskers.unl.edu

Order detail ID: 70145473
Order License ID: 3977690523495
ISSN: 1940-3372
Publication Type: e-Journal
Volume: 8
Issue: 3
Publisher: CROP SCIENCE SOCIETY OF AMERICA
Author/Editor: Crop Science Society of America
Permission Status: Granted

Permission type: Republish or display content
Type of use: Thesis/Dissertation

2.1 ABBREVIATIONS

GRIN, Germplasm Resource Information Network; GWA, genome-wide association; GWAS, genome-wide association study; LD, linkage disequilibrium; LG, linkage group; MG, maturity group; MAF, minor allele frequency; PC, principal component; PCA, principal component analysis; QTL, quantitative trait loci; SNP, single nucleotide polymorphism; SP, subpopulation; USDA, United State Department of Agriculture.

2.2 ABSTRACT

Population structure analyses and genome-wide-association studies (GWAS) conducted on crop germplasm collections provide valuable information on the frequency and distribution of alleles governing economically important traits. The value of these analyses is substantially enhanced when the accession numbers can be increased from 1,000 to~ 10,000 or more. In this research we conducted the first comprehensive analysis of population structure on the collection of 14,000 soybean accessions (*Glycine max* and *G. soja*) using a 50K SNP chip. Accessions originating from Japan were relatively homogenous and distinct from the Korean accessions. As a whole, both Japanese and Korean accessions diverged from the Chinese accessions. The ancestry of founders of the American accessions derived mostly from two Chinese subpopulations, which reflects the composition of the American accessions as a whole. A 12,000 accession GWAS conducted on seed protein and oil is the largest reported to date in plants and identified SNPs with strong signals on chromosomes 20 and 15. A chromosome 20 region previously reported to be important for protein and oil content was further narrowed and now contains only three plausible candidate genes. The haplotype effects show a strong negative relationship between oil and protein at this locus, indicating negative pleiotropic effects or multiple closely linked loci in repulsion phase linkage. The vast majority of accessions carry the haplotype allele conferring lower protein and higher oil. Our results provide a fuller understanding of the distribution of genetic variation contained within the USDA soybean collection and how it relates to phenotypic variation for economically important traits.

2.3 INTRODUCTION

Soybean (*Glycine max* L. Merr.) is an important crop worldwide and a major source of protein and oil for human food, animal feed, and industrial products (Wilson, 2008). The percentages of protein and oil content, while influenced by both genotype and environment, typically average *ca.* 40% and 20%, respectively. Increasing the relative oil content in soybean seed is complicated by its high negative correlation to protein content (Brummer et al., 1997; Burton, 1987; Clemente and Cahoon, 2009; Cober and Voldeng, 2000; Wilcox, 1998) caused by either pleiotropic effects or linkage (Chung et al., 2003). Moreover, total seed yield is often negatively correlated with seed protein, although the correlation is weaker than that between protein and oil (Chung et al., 2003). Dissecting the genetic bases underlying seed oil and protein content, and eventually recombining them in desired genetic backgrounds continues to be a challenge to soybean breeders.

Given the importance of oil and protein content, the genes or quantitative trait loci (QTL) underlying these traits have undergone intensive investigations (Bolon et al., 2010; Chung et al., 2003; Hwang et al., 2014; Vaughn et al., 2014). However, most of what we know about the genetic architecture of seed protein and oil content is based on traditional QTL linkage analysis of populations derived from crosses of two parents with contrasting phenotypes. More than 50 QTL have been reported as controlling seed oil and protein content in a number of QTL mapping studies (www.soybase.org). Among these QTL, a region mapped to soybean linkage group I (LG-I) has consistently shown the strongest association with percent protein composition (Diers et al, 1992; Chung et al, 2003; Nichols et al, 2006). The LG-I QTL is of particular interest due to its large additive effect detected in many mapping populations (Csanadi et al., 2001; Diers et al., 1992; Sebolt et al., 2000)

and across multiple environments (Brummer et al., 1997). Nichols et al. (2006) narrowed down this region to a 3 cM interval using BC_5F_5-derived near isogenic lines. This genetic map interval has a corresponding physical distance of 8.4Mbp, from 24.54Mb to 32.92Mb on chromosome 20 (Bolon et al., 2010). Hwang et al. (2014) further narrowed down the candidate region to a 3 Mb region located between 27.6 Mb to 30.0 Mb on the same chromosome. Vaughn et al. (2014), however, mapped this same QTL approximately 1 Mb downstream of the region that Hwang et al (2014) identified. The size of the narrowed region defined by Hwang et al. (2014) and Vaughn et al. (2104) is still too large for targeting candidate genes for cloning.

Mapping resolution can be greatly enhanced when accession numbers are increased from ~1K to~ 10K or more (Korte and Farlow, 2013). Remarkably, the USDA Agricultural Research Service Soybean Genomics Group has genotyped the entire USDA Soybean Germplasm Collection with the Illumina Infinium SoySNP50K iSelect Beadchip (http://www.soybase.org/dlpages/#snp50k). The availability of this information will provide soybean researchers with a deeper understanding of the genetic variation contained in the germplasm collection, and holds potential to pinpoint important loci controlling traits of interest through genome-wide association (GWA) analysis. Compared to traditional linkage mapping, GWA analysis provides much greater mapping resolution and evaluates greater allelic diversity simultaneously (Myles et al., 2009; Yu and Buckler, 2006; Zhu et al., 2008). Harnessing the genetic variation contained in crop germplasm collections for mapping QTL through GWA has been found successfully conducted in several crop species, including barley (Munoz-Amatriain et al., 2014), maize (Romay et al., 2013), rice (Huang et al., 2010) and wheat (Cavanagh et al., 2013). To date, only a small fraction of

the soybean germplasm repositories around the world has been explored through GWAS and such studies have typically included relatively small (i.e., < 1000) numbers of accessions in any given population (Hwang et al., 2014; Sonah et al., 2015; Vaughn et al., 2014).

Here, we report results from the first analysis of population structure on the entire collection of 14K unique soybean accessions, which included *G. max* and *G. soja* accessions. A GWAS for seed protein and oil content on over 12K unique *G. max* accessions was performed, which is the largest GWAS conducted in plants reported to date. We determined the distribution of favorable alleles among subpopulations defined by world region and maturity group (MG) of the 12K accessions. The results reported herein provide a fuller understanding of the distribution of genetic variation contained within the collection and how it relates to phenotypic variation for economically important traits.

2.4 MATERIALS AND METHODS

2.4.1 Plant Materials

The accessions used in this study are from the U.S. Department of Agriculture (USDA) Soybean Germplasm Collection. The entire collection consists of nearly 22,000 accessions, including modern and land race cultivars (*G. max*); wild relatives of soybean *(G. soja*); and perennial *Glycine* (www.soybase.org). From the whole set, we selected only the annual accessions and dropped accessions determined to be genotypic duplicates (e.g., near-isogenic lines, non-USA duplicates). This filtering left 14,430 unique accessions (Supplementary Fig. S1) collected from 85 countries (Supplementary Table S1, Supplementary Fig. S2a) representing the range of photoperiod/temperature latitudinal

adaptation as defined by a maturity group (MG) Roman numeral designation (Supplementary Table S2 and Supplementary Fig. S2b). Further information for each accession (accession name, accession number, country of origin etc.) can be found in the USDA Germplasm Resources Information Network (GRIN) database (www.ars-grin.gov).

2.4.2 Genotype and Phenotype Data

Genotype data consisted of 52,041 SNPs scored on 14,430 germplasm accessions using the Illumina Infinium SoySNP50K BeadChip as described by Song et al. (2013). SNP genotyping was conducted on the Illumina platform by following the InfiniumH HD Assay Ultra Protocol (Illumina, Inc. San Diego, CA). SNPs were scored using the GenomeStudio Genotyping Module v1.8.4 (Illumina, Inc. San Diego, CA). The SNP data is publicly available at http://www.soybase.org/dlpages/index.php. Markers with MAF < 0.05 were removed from the genotype dataset, leaving 36,513 SNPs for the population structure analysis.

Existing oil and protein content data made available by the USDA GRIN was used for the analysis. The phenotype data were originally obtained from field evaluations conducted by USDA-ARS germplasm curation staff and their collaborators. The field evaluations were conducted at various locations at which accessions from one or more MG classes had adaptation, and such field trials often spanned several years. Details and publication references relative to the methods used to quantify soybean seed protein and oil content are provided in GRIN (http://www.ars-grin.gov/cgi-bin/npgs/html/desc_form.pl?51; scroll down to oil and protein in list of descriptors). In brief, since 1990, accessions with yellow seed coats have been evaluated for protein and

oil concentrations using the near-infrared reflectance (NIR) method on a whole-seed sample. For those accessions with entirely pigmented or exceptionally mottled seed coats, seed protein was quantified with the Kjeldahl method and seed oil with the Butt extraction method. The Kjeldahl and Butt methods were also used in all pre-1990 evaluations. Using this existing phenotypic information, a total of 12,116 *G. max* accessions were identified that had seed oil and protein data (Supplementary Fig. S1). The phenotypic data used for the GWAS is summarized by accession number in Supplementary Table S3.

2.4.3 Population Structure

The model-based clustering algorithm of ADMIXTURE v1.22 was used (Alexander et al., 2009) to investigate subpopulation structure of the 14,430 soybean accessions. ADMIXTURE identifies K genetic clusters, where K is specified by the user, from the provided SNP data. For each individual, the ADMIXTURE method estimates the probability of membership to each cluster. A preliminary analysis was performed in multiple runs by inputting successive values of K from 3 to 20. This tested range of K inputs was based on the results of several studies that estimated the number of subpopulations (Hwang et al., 2014; Hyten et al., 2007; Sonah et al., 2015). A 10-fold cross-validation procedure was performed with 30 different fixed initial seeds for each K values. The most likely K value was determined using ADMIXTURE's cross-validation values. The software CLUMPP (Jakobsson and Rosenberg, 2007) was used to obtain the optimal alignments of 30 replicates for each K-value. Individual genotype membership proportions were averaged across runs according to the permutation with the greatest symmetric similarity coefficient. The output from CLUMPP for the optimal K was used to make plots using the cluster visualization program in R. To verify the proportion of correct

and incorrect classifications, we performed a linear discriminant analysis using the ancestry estimates for each accession with K=5. Principal-component analysis was also conducted to summarize the genetic structure and variation present in the collection. The hierarchical *F* statistics were used to estimate proportion of genetic variance explained by world region and MG class using ancestry estimates for K=5 and calculated using the *hierfstat* R package (Goudet, 2005).

2.4.4 Genome-wide Association Analysis

Marker-trait associations were tested using the linear mixed model

$$y = X\beta + C\gamma + Zu + e$$

where y is a vector of phenotypes; β is a vector of fixed marker effects; γ is a vector of subpopulation effects; u is a vector of polygenic effects caused by relatedness where $\mathbf{u} \sim MVN(0, \mathbf{K}\sigma_u^2)$; e is a vector of residuals where $\mathbf{e} \sim MVN(0, \mathbf{I}\sigma_e^2)$; X is a marker matrix; C is an incidence matrix containing membership proportions to each of the five genetic clusters identified by the ADMIXTURE analysis; and Z is the corresponding design matrix for u. K is the realized relationship matrix estimated internally in the Factored Spectrally Transformed Linear Mixed Models (FaST-LMM) using the SNP data (Lippert et al., 2011).

The above model was implemented using the FaST-LMM algorithm (Lippert et al., 2011). This program is designed to accommodate large datasets with reduced computational time (Lippert et al., 2011). FaST-LMM uses either maximum likelihood (ML) or restricted maximum likelihood (REML). Maximum likelihood (ML) was used for

this study because it has been found to be more reliable (Eu-ahsunthornwattana et al., 2014). FaST-LMM uses an exact method in which the additive genetic and residual variance components are re-estimated for each SNP in a model, including the marker effect, rather than being estimated under the null hypothesis.

Association analyses were conducted both across and within groups of accessions classified by either MG or world region class (Supplementary Tables S1 and S2). GWA mapping across all groups was conducted using SNP with MAF>0.01, and population structure was accounted for using both γ and u. GWA mapping within groups was performed using SNP with MAF>0.05; γ was ignored and only u was fitted as a random effect as described above.

The method of Li and Ji (2005) was used to calculate a comparison-wise error rate to control the experiment-wise error rate. Briefly, the correlation matrix and eigenvalue decomposition among 36,513 SNPs were calculated to determine effective number of independent tests (M_{eff}). The test criteria was then adjusted using the M_{eff} with the correction (Sidak, 1967) below

$$\alpha_p = 1 - (1- \alpha_e)^{1/Meff},$$

where α_p is the comparison-wise error rate and α_e is the experiment-wise error rate. An $\alpha_e = 0.05$ was used in this study.

Multiple linear regression was used to estimate the proportion of phenotypic variance accounted for by significant SNPs after accounting for population structure effects. Windows of 500 kb were used to define SNPs tagging a locus. Only the most significant SNPs present within a 500 kb window was used to tag that locus.

2.4.5 Haplotype Analysis

Haplotype blocks were constructed using the four gamete method (4gamete) (Wang et al., 2002) implemented in the software Haploview (Barrett et al., 2005). The 4gamete method creates block boundaries where there is evidence of recombination between adjacent SNPs based on the presence of all four gametic types. A cutoff of 1% was used, meaning that if addition of a SNP to a block resulted in a recombinant allele at a frequency exceeding 1%, then that SNP was not included in the block.

The frequencies of identified haplotype alleles were estimated in all accessions and within each subpopulation. At each haplotype block, a Fisher's exact test was used to test the null hypothesis that the frequency of the haplotype alleles for seed oil and protein content was the same across all subpopulations.

2.5 RESULTS AND DISCUSSION

2.5.1 The USDA Soybean Germplasm Collection

Of the 19,652 genotyped accessions in the publicly available SNP data set, 659 near-isogenic lines of multiple *G. max* cultivars were removed, but the recurrent and donor parents of those NILs were retained. An additional 4207 *G. max* and 362 *G. soja* accessions were SNP-genotype duplicates (i.e. 24%) and thus were removed. The high rate of redundant and highly similar accessions detected in the USDA Soybean Germplasm Collection is not surprising because genetic redundancy is a common problem in germplasm collections (Food and Agriculture Organization, 2010; McCouch et al., 2012). It is estimated that in world-wide collections, only one-third of the total number of accessions conserved *ex situ* are distinct and duplications occur within and between genebanks of the same crop (Food and Agriculture Organization, 2010; McCouch et al.,

2012). The major cause is the unwitting submission of the same accession with different names and designation. Based on phenotype alone, it is not possible to identify redundant accessions, yet the maintenance of duplicated materials invokes unnecessary and costly efforts. After eliminating all genotypic duplicates (one accession per set of duplicates was retained), a final set of 14,430 genotyped accessions (i.e., 13,624 *G. max* and 806 *G. soja* accessions) was available for population structure analysis (Supplementary Fig. S1).

Geographic origin and MG were the principal determinants of population structure within the soybean germplasm collection. Accessions collected from China (36%), North and South Korea (19%), Japan (17%), North and South America (9%), South and Southeast Asia (8%), Europe (5%), and Russia (5%) make up the vast majority of the collection (Supplementary Fig. S2a). Soybeans are classified into 13 unique MG Roman Numeral groups from very early to very late (000, 00, 0, I, II, III, IV, V, VI, VII, VIII, IX and X), based on temperature and photoperiod response to latitude. Maturity Group numbers of 000, 00, and 0 were combined, as were MG numbers VII, VIII, IX, and X , in order to reduce the 13 numbered MGs to just eight more manageable MG "classes" for use in the population structure analysis (Supplementary Fig. S2b). Maturity Groups II, III, IV and V represent over 50% of the entire collection.

2.5.2 Population Structure

Using ADMIXTURE (Alexander et al., 2009) and principal component analysis to infer population structure, we observed a clear subpopulation (SP) structure within the soybean germplasm collection. The total amount of genetic variation explained by the first three PCs was 16%. The first three principal components (PC) visually differentiate

accessions by species (*G. soja* vs *G. max*) and world region (Fig. 1). The estimated cross-validation (CV) error from ADMIXTURE and correspondent ΔCV values suggested the presence of four to eight natural SPs (K=4-8) within the collection (Supplemental Fig. S3). A K value of five was ultimately chosen because higher K values resulted in subpopulations in which no accessions belonged (based on an 80% membership cutoff). Accessions within a subpopulation with membership coefficients of <0.8 were considered admixed. The graphs of ancestry estimate for each accession for five subpopulations was plotted by world region and MG class (Fig. 2).

Subpopulation 1 represents the *G. soja* accessions which can be traced from China, Japan, Korea and Russia (Figs. 1, 2a; Supplementary Tables S4, S5). Subpopulation 1 is well represented in the MG classes. Notably, a single *G. max* accession with GRIN ID of PI 549045A falls into the *G. soja* SP. Although this accession has many wild soybean characteristics, it has a plant type very atypical of wild soybean, which caused it to be classified as *G. max.* Subpopulations 2 to 5 represent the *G. max* accessions (Figs. 2a, 2b; Supplementary Table S4), which consisted of a mixture of accessions from different world regions and MG classes (Supplementary Tables S5 and S6). Subpopulation 3 is composed predominantly of accessions of the late MG class (i.e., V, VI, and VII-X), which are mostly from China and SSE Asia, whereas SP5 contains a significant proportion of early MG accessions (i.e., 000-0, I and II), mainly originating from China and Far East Russia. Subpopulation 4 forms a unique SP comprised primarily of accessions from Japan (61%) and Korea (34%). Subpopulation 2 contains only 12 accessions from China and Europe, mainly from MG II plus a single accession from MG III. Subpopulation 2 had the highest mean ancestry shared to accessions from America (31%) and Europe (30%)

(Supplementary Table S7). Nearly two-thirds of the accessions in the soybean germplasm collection are admixed (Supplementary Tables S5 and S6). A large portion of admixed accessions generally traced to China, Korea, America and Europe. Notably, more than 90% of accessions from America and Europe are admixed. The proportion of individuals for each world region and MG within each of the five subpopulations was not equal, indicating different degrees of allelic diversity across subpopulations. Japanese accessions were more homogenous and mostly belonged to SP4. Similarly, the overrepresentation of accessions from some MG classes in SP3 and SP5 was due to the sensitivity of soybean to photoperiod and temperature, which restricts adaptability to compatible regions of latitude.

Another interesting result from the population structure analysis is the relationship between China, SSE Asia, Japan, and Korea accessions. Accessions from Japan and Korea were more closely related to each other than with accessions from China. Accessions from Japan form a unique subpopulation (SP4), whereas those from Korea consist of mostly an admixture between SP4 and SP5 (Fig. 2a), possibly reflecting the migration of soybean from northeast China to Korea. The homogeneous subpopulation structure among the Japan accessions could be due to the isolation of Japan by sea from China and Korea and the fact that Japan was never conquered by China (Hall, 1988). The earlier MG classes appear to contain more SP5 ancestry (Fig. 2b), reflecting the latitude of where North Korea is connected to China relative to ancient soybean trade routes. Subpopulation 3 is clearly a significant fraction in China and SSE Asia, suggesting a substantial movement of *G. max* accessions into nearby Asian countries.

Based on population structure results, we then evaluated the genetic relationships of the major ancestors of American soybean (Gizlice et al., 1994; Li and Nelson, 2001;

Ude et al., 2003) to accessions within the collection. Two of 34 USA soybean ancestors, the well-known ancestors A.K. Harrow and Illini, were found to be completely SNP identical. The majority of American accession ancestry belong to SP2 and SP5 (59%), whereas only 44% of China accessions ancestry belong to these SPs (Supplementary Table S7). As expected, the ancestors of American soybean, which originated from China, mostly share ancestry with SP2 and SP5 (49%) (Fig. 3 and Supplementary Table S7) reflecting the ancestry of American germplasm as a whole. This analysis is complicated by the fact that ancestors of American soybean germplasm contributed at different pedigree levels (Fig. 3) (Gizlice et al., 1994; Li and Nelson, 2001; Ude et al., 2003), coupled with the fact that the American soybean germplasm resulted from a severe population bottleneck when soybeans were introduced to North America (Hyten et al., 2006).

Hierarchical F statistics, calculated using ancestry estimates for K=5, showed that genetic differentiation explained by world region (~10%) was higher than that explained by MG (~5%). The results of discriminant analysis supports this finding with the overall correct classification being greater for geographic origin (53%) than for MG class (35%) based on the GRIN classifications. Although the amount of total variation explained is small, these results suggest that population structure in the USDA Germplasm Collection is driven more by world region than MG.

2.5.3 Genome-wide Association Study for Protein and Oil

A GWA study was conducted solely on the *G. max* accessions to avoid confounding effects of strong subpopulation structure that would arise by combining the *G. max* and the *G. soja* accessions. Genotype data and seed oil and protein content data were available for 12,116

G. max accessions (Supplementary Fig. S1). Phenotype data were obtained from the GRIN soybean database (Supplementary Table S3). As expected, protein had significant negative correlation to oil (r = -0.62; $p<0.0001$) across world region and MG. The same negative relationship was observed for oil *versus* protein within world region or MG classes.

A GWAS was performed using all 12,116 *G. max* accessions (Fig.4) and then for subsets of accessions based on world region and MG class (Supplementary Figs. S4a, S4b and S5a, S5b). A total of 19 significant associations (-logP >5. 17) were identified for protein (Fig. 4 and Table 1). Clusters of highly significant markers were present on chromosomes 15 (3.82-3.96Mb) and 20 (29.59-31.97Mb), which collectively explained 7% of the phenotypic variance for protein. GWAS for oil detected 18 significantly associated SNP markers (-logP>5. 17), with the strongest association detected at 3.82Mb on chromosome 15 (Fig. 4 and Table 1). Collectively, the three QTL identified for oil explained 6% of the phenotypic variance. The major associations detected on chromosomes 15 (LG-E) and 20 (LG-I) were highly significant for both protein and oil and had map positions close to those identified for the corresponding LGs identified in the very first soybean QTL mapping experiment (Diers et al., 1992). These two QTL have been detected repeatedly in linkage mapping studies since then (http://www.soybase.org/search/qtllist_by_symbol.php). Allelic effect estimates of SNP markers showed negative genic relationship between protein and oil content (Table 1), i.e., the SNP allele associated with increased protein content was also always associated with decreased oil content and *vice versa*.

A GWAS within the world region identified at least one or both of the strongest regions on chromosomes 15 and 20 for protein and oil, except within the America and SSE

Asia accession subsets (Supplementary Figs. S4a, S4b and S5a, S5b). The strongest associations for protein (-logP=14.52) and oil (-logP=14.22) among world region classes were in Korea (Supplementary Figs. S4a and S4b). Similarly, GWAS results based on different MG classes detected the same genomic regions for protein and oil on either chromosomes 15 and 20, except in the cases of the MG I class or MG II class for protein and for the MG IV class for oil (Supplementary Figs. S5a and S5b). The strongest associations for both traits were identified in the late MG classes of V (-logP for oil=13.95) and VII-X (-logP for protein=19.83). The chromosomal bp resolution of the identified QTL on chromosomes 15 and 20 varied among world region and MG classes. For example, on chromosome 20, the resolution of associated SNPs for protein and oil spanned only 8 Mb region between 26 – 32 Mb, which corresponds to the narrowed region defined by Bolon et al. (2010). The highest resolution was achieved when all 12K accessions were combined for GWAS, demonstrating the advantage of exploiting diversity and thus greater historical recombination for increased resolution.

In the recent GWA analysis, Vaughn et al. (2014) used accessions almost exclusively from Korea and which were mostly of MG V. The GWA results in this study using accessions from Korea and MG V had the associated SNPs for protein identified between 29.20 – 31.97 Mb, with the strongest associated SNP similar to the one identified by Vaughn et al. (2014) at 31,972,955 bp on chromosome 20. However, the association at 31,972,955 bp is entirely not detected when the GWA analysis was limited to other world region and MG classes. Rather, an association was detected at an even higher level of significance for protein using late MG class VII-X at 31.24 Mb (-logP=19.83) on chromosome 20, which is consistent to the most significant SNP detected using all 12,116

G. max accessions. Similarly, the significant SNP reported by Hwang et al. (2014) between 29.39–29.98 Mb on chromosome 20 was missed among MGs, but that SNP was identified in MG classes I, III and IV if –logP threshold was lowered to 4.0. Although the associations were detected when the threshold stringency was reduced, the association threshold of –logP=4 is still more stringent than the -logP=3 that Hwang et al. (2014) used in GWA analysis using 298 *G. max* accessions from MGs II, III and IV. The varying results of GWAS within subpopulation classes can be due to differing levels of recombination, diversity as well as rate of LD decay across the assembled panels. Notably, as mentioned above, the highest resolution was achieved when the GWA analysis was conducted using all *G. max* accessions.

2.5.4 Refining the Candidate Region for Protein and Oil

The levels of significance for protein and oil QTL were quite high in this study and thus could be used to delineate a narrower bp region for the identification of candidate genes. This was clearly illustrated in the region detected on chromosome 20 for both protein and oil content, which co-located with a previously reported QTL responsible for major pleiotropic effects on protein and oil content (Bolon et al., 2010; Hwang et al., 2014; Nichols et al., 2006). A previous study showed that the QTL resides in an 8.4 Mbp region located between 24.5-32.9 Mb on chromosome 20 (Bolon et al., 2010). Subsequently, that region was narrowed further to a 2.4 Mb region located between 27.6-30.0 Mb (Hwang et al., 2014). Vaughn et al. (2014), however, mapped this same QTL approximately 1 Mb downstream of the region that Hwang et al (2014) identified, with the most significant SNP identified at 31,972,955 bp. To further refine this candidate region, a haplotype analysis

was performed in this study covering the genomic regions identified by Hwang et al. (2014) and Vaughn et al. (2014) that spans a circa 4.5 Mb region. Within the entire collection, LD in that 4.5 Mb region decays from r^2=0.50 to r^2=0.40 within 250 Kb, then decays more slowly to r^2=0.20 from 500 Kb to 1000 Kb. A haplotype block analysis of this 4.5 Mb region identified five haplotype blocks (Fig. 5), with highly significant SNPs located in the third, fourth and fifth blocks. The fourth and fifth blocks contain significant SNPs associated for both oil and protein while the third block includes the SNP associated with protein only (Fig. 5; Table 1).

The third block contains the region defined by Hwang et al. (2014) which is further narrowed down to less than 1 Mb region in this study and spans between 29.06–30.04 Mb. This region now encompasses only three (Glyma20g21030, Glyma20g21040 and Glyma20g21080) of the original 12 potential candidate genes (Fig.5a) (Bolon et al., 2010). Glyma20g21040 has no known function, while Glyma20g21030 is annotated as a putative ammonium transporter (AMT1), which catalyzes the transfer of ammonium from one side of a membrane to the other (Sohlenkamp et al., 2002). Although the AMT1 annotation reflects a partial sequence homology to the Arabidopsis gene, it may be that this gene in *G. max* was recruited for the transport of N from the female plant integument tissue (i.e., seed coat) to the developing embryo. Glyma20g21080 is another candidate gene within the third block region which is annotated as ATP synthase D chain. An indirect relationship of ATP synthase levels on the accumulation of storage proteins, lipid biosynthesis, and photosynthesis in the seed have been documented (Borisjuk et al., 2003; Borisjuk et al., 2005; Rolletschek et al., 2003; Rolletschek et al., 2005). However, the potential role of

ATP synthase as a candidate gene on energy status and accumulation of storage product in the soybean seed needs further elucidation (Bolon et al., 2010).

The fourth block spans 550 Kb between 30.38-30.93; LD decays within 250 kb with r^2=0.24 on average. This region is the focus of another group who has fine-mapped the location of high protein QTL allele present in high protein *G. soja* PI 468916 (Brian Diers, Pers. Commun.). Despite the fact that this allele comes from *G. soja* progenitor, this *G. soja* allele has likely been dispersed throughout the soybean germplasm (Diers et al., 1992; Sebolt et al., 2000), given the discovery of this allele in high-/low-protein *G. max* accessions (Chung et al., 2003; Fasoula et al., 2004; Wilcox and Cavins, 1995; Wilcox, 1998). The most plausible candidate gene within the fourth block region is the Glyma20g21361 that has been annotated as Conserved Oligomeric Golgi Complex (subunit 6), which is involved in the intra- and inter-cellular vesicle-mediated transfer and storage of proteins. Glyma20g21540 is another gene within this fourth block region which encodes protein of unknown function (i.e., annotated as putative uncharacterized proteins).

A cluster of eight highly significant SNPs associated for protein and oil was located in the fifth block that spans 900 Kb region between 31.15-32.05 Mb (Figure 5; Table 1). The most significant SNP (BARC_1.01_Gm_20_31243150_C_T) for protein identified in the present study was in complete LD with the strongest SNP (BARC_1.01_Gm_20_31972955_G_A) associated for protein detected by Vaughn et al. (2014). The LD decay within this region was calculated to be 400 kb on average with r^2=0.32 and a candidate gene was identified within the target region. Only one of the annotated genes (Glyma20g21780) identified by Bolon et al. (2010) was located within the 900 Kb defined region in the fifth block. The gene Glyma20g21780 located at 31.38 Mb

encodes an ethylene receptor and was demonstrated to be involved in signal transduction and protein histidine kinase activity (Bolon et al., 2010; www.soybase.org).

On chromosome 15, a significant region was detected for protein and oil across all accessions with a high resolution of less than 150 Kb between 3.82-3.96 Mb. The LD decay within this region was estimated at 165 Kb with r^2=0.23 on average. This detected QTL co-localizes with the three candidate genes (Glyma15g05470, Glyma15g05760 and Glyma15g05770), which could be related to seed protein and oil levels. Glyma15G05470 encodes a RAG1-activating protein which is a soybean ortholog of the Arabidopsis Nodulin MtN3 family protein involved in sucrose transmembrane transporter activity. Glyma15g05760 encodes for sulfate transporter, where in soybean, sulfate accumulates in pods and decreases with the onset of grain enlargement (Sunarpi, 1996; Tabe and Droux, 2001). Glyma15g05770 has a protease inhibitor activity and is involved in lipid transfer for seed storage protein (Wang et al., 2007). The synthesis of storage products during seed development is coordinated with carbohydrate and nitrogen metabolic processes involving many transporters, including ammonium and sulfate transporters (Weber et al., 1998).

Haplotype alleles on chromosome 20 displayed a negative relationship between the protein and oil content (Table 2). Given the higher mapping resolution and presence of few likely candidate genes in the detected genomic regions, we hypothesize pleiotropic gene effects underlie the observed negative correlation between oil and protein at these loci. Nevertheless, the possibility of multiple genes in very tight repulsion-phase linkage cannot be excluded because the sizes of the haplotype blocks in the chromosome 20 region are still greater than 500 Kb. It is interesting to note that the haplotypes with the lowest average protein, and conversely the highest average oil (H1, H4 and H7), are the predominant

haplotypes, especially among Ancestors, America, Japan, and Korea accessions (Table 2). Haplotype H7 in block B5, conferring high protein, is very rare among germplasm accessions collection in general, but is relatively less rare among the Korea accessions, perhaps due to selection on protein because of cultural and culinary preferences in that country. The two haplotype alleles in block B5 have larger difference in protein content (3.82%) compared to the haplotype alleles at block B3 (1.27%). Based on a country-specific GWAS Manhattan plots, it appears this haplotype among the Korea accessions is providing the variation leading to the SNP-protein associations we detected on chromosome 20 (Supplementary Fig. S4a).

This study also demonstrates the potential usefulness of extensive phenotypic data that has been collected on germplasm collections, but which has not yet been fully utilized to mine favorable alleles not present in the narrow germplasm pool currently underpinning modern soybean improvement. By maximizing the number of accessions included in the GWAS, we were able to exploit a greater number of historical recombinants, leading to increased power and resolution. A potential limitation in the analysis conducted here is the nature of the phenotype data used for GWAS. The seed protein and oil data obtained from GRIN were derived from multiple field evaluations of the USDA Soybean Germplasm Collection conducted over the past 50 years at various latitude-specific locations. To assess the reliability of the GRIN phenotypic data, we conducted another GWAS for oil and protein content that was limited to 9,861 unique *G. max* accessions for which raw data was available from the GRIN database. A linear model accounting for environmental effects was applied and adjusted phenotypic values were input into the GWAS model. We detected only very small differences between GWAS results using adjusted data (Supplementary

Fig. S6) and GWAS results using non-adjusted data (Fig. 4). The only exception was a few chromosomal positions in which the SNP associations straddled the statistical significance threshold.

Despite the success of the 12K GWAS in narrowing down the major candidate region, it was surprising how few genomic regions were found to be associated with protein and oil given the large number of QTL reported for these traits using biparental linkage mapping populations (www.soybase.org). Only six regions were identified for both protein and oil, collectively explaining only 6-7 percent of the phenotypic variation. Many more associations were expected given the size of the panel used for the GWAS. These results agree with Vaughn et al. (2014) who used a similar population and phenotypic dataset, albeit it smaller and less genetically diverse. Hwang et al. (2014), using a much smaller panel size, identified many more associations using a more relaxed significance threshold. If we relaxed the threshold to that used by Hwang et al. (2014) the number of associations for protein and oil would be increased by four fold, with associations on 14 chromosomes. The lack of associations passing the more stringent statistical threshold used in this study might be related to the distribution of allelic effects and frequencies in the collection. It is entirely possible that the genetic variation for these traits is controlled by a multitude of rare or low-frequency alleles, which are difficult to identify in GWAS. The QTL on chromosome 20 may be an indication of such a genetic architecture. It can be seen that the haplotype alleles at blocks 4 and 5 that increase protein (and lower oil) are of very low frequency in the collection (Table 2). It is unlikely this QTL would be detected were it not for its relatively large effect. Other possible reasons for the lack of QTL detection include confounding of QTL allele frequencies with population structure (Rincent et al., 2014). If

frequency of alleles affecting protein and oil are confounded with population structure, then correcting for population structure using the mixed linear model would reduce detection power. Haplotype allele frequencies do differ between world regions at the chromosome 20 QTL. Finally, epistasis is always a possible cause. Determining the relative contribution of all possible causes to this "missing heritability" problem of soybean oil and protein was outside the scope of this research but it certainly deserves further study.

The wealth of phenotypic diversity available in the soybean germplasm collection should be mined to help meet the demands of food production in the face of climate change and ever-evolving pathogens. The results reported herein, and others surely to flow from this valuable resource, provide a fuller understanding of the distribution of genetic variation contained within the collection and its relation to phenotypic variation for economically important traits. Further characterization of the phenotypic diversity and its relationship to the genomic diversity will ultimately facilitate a more efficient and effective introgression of this diversity into elite varieties for continued genetic improvement.

2.6 REFERENCES

Alexander, D.H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655-1664.

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

Bolon, Y.T., B. Joseph, S.B. Cannon, M.A. Graham, B.W. Diers, A.D. Farmer, G.D. May, G.J. Muehlbauer, J.E. Specht, Z.J. Tu, N. Weeks, W.W. Xu, R.C. Shoemaker, and C.P. Vance. 2010. Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC Plant. Biol. 10:41-2229-10-41.

Borisjuk, L., H. Rolletschek, S. Walenta, R. Panitz, U. Wobus, and H. Weber. 2003. Energy status and its control on embryogenesis of legumes: ATP distribution within vicia faba embryos is developmentally regulated and correlated with photosynthetic capacity. Plant Journal 36:318-329.

Borisjuk, L., T.H. Nguyen, T. Neuberger, T. Rutten, H. Tschiersch, B. Claus, I. Feussner, A.G. Webb, P. Jakob, H. Weber, U. Wobus, and H. Rolletschek. 2005. Gradients of lipid storage, photosynthesis and plastid differentiation in developing soybean seeds. New Phytol. 167:761-776.

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

Burton, J.W. 1987. Quantitative genetics results relevant to soybean breeding. Wilcox, J.R.(Ed.).Agronomy: A Series of Monographs, no.16.Soybeans: Improvement, Production, and Uses, Second Edition.Xxii+888p.American Society of Agronomy, Inc., Crop Science Society of America, Inc., Soil Science Society of America, Inc.Publi(TRUNCATED)211-248.

Cavanagh, C.R., S. Chao, S. Wang, B.E. Huang, S. Stephen, S. Kiani, K. Forrest, C. Saintenac, G.L. Brown-Guedira, A. Akhunova, D. See, G. Bai, M. Pumphrey, L. Tomar, D. Wong, S. Kong, M. Reynolds, M.L. da Silva, H. Bockelman, L. Talbert, J.A. Anderson, S. Dreisigacker, S. Baenziger, A. Carter, V. Korzun, P.L. Morrell, J. Dubcovsky, M.K. Morell, M.E. Sorrells, M.J. Hayden, and E. Akhunov. 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc. Natl. Acad. Sci. U. S. A. 110:8057-8062.

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

Clemente, T.E., and E.B. Cahoon. 2009. Soybean oil: Genetic approaches for modification of functionality and total content. Plant Physiol. 151:1030-1040.

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

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

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

Eu-ahsunthornwattana, J., E.N. Miller, M. Fakiola, S.M.B. Jeronimo, J.M. Blackwell, H.J. Cordell, and Wellcome Trust Case Control. 2014. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. Plos Genetics 10:e1004445.

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

Food and Agriculture Organization. 2010. The second report on the state of the world's plant genetic resources for food and agriculture.

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

Goudet, J. 2005. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes 5:184-186.

Hall, J.W. 1988. The cambridge history of japan. Cambridge University Press, Cambridge Cambridgeshire ; New York.

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, C. Li, C. Zhu, T. Lu, Z. Zhang, M. Li, D. Fan, Y. Guo, A. Wang, L. Wang, L. Deng, W. Li, Y. Lu, Q. Weng, K. Liu, T. Huang, T. Zhou, Y. Jing, W. Li, Z. Lin, E.S. Buckler, Q. Qian, Q.F. Zhang, J. Li, and B. Han. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961-967.

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

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

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

Jakobsson, M., and N.A. Rosenberg. 2007. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801-1806.

Korte, A., and A. Farlow. 2013. The advantages and limitations of trait analysis with GWAS: A review. Plant Methods 9:(22 July 2013)-(22 July 2013).

Li, J., and L. Ji. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb) 95:221-227.

Li, Z.L., and R.L. Nelson. 2001. Genetic diversity among soybean accessions from three countries measured by RAPDs. Crop Sci. 41:1337-1347.

Lippert, C., J. Listgarten, Y. Liu, C.M. Kadie, R.I. Davidson, and D. Heckerman. 2011. FaST linear mixed models for genome-wide association studies. Nature Methods 8:833-U94.

McCouch, S.R., K.L. McNally, W. Wang, and R.S. Hamilton. 2012. Genomics of gene banks: A case study in rice. Am. J. Bot. 99:407-423.

Munoz-Amatriain, M., A. Cuesta-Marcos, J.B. Endelman, J. Comadran, J.M. Bonman, H.E. Bockelman, S. Chao, J. Russell, R. Waugh, P.M. Hayes, and G.J. Muehlbauer. 2014. The USDA barley core collection: Genetic diversity, population structure, and potential for genome-wide association studies. Plos One 9:e94688.

Myles, S., J. Peiffer, P.J. Brown, E.S. Ersoz, Z. Zhang, D.E. Costich, and E.S. Buckler. 2009. Association mapping: Critical considerations shift from genotyping to experimental design. Plant Cell 21:2194-2202.

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

Rincent, R, L. Moreau, H. Monod, E. Kuhn, A. Melchinger, R.A. Malvar, J. Moreno-Gonzalez, S. Nicolas, D. Madur, V. Combes, F. Dumas, T. Altmann, D. Brunel, M. Ouzunova, P. Flament, P. Dubreuil, A. Charcosset, and T. Mary-Huard. 2014. Recovering power in association mapping panels with variable levels of linkage disequilibrium. Genetics 197: 375-387.

Rolletschek, H., H. Weber, and L. Borisjuk. 2003. Energy status and its control on embryogenesis of legumes. embryo photosynthesis contributes to oxygen supply and is coupled to biosynthetic fluxes. Plant Physiol. 132:1196-1206.

Rolletschek, H., R. Radchuk, C. Klukas, F. Schreiber, U. Wobus, and L. Borisjuk. 2005. Evidence of a key role for photosynthetic oxygen release in oil storage in developing soybean seeds. New Phytol. 167:777-786.

Romay, M.C., M.J. Millard, J.C. Glaubitz, J.A. Peiffer, K.L. Swarts, T.M. Casstevens, R.J. Elshire, C.B. Acharya, S.E. Mitchell, S.A. Flint-Garcia, M.D. McMullen, J.B. Holland, E.S. Buckler, and C.A. Gardner. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14:R55-2013-14-6-r55.

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

Sidak, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of American Statistical Association 62:626-633.

Sohlenkamp, C., C.C. Wood, G.W. Roeb, and M.K. Udvardi. 2002. Characterization of arabidopsis AtAMT2, a high-affinity ammonium transporter of the plasma membrane. Plant Physiol. 130:1788-1796.

Sonah, H., L. O'Donoughue, E. Cober, I. Rajcan, and F. Belzile. 2015. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant. Biotechnol. J. 13:211-221.

Sunarpi, J.W.A. 1996. Effect of sulfur nutrition on the redistribution of sulfur in vegetative soybean plants. Plant Physiology (Rockville) 112:623-631.

Tabe, L.M., and M. Droux. 2001. Sulfur assimilation in developing lupin cotyledons could contribute significantly to the accumulation of organic sulfur reserves in the seed. Plant Physiol. 126:176-187.

Ude, G.N., W.J. Kenworthy, J.M. Costa, P.B. Cregan, and J. Alvernaz. 2003. Genetic diversity of soybean cultivars from china, japan, north america, and north american ancestral lines determined by amplified fragment length polymorphism. Crop Sci. 43:1858-1867.

Vaughn, J.N., R.L. Nelson, Q. Song, P.B. Cregan, and Z. Li. 2014. The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. G3-Genes Genomes Genetics 4:2283-2294.

Wang, H., B. Zhang, Y. Hao, J. Huang, A. Tian, Y. Liao, J. Zhang, and S. Chen. 2007. The soybean dof-type transcription factor genes, GmDof4 and GmDof11, enhance lipid content in the seeds of transgenic arabidopsis plants. Plant Journal 52:716-729.

Wang, N., J.M. Akey, K. Zhang, R. Chakraborty, and L. Jin. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. Am. J. Hum. Genet. 71:1227-1234.

Weber, H., S. Golombek, U. Heim, L. Borisjuk, R. Panitz, R. Manteuffel, and U. Wobus. 1998. Integration of carbohydrate and nitrogen metabolism during legume seed development: Implications for storage product synthesis. J. Plant Physiol. 152:641-648.

Wilcox, J.R. 1998. Increasing seed protein in soybean with eight cycles of recurrent selection. Crop Sci. 38:1536-1540.

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

Wilson, R.F. 2008. Soybean: Market driven research needs. In: G.Stacey, editor, Genetics and genomics of soybean. Springer Science+Business Media, New York. p. 3–15.

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

Zhu, C., M. Gore, E.S. Buckler, and J. Yu. 2008. Status and prospects of association mapping in plants. Plant Genome 1:5-20.

2.7 FIGURES

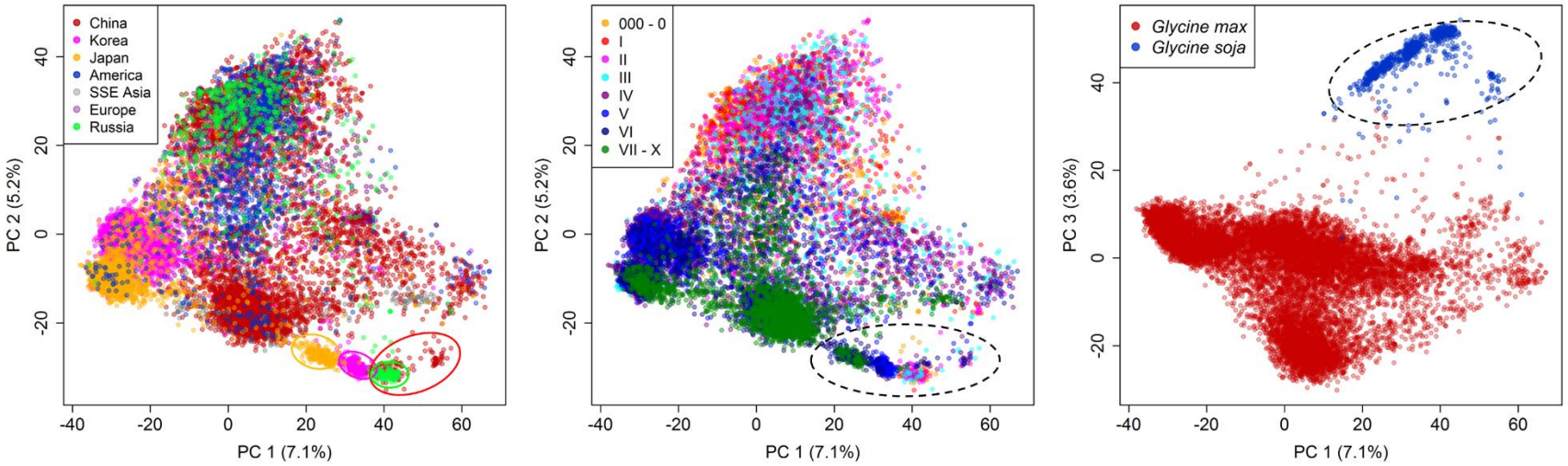


Figure 1. Principal component analysis of 14,430 accessions of the soybean germplasm collection. The *G. soja* accessions are demarked with colored circles in the world region panel and a dashed ellipse in the other two panels.

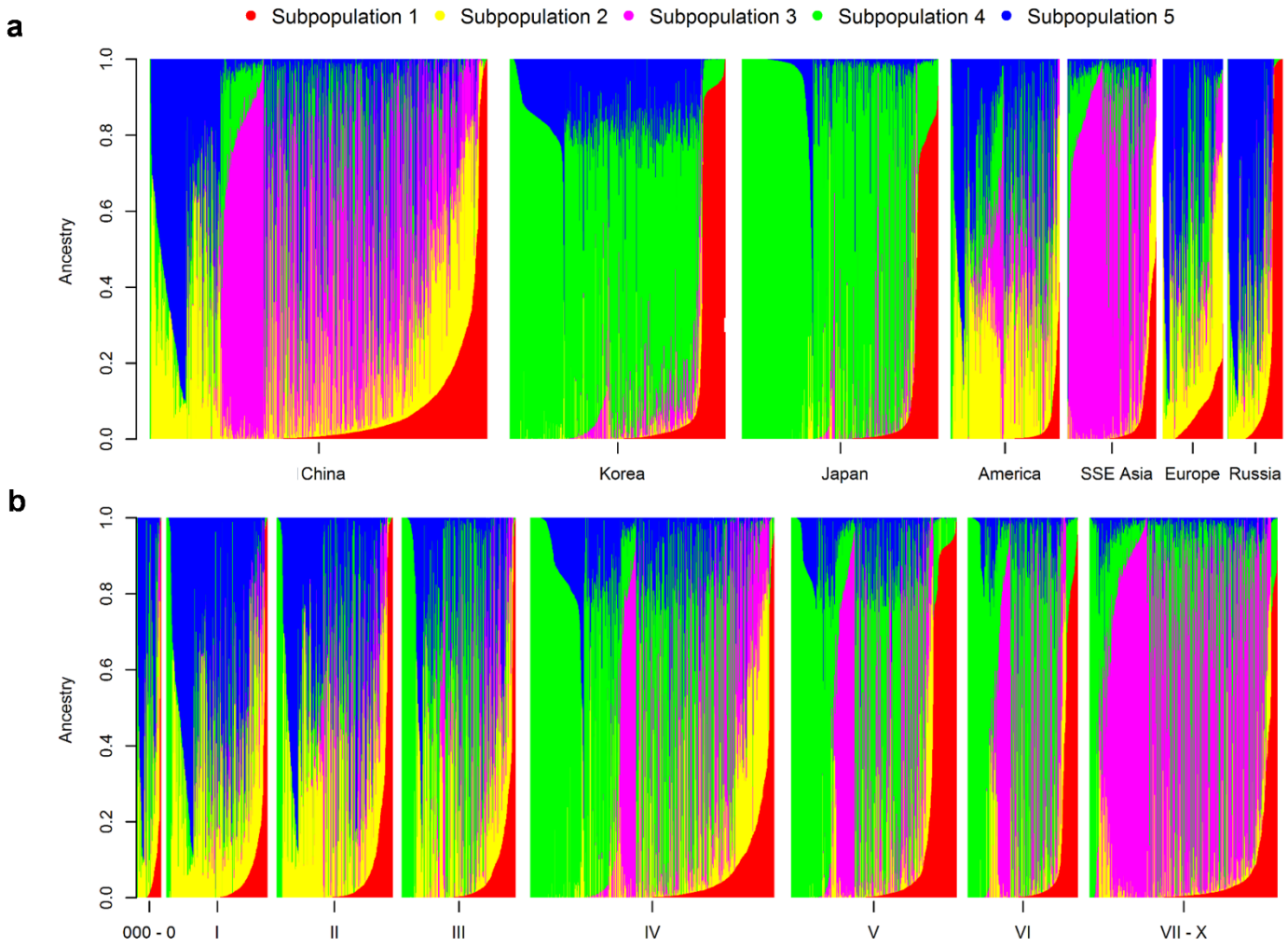


Figure 2. Population structure in the soybean germplasm collection inferred by ADMIXTURE. The number of clusters (K) present in the entire population of 14,430 accessions was judged to be K=5. Each colored vertical line in the world region (panel a) or MG class (panel b) represents an individual accession that was assigned proportionally to the one of the five clusters.

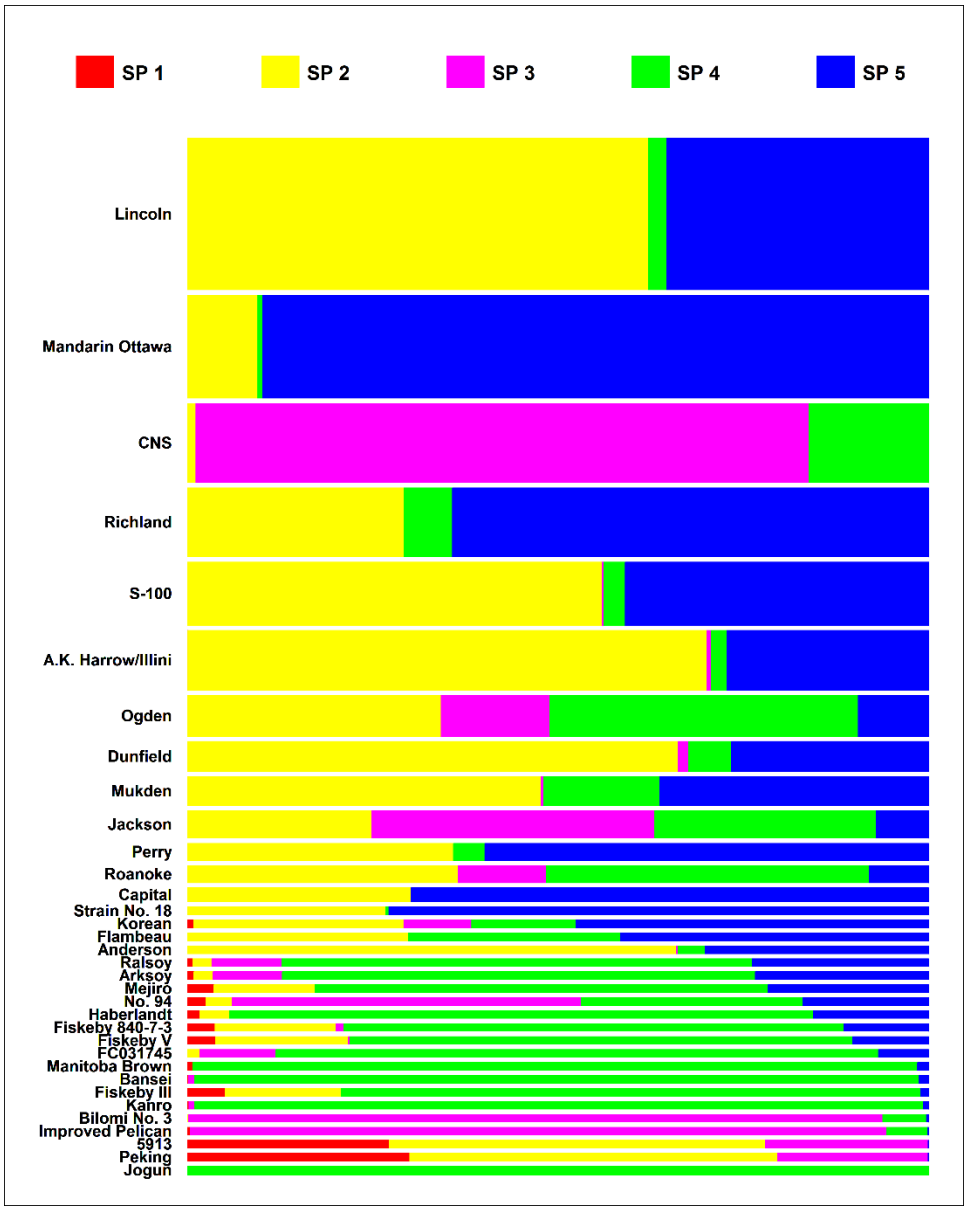


Figure 3. Plot of ancestry estimates inferred by ADMIXTURE for 34 USA soybean ancestors. Each colored vertical bar represents an ancestral accession that was assigned proportionally to the K clusters (K=5) with the proportions represented by the relative lengths of the K colors. The bar width reflects the percentage (>2%) contribution of the major ancestors according to (Gizlice et al., 1994; Li and Nelson, 2001; Ude et al., 2003).

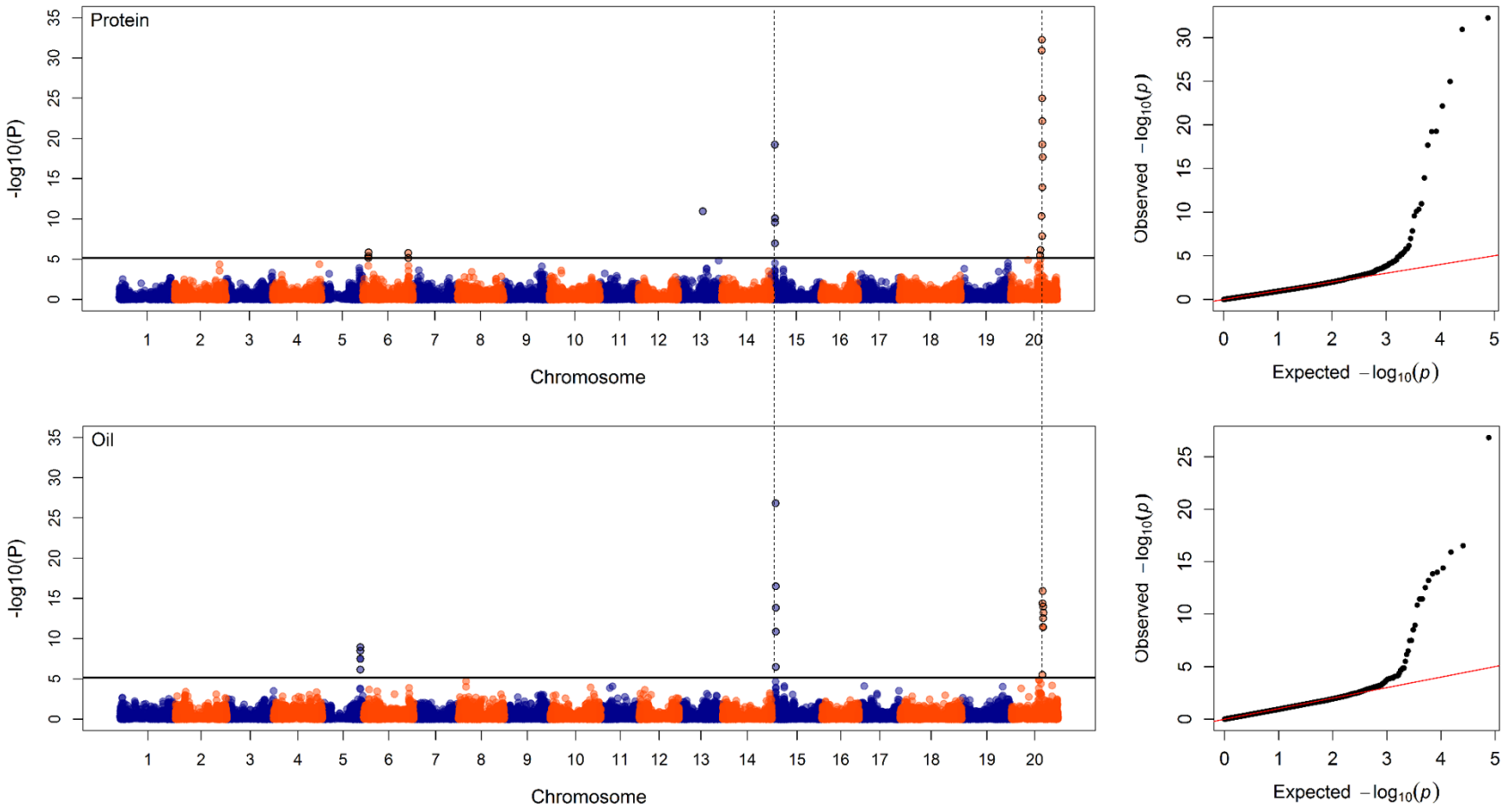
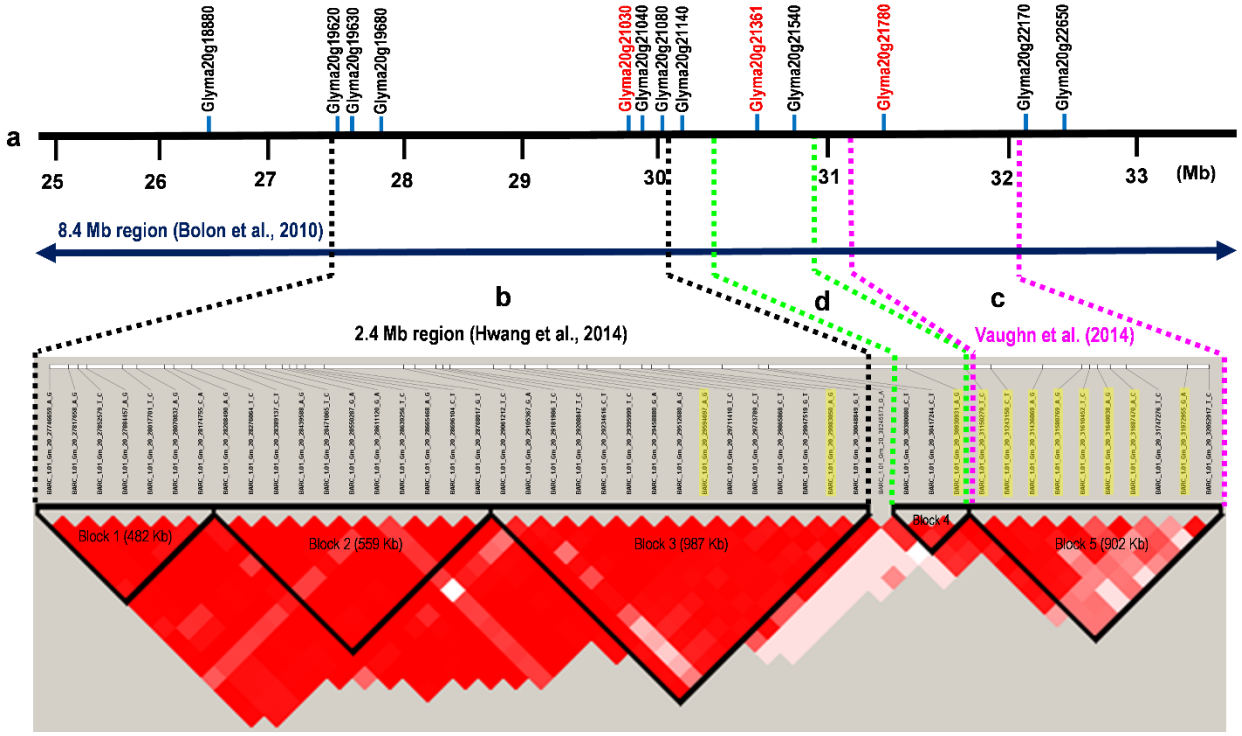


Figure 4. Genome-wide association scans for 12,116 *G. max* accessions for seed protein and oil content. Manhattan plots show the associations for seed protein and oil with SNP markers that are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line denotes the calculated threshold value for declaring significant association. The dashed vertical lines indicate a significant association for both seed protein and oil content had a co-localized chromosomal position.



\

Figure 5. Haplotype analysis of the chromosome 20 region that harbors a major pleiotropic seed protein/oil QTL. The figure shows (a) the 8.4 Mb region (navy blue line capped with arrows) defined by Bolon et al. (2010), (b) a recent narrowing of that region by Hwang et al. (2014) to 2.4 Mb region (black dashed lines), and (c) a 900 Kb region (magenta-dashed lines) identified by Vaughn et al. (2014) that is located to the right of the Hwang et al. (2014) region, and (d) a 550 Kb (green dashed lines) between the two foregoing regions. Potential candidate genes in the region are listed at the top of the figure (i.e., Glyma names), with the red-font ones being the most plausible candidate genes). Using the 4-gamete rule in the present study, five blocks were defined within that 4.5 Mbp region comprising the three sub-segments. Statistically significant SNPs are highlighted in yellow.

2.8 TABLES

Table 1. Allelic effects estimates of SNP markers significantly associated (-logP > 5.17) for seed protein and oil content QTL. The SNP markers are listed by chromosome number.

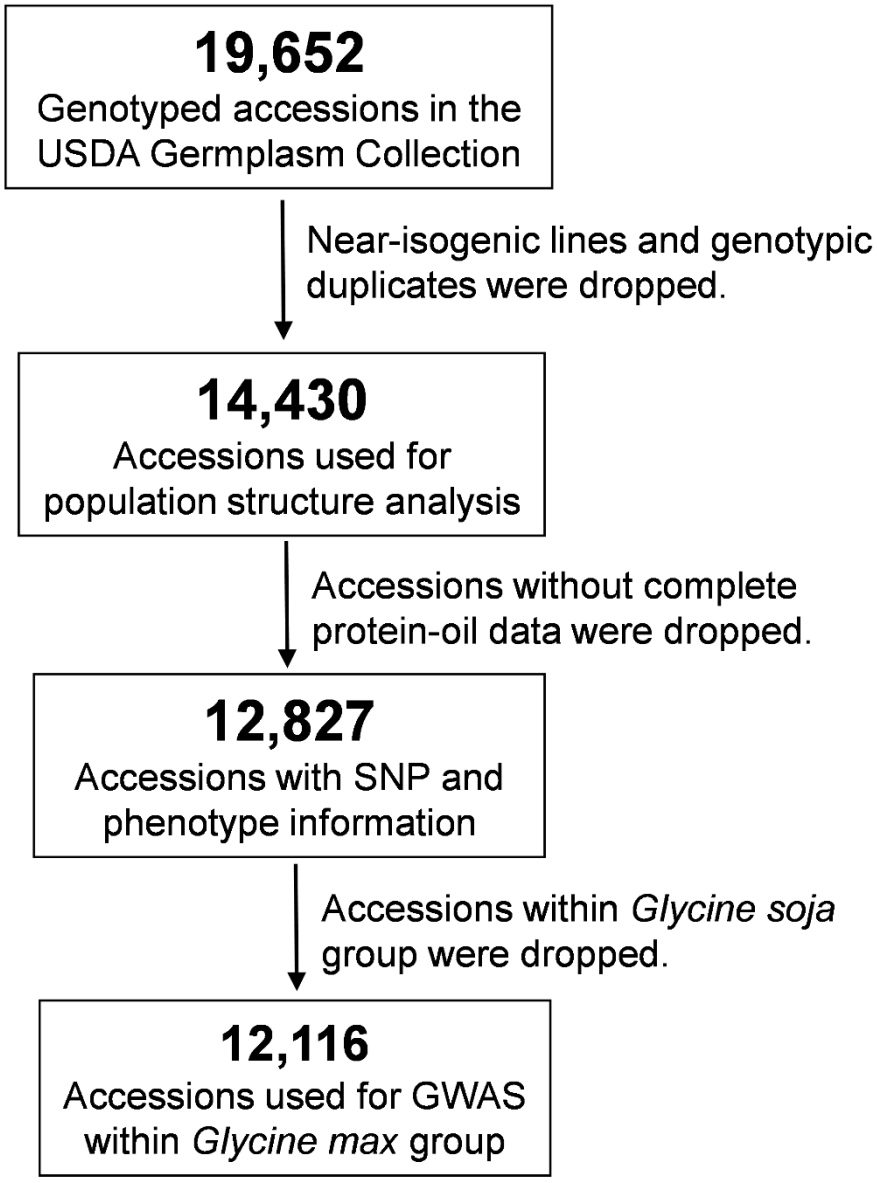
		Seed Protein Content		Seed Oil Content	
SNP	Chromosome	-logP	Allelic Effect Estimate (%)	-logP	Allelic Effect Estimate (%)
BARC_1.01_Gm_20_29594697_A_G	20	5.49	0.22	ns	ns
BARC_1.01_Gm_20_29983050_A_G	20	6.16	0.23	ns	ns
BARC_1.01_Gm_20_30930931_A_G	20	10.35	0.20	5.51	-0.09
BARC_1.01_Gm_20_31150279_T_C	20	30.94	0.40	14.39	-0.18
BARC_1.01_Gm_20_31243150_C_T	20	32.28	0.40	15.91	-0.18
BARC_1.01_Gm_20_31436069_A_G	20	7.87	0.17	ns	ns
BARC_1.01_Gm_20_31580769_A_G	20	24.98	0.36	11.45	-0.16
BARC_1.01_Gm_20_31610452_T_C	20	22.16	0.31	13.99	-0.16
BARC_1.01_Gm_20_31640038_A_G	20	13.94	0.24	11.43	-0.14
BARC_1.01_Gm_20_31687470_A_C	20	19.25	0.30	12.53	-0.16
BARC_1.01_Gm_20_31972955_G_A	20	17.68	0.26	13.22	-0.14
BARC_1.01_Gm_15_3828587_A_G	15	19.23	0.27	26.83	-0.21
BARC_1.01_Gm_15_3833574_A_G	15	ns	ns	6.49	0.10
BARC_1.01_Gm_15_3918803_A_C	15	9.57	0.20	13.84	-0.16
BARC_1.01_Gm_15_3919945_G_A	15	10.10	0.21	16.52	-0.17
BARC_1.01_Gm_15_3967324_A_G	15	6.97	0.17	10.87	-0.14
BARC_1.01_Gm_13_24858209_A_G	13	10.95	-0.20	ns	ns
BARC_1.01_Gm_06_5591484_T_C	6	5.37	0.13	ns	ns
BARC_1.01_Gm_06_5660542_A_G	6	5.84	0.15	ns	ns
BARC_1.01_Gm_06_46040638_C_T	6	5.78	0.17	ns	ns
BARC_1.01_Gm_05_38495217_A_C	5	ns	ns	8.50	-0.17
BARC_1.01_Gm_05_38495666_C_T	5	ns	ns	7.50	-0.15
BARC_1.01_Gm_05_38519280_G_A	5	ns	ns	6.16	-0.14
BARC_1.01_Gm_05_38543317_T_C	5	ns	ns	7.49	-0.15
BARC_1.01_Gm_05_38569452_T_G	5	ns	ns	8.95	-0.16

ns - SNP marker is not signicantly associated for a trait using -logP threshold of 5.17.

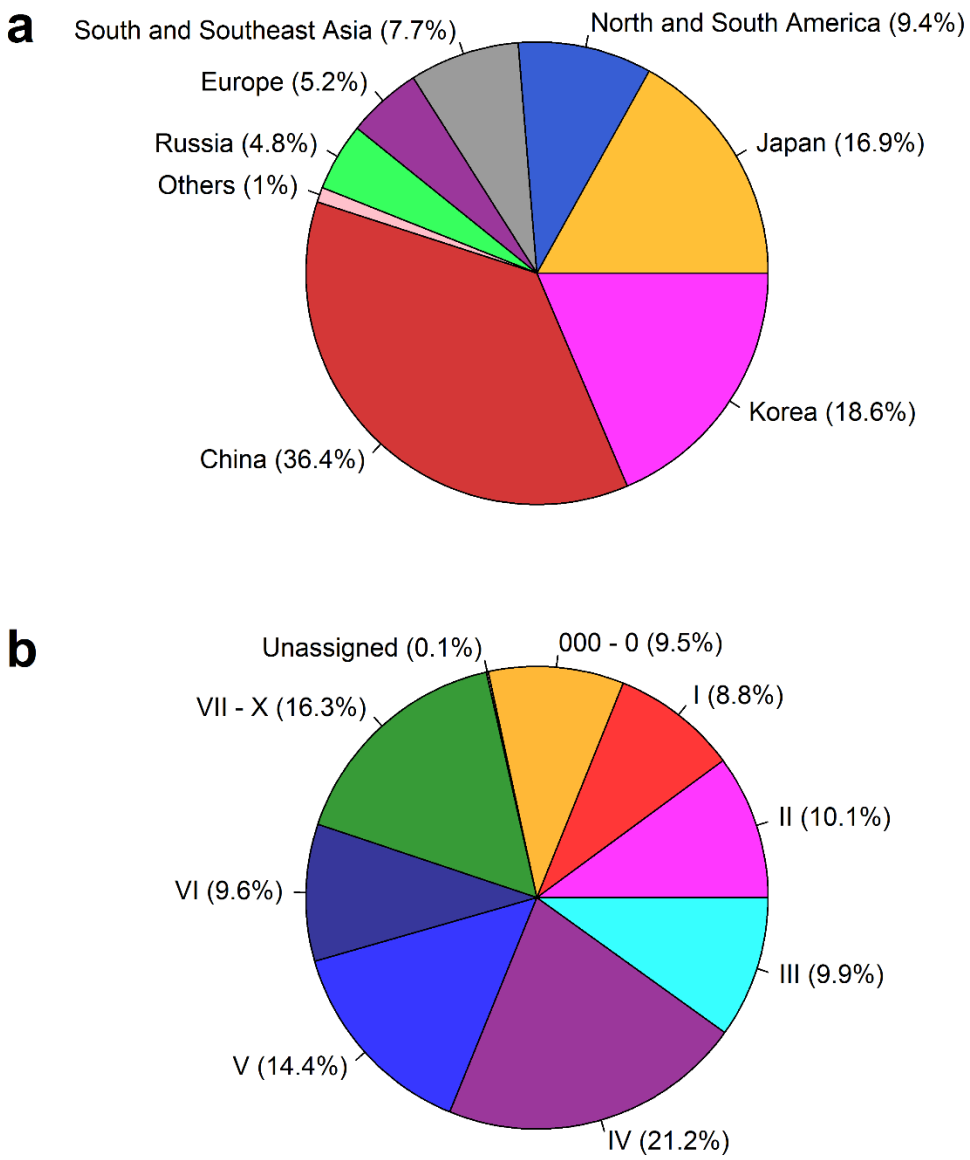
Table 2. Seed oil and protein means and country frequencies for the major haplotypes observed in candidate gene region on chromosome 20.

				Haplotype Frequency									Mean (%)	
Block No.	**Position (Mb)**	**Haplotype No.**	**Haplotype**	**All**	**China**	**Korea**	**Japan**	**America**	**SSE Asia**	**Europe**	**Russia**	**US Ancestors**	**Oil**	**Protein**
B3	29.06 - 30.04	H1	TGTTCTGAATCCGAG	0.830	0.780	0.906	0.896	0.854	0.526	0.941	0.817	0.852	18.00	43.78
B3	29.06 - 30.04	H2	CGCCTCAGGCTTTGT	0.097	0.161	0.005	0.010	0.088	0.382	0.016	0.097	0.118	17.05	45.05
B3	29.06 - 30.04	H3	TACCCCGGGTCCGAG	0.006	0.008	0.000	0.004	0.001	0.000	0.019	0.027	0.000	15.83	44.32
B4	30.38 - 30. 93	H4	CCA	0.594	0.457	0.794	0.714	0.694	0.367	0.548	0.667	0.650	18.01	43.79
B4	30.38 - 30. 93	H5	TTA	0.343	0.484	0.141	0.225	0.268	0.500	0.403	0.263	0.294	17.72	44.04
B4	30.38 - 30. 93	H6	CCG	0.027	0.020	0.042	0.024	0.011	0.075	0.020	0.021	0.000	15.20	47.45
B5	31.15 - 32.05	H7	TCAATAATGT	0.923	0.938	0.890	0.913	0.966	0.856	0.963	0.930	0.970	17.92	43.87
B5	31.15 - 32.05	H8	CTAGCGCTAT	0.013	0.000	0.045	0.020	0.005	0.003	0.000	0.000	0.030	14.92	47.69
Total *Glycine max* Accessions				12116	4744	2365	2117	816	610	735	513	34	17.8	44.0

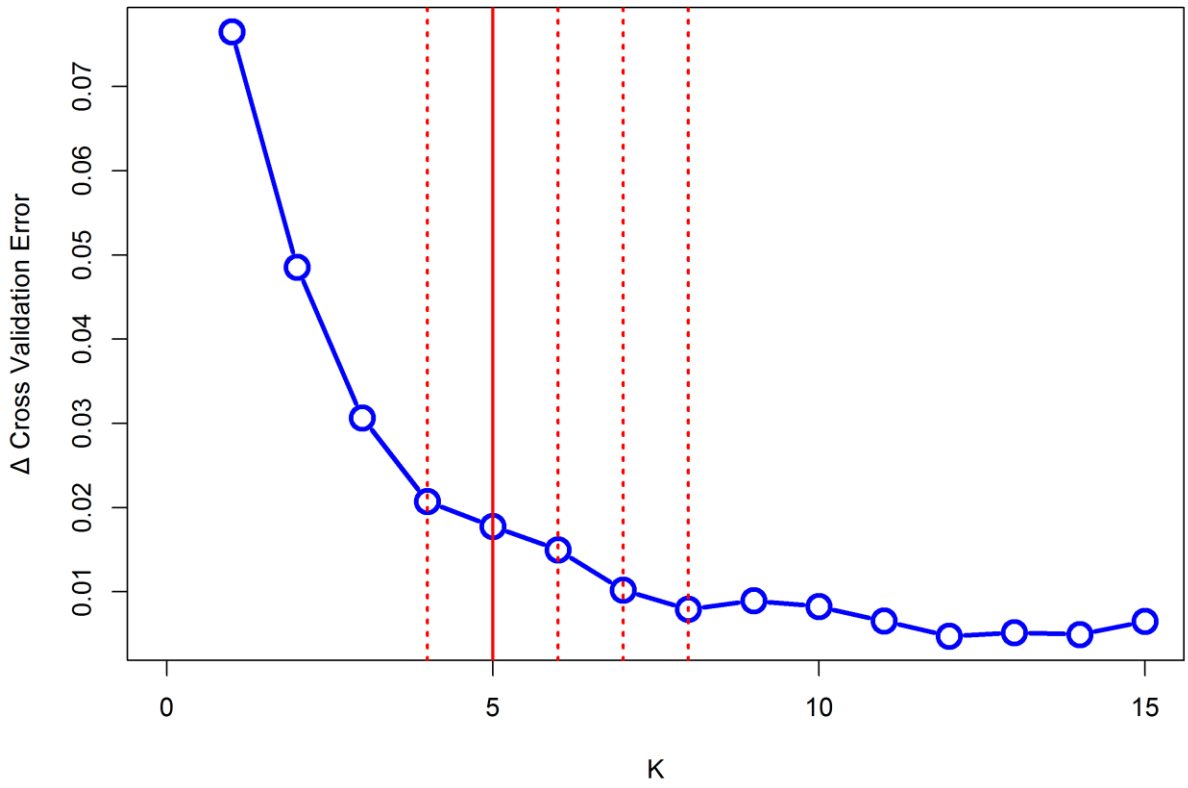
2.9 APPENDIX



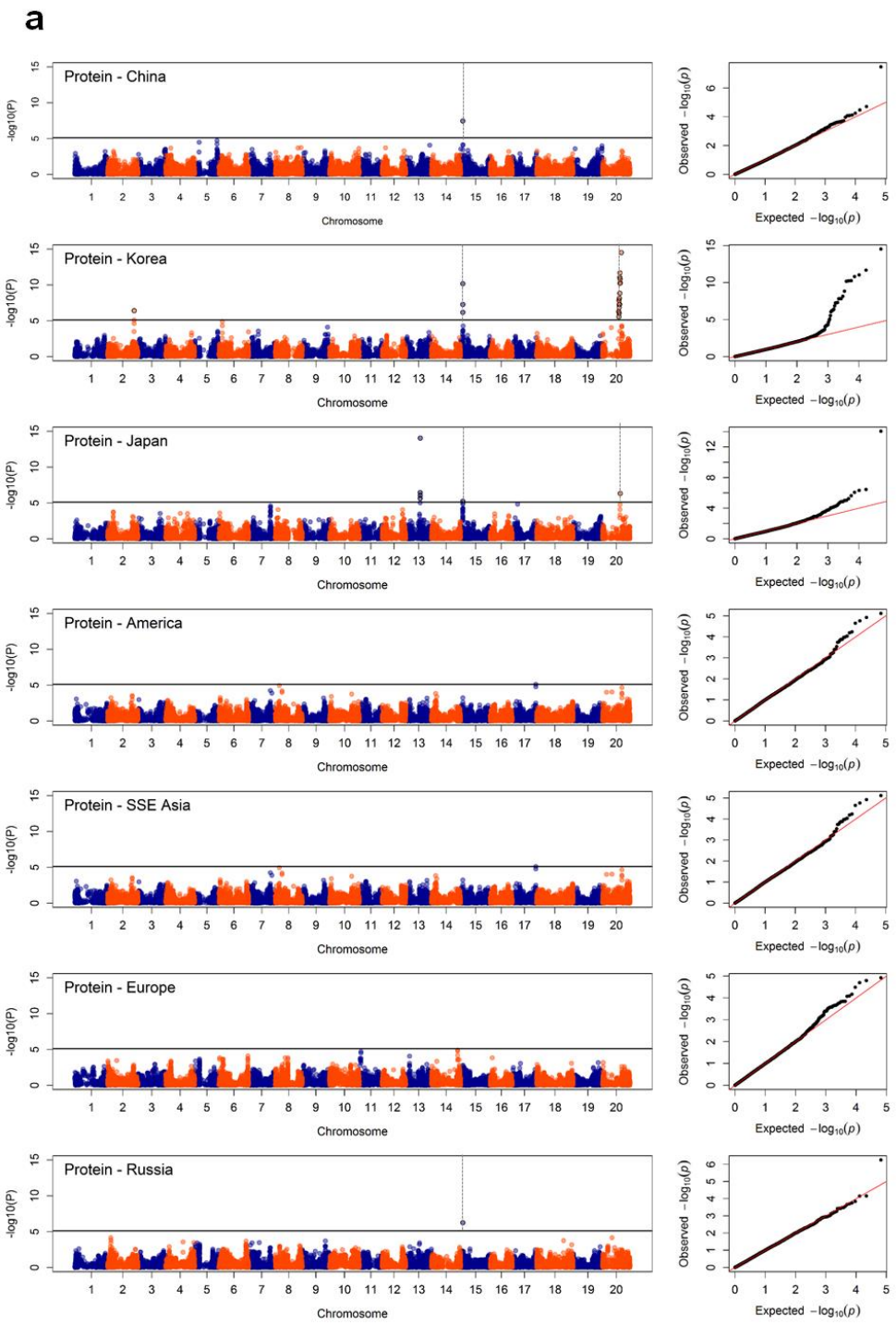
Supplementary Figure S1. The stepwise filtering of *G. max* and *G. soja* accessions held in the USDA Germplasm Collection for analysis of population structure and genome-wide association mapping.



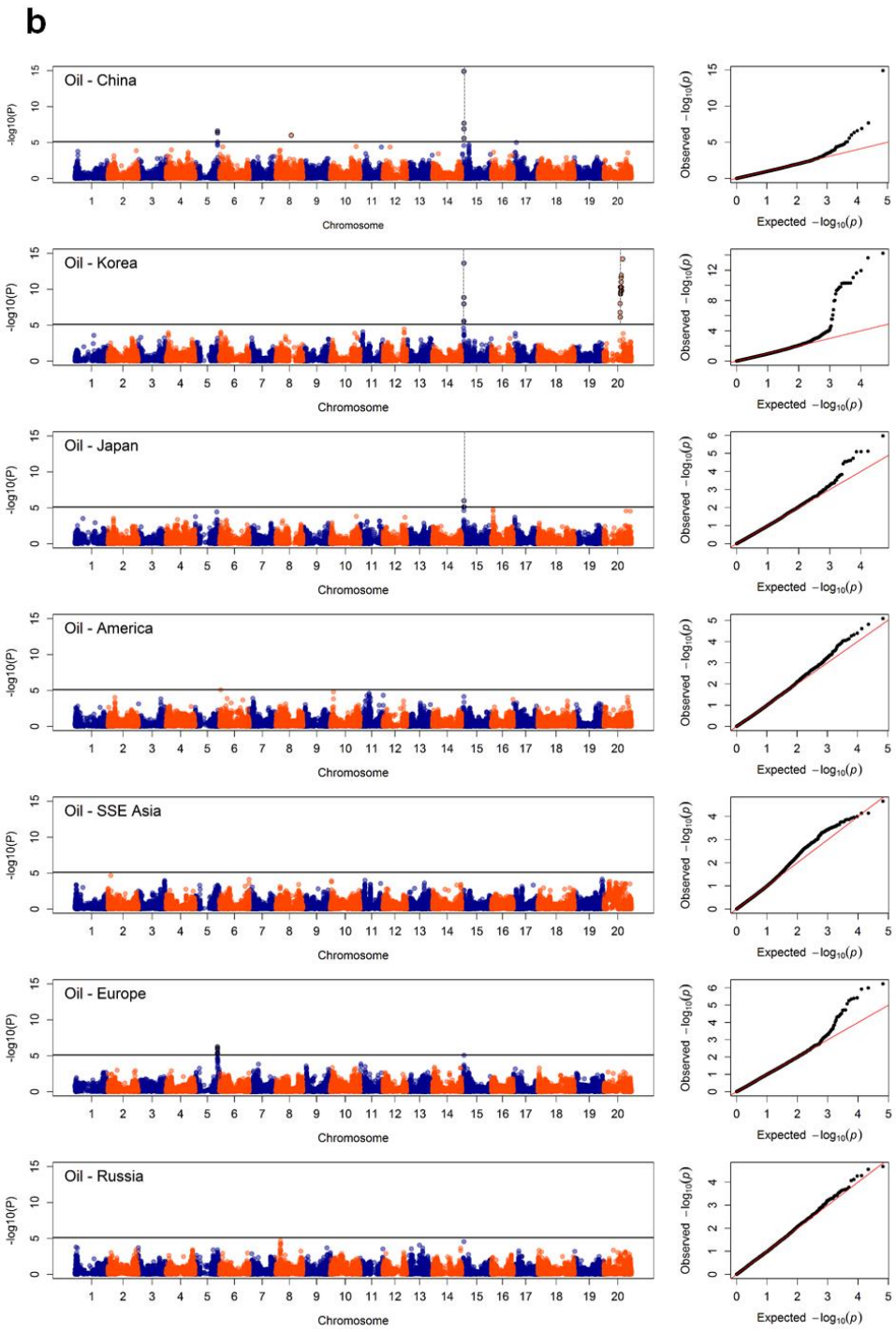
Supplementary Figure S2. Percentage distribution of the 14,430 soybean accessions used in the population structure analysis according to world region (panel a) and maturity group class (panel b).



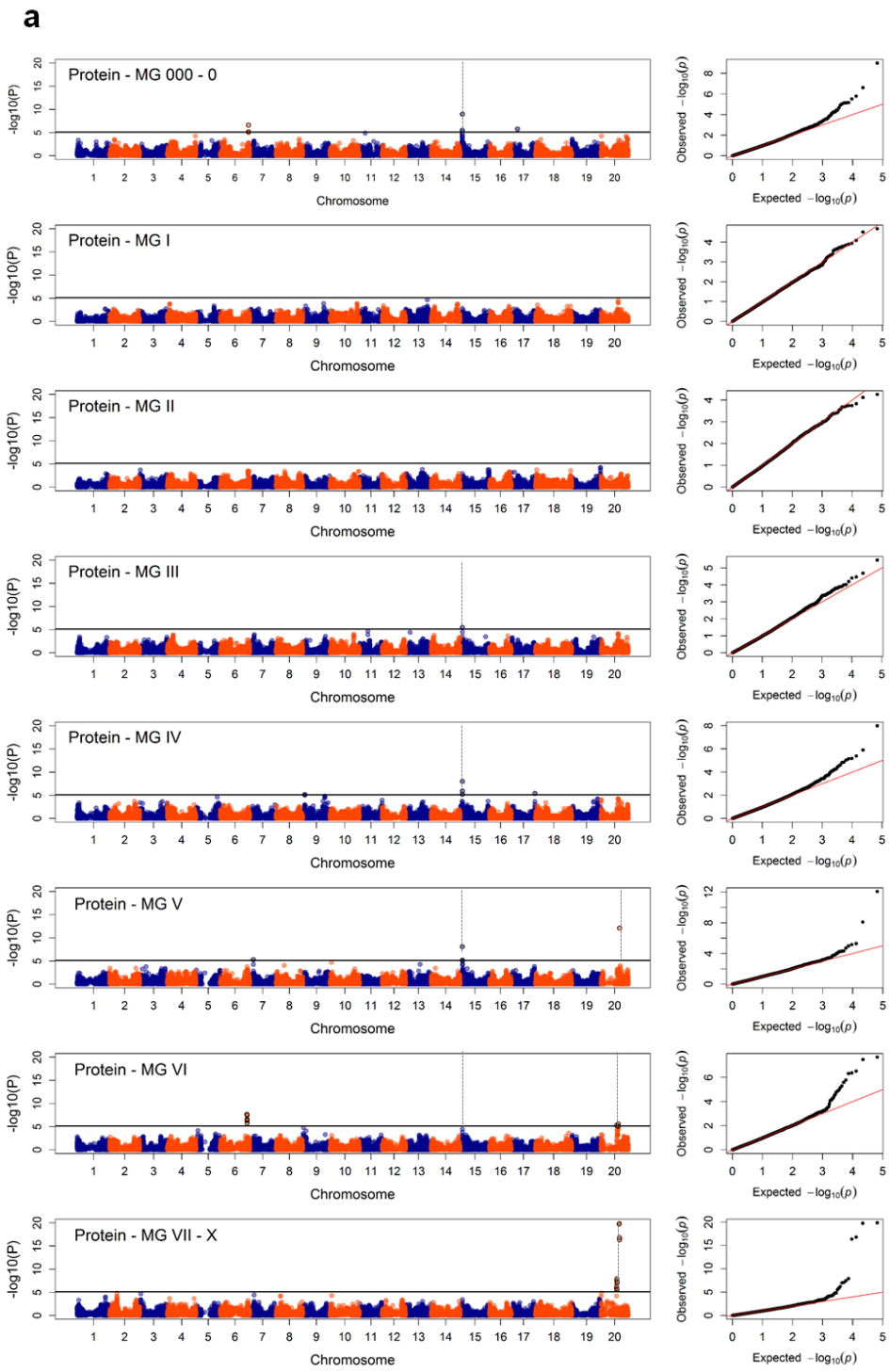
Supplementary Figure S3. Exploration of the optimal number of genetic subpopulations (K) using Δ cross-validation error values in the soybean germplasm collection. A solid line denotes the choice of K=5 which represents the most likely number of subpopulations within the soybean germplasm collection.



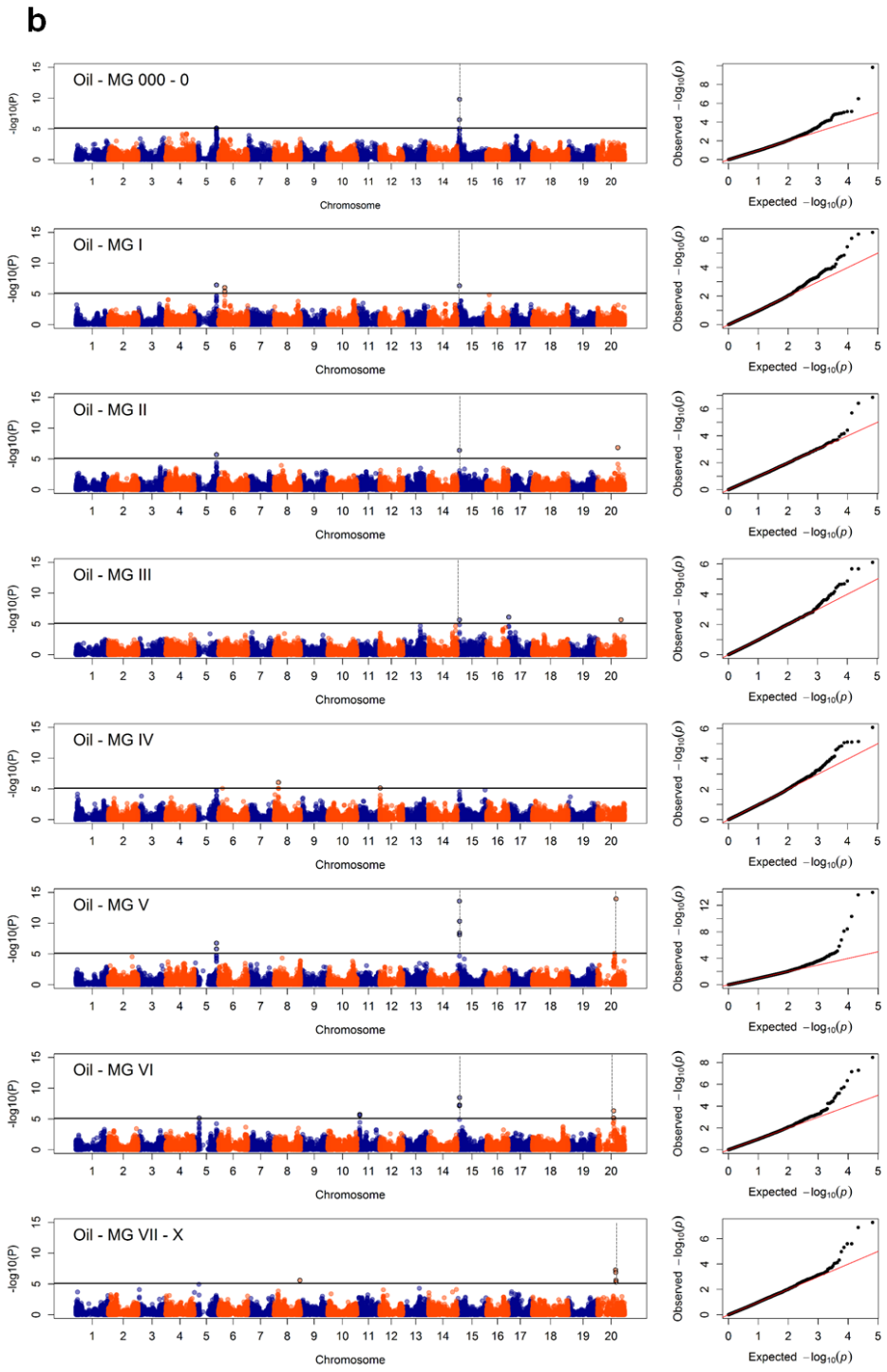
Supplementary Figure S4. Genome-wide association study for protein (panel a) and oil (panel b) within each world region class. Manhattan plot for association within each subpopulation (left) and quantile-quantile plots of -$\log_{10}$(P value) (right) are vertically arranged in each panel. Markers are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line indicates the calculated threshold value for declaring a significant association. The dashed vertical lines indicate the same significant associations detected in two or more MGs that co-localized with the significant SNP detected for either protein or oil using 12, 116 *G. max* accessions.



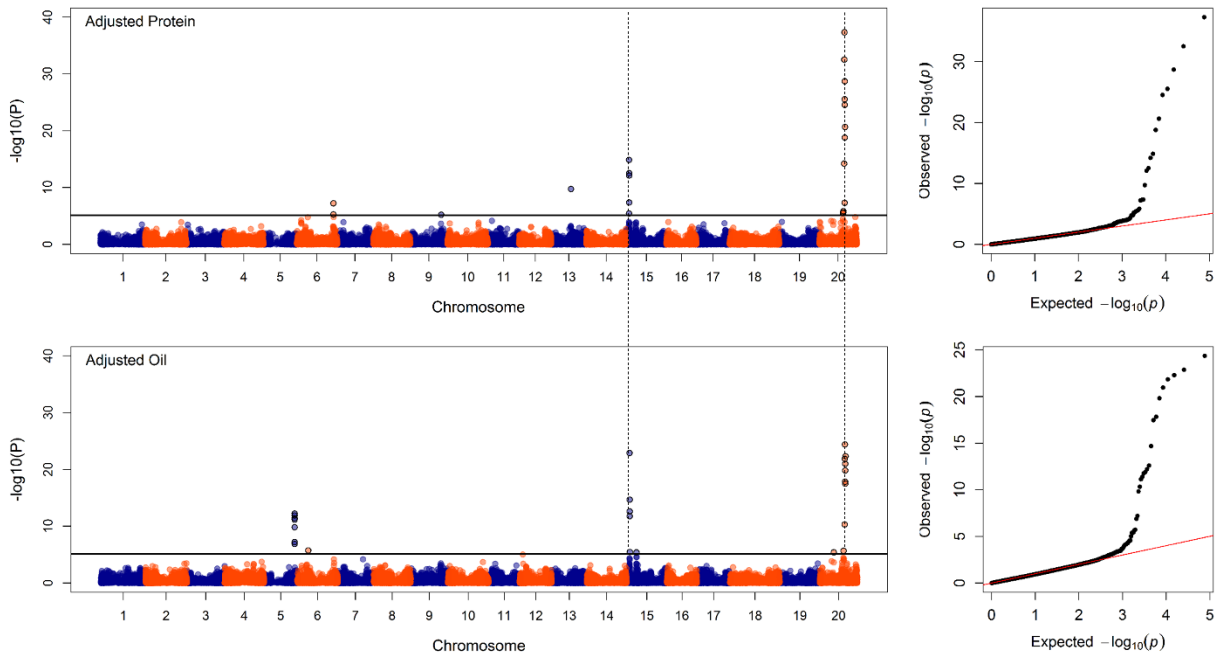
Supplementary Figure S4. Genome-wide association study for protein (panel a) and oil (panel b) within each world region class. Manhattan plot for association within each subpopulation (left) and quantile-quantile plots of $-\log_{10}$(P value) (right) are vertically arranged in each panel. Markers are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line indicates the calculated threshold value for declaring a significant association. The dashed vertical lines indicate the same significant associations detected in two or more MGs that co-localized with the significant SNP detected for either protein or oil using 12, 116 *G. max* accessions.



Supplementary Figure S5. Genome-wide association study for protein (panel a) and oil (panel b) within each MG class. Manhattan plot for association within each subpopulation (left) and quantile-quantile plots of -$\log_{10}$ (P value) (right). Markers are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line indicates the calculated threshold value for declaring significant association. The dashed vertical lines indicate the same significant associations detected in two or more MGs that co-localized with the significant SNP detected for either protein or oil using 12, 116 *G. max* accessions.



Supplementary Figure S5. Genome-wide association study for protein (panel a) and oil (panel b) within each MG class. Manhattan plot for association within each subpopulation (left) and quantile-quantile plots of -$\log_{10}$(P value) (right). Markers are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line indicates the calculated threshold value for declaring significant association. The dashed vertical lines indicate the same significant associations detected in two or more MGs that co-localized with the significant SNP detected for either protein or oil using 12, 116 *G. max* accessions.



Supplementary Figure S6. Genome-wide association scans for *G. max* accessions using adjusted phenotype data for seed oil and protein content. Manhattan plots show the associations for seed protein and oil with SNP markers that are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line denotes the calculated threshold value for declaring significant association. The dashed vertical lines indicate that the significant association positions on chromosome 15 and 20 for protein were the same as for those oil.

4.1.2 SUPPLEMENTARY TABLES

Supplementary Table S1. Distribution of 14,430 soybean accessions among seven arbitrarily chosen world regions (column names) by collation of accessions originating from 13 different specific world sub-regions (row names).

	World Region:								
World Sub-Region	China	Korea	Japan	America	SSE Asia	Europe	Russia	Others	Total
China	5216	0	0	0	0	0	0	0	5216
Korea	0	2665	0	0	0	0	0	0	2665
Japan	0	0	2426	0	0	0	0	0	2426
North America	0	0	0	1115	0	0	0	0	1115
South America	0	0	0	229	0	0	0	0	229
South Asia	0	0	0	0	280	0	0	0	280
Southeast Asia	0	0	0	0	819	0	0	0	819
Europe	0	0	0	0	0	749	0	0	749
Russia	0	0	0	0	0	0	682	0	682
Africa	0	0	0	0	0	0	0	148	148
Central Asia	0	0	0	0	0	0	0	18	18
Western Asia	0	0	0	0	0	0	0	10	10
Australasia	0	0	0	0	0	0	0	9	9
Others	0	0	0	0	0	0	0	64	64
Total	5216	2665	2426	1344	1099	749	682	249	14430

Supplementary Table S2. Distribution of 14,430 soybean accessions among eight maturity group MG classes. Of the 13 individual MGs (i.e. row numbers), the earliest maturing three (000, 00, 0), and the latest maturing three (VIII, IX, X) were combined (because of low accession numbers) to create just eight MG classes.

	Maturity Group Class:									
Maturity Group	000 - 0	I	II	III	IV	V	VI	VII - X	Unassigned	Total
000	118	0	0	0	0	0	0	0	0	118
00	405	0	0	0	0	0	0	0	0	405
0	845	0	0	0	0	0	0	0	0	845
I	0	1271	0	0	0	0	0	0	0	1271
II	0	0	1457	0	0	0	0	0	0	1457
III	0	0	0	1430	0	0	0	0	0	1430
IV	0	0	0	0	3066	0	0	0	0	3066
V	0	0	0	0	0	2080	0	0	0	2080
VI	0	0	0	0	0	0	1381	0	0	1381
VII	0	0	0	0	0	0	0	857	0	857
VIII	0	0	0	0	0	0	0	820	0	820
IX	0	0	0	0	0	0	0	603	0	603
X	0	0	0	0	0	0	0	77	0	77
Unassigned	0	0	0	0	0	0	0	0	20	20
Total	1368	1271	1457	1430	3066	2080	1381	2357	20	14430

Supplementary Table S3. Accession identification numbers and seed oil and protein content phenotypes for 12,116 G max accessions used for GWAS. Because of file size issue, this dataset was partially displayed in this section but the full dataset was made available at The Plant Genome website: https://dl.sciencesocieties.org/publications/tpg/tocs/8/3, navigate to Supplementary Table 3.

No.	Accession ID	Oil (%)	Protein (%)
1	FC001547	20.0	42.8
2	FC002109	18.6	41.9
3	FC003548	19.2	40.0
4	FC003609	20.0	43.2
5	FC003654-1	19.1	42.5
6	FC003659	17.3	45.9
7	FC003981	18.2	42.3
8	FC004002N	20.9	41.9
9	FC004007B	19.4	44.1
10	FC019976-1	18.9	43.2
11	FC019976-2	18.3	43.2
12	FC019979-1	19.1	43.0
13	FC019979-2	17.8	43.1
14	FC019979-3	19.8	41.3
15	FC019979-4	18.5	43.3
16	FC019979-5	19.9	41.6
17	FC019979-6	20.4	42.0
18	FC019979-7	17.4	43.0
19	FC029219	19.5	43.8
20	FC030265	18.6	43.9
21	FC030282	16.1	43.3
22	FC030683	17.7	45.8
23	FC030684	18.3	47.3
24	FC030689	17.3	42.5
25	FC030691	17.0	47.3
26	FC030692	17.7	45.6
27	FC030694	19.5	43.0
28	FC030967	16.4	44.8
29	FC031122	17.2	44.6
30	FC031416	14.6	43.1

Supplementary Table S4. Number of accessions within each subpopulation assigned to a given species based on membership coefficient criterion of >0.8. Accessions with a membership coefficient <0.8 were considered admixed.

	Subpopulation Number:				
Species	1	2	3	4	5
Glycine soja	663	0	0	0	0
Glycine max	1 [a]	12	1583	2797	334

[a] Genotypically, a *G. soja* , but phenotypically classfied as a *G. max* accession (PI549045A).

Supplementary Table S5. Number of accessions within each subpopulation assigned to each world region class based on membership coefficient criterion of >0.8. Accessions with a membership coefficient <0.8 were assigned to the admixed group.

	Subpopulation Number:					
World Region	1	2	3	4	5	Admixed
China	130	11	928	39	185	3923
Korea	277	0	0	952	0	1436
Japan	140	0	12	1701	4	569
America	0	0	43	54	16	1231
SSE Asia	0	0	560	11	1	527
Europe	0	1	0	27	39	682
Russia	117	0	0	2	86	477
Others	0	0	40	11	3	195

Supplementary Table S6. Number of accessions within each subpopulation assigned to each MG class based on membership coefficient criterion of >0.8. Accessions with a membership coefficient <0.8 were assigned to the admixed group.

	Subpopulation Number:					
Maturity Group Class	1	2	3	4	5	Admixed
000 - 0	63	0	1	70	159	1075
I	34	0	1	78	100	1058
II	64	11	19	142	53	1168
III	38	1	26	320	16	1029
IV	55	0	163	853	5	1990
V	249	0	237	568	0	1026
VI	95	0	195	459	0	632
VII - X	58	0	940	307	0	1052
Unassigned	8	0	1	0	1	10

Supplementary Table S7. Mean ancestry estimates of accessions within each world region relative to the subpopulation assignment. Estimates also provided for the 34 accessions that comprise a group known as USA soybean ancestors (Gizlice et al., 1994; Li and Nelson, 2001; Ude et al., 2003).

	Subpopulation Number				
World Region	1	2	3	4	5
China	0.07	0.22	0.36	0.13	0.22
Korea	0.11	0.03	0.04	0.67	0.16
Japan	0.10	0.03	0.03	0.79	0.05
America	0.02	0.31	0.17	0.22	0.28
SSE Asia	0.05	0.05	0.72	0.13	0.04
Europe	0.07	0.30	0.07	0.23	0.33
Russia	0.22	0.20	0.04	0.09	0.45
Others	0.03	0.19	0.32	0.27	0.20
Ancestor	0.02	0.25	0.14	0.36	0.24

CHAPTER 3: GENOME-WIDE ASSOCIATION MAPPING OF QUALITATIVELY INHERITED TRAITS IN A LARGE GERMPLASM COLLECTION

This chapter has been accepted for publication*: Bandillo N, Lorenz A, Graef G, Jarquin D, Hyten D, Nelson R, Specht J (2016). Genome-wide association mapping of qualitatively inherited traits in a germplasm collection.* ***Plant Genome.***

3.1. ABBREVIATIONS: CLG, classical linkage group; cM, centi-Morgan; GRIN, Germplasm Resource Information Network; GBS, genotyping by sequencing; GWA, genome-wide association; GWA, HPS, hydrophobic protein from soybean; genome-wide association; LD, linkage disequilibrium; LG, linkage group; MG, maturity group; MLG, molecular linkage group; MAF, minor allele frequency; MLM, mixed linear model; NIL, near-isogenic line; SNP, single nucleotide polymorphism.

3.2. ABSTRACT

Genome-wide association (GWA) has been used as a tool for dissecting the genetic architecture of quantitatively inherited traits. We demonstrate here that GWA can also be highly useful for detecting the genomic locations of major genes governing categorically defined phenotype variants that exist for qualitatively inherited traits in a germplasm collection. GWA mapping was applied to categorical phenotypic data available for ten descriptive traits in a collection of ~13,000 *Glycine max* (L.) Merr. accessions that had been genotyped with a 50K SNP chip. A GWA on a panel of accessions of this magnitude offered substantial statistical power and mapping resolution, and we found that GWA mapping resulted in the identification of strong SNP signals for 23 known genes as well as several heretofore unknown genes controlling the phenotypic variants in those traits. Because some of those genes had been cloned, we were able to show that the narrow SNP signal regions we detected for the phenotypic variants had chromosomal bp spans that, with few exceptions, bracketed the bp region of the cloned gene coding sequences, despite variation in SNP number/distribution of chip SNP set. Our GWA results identified very narrow regions that likely contained the trait-governing candidate genes, and we provide insights on how to deal with digenic traits for which linkage or epistasis can influence the outcome. In essence, GWA mapping aimed at qualitatively inherited traits can provide a convenient path for rapidly generating high-resolution positioning of many yet to be mapped genes on the soybean genomic sequence map.

3.3. INTRODUCTION

In the USA, there are 30 USDA-ARS National Plant Germplasm System (NPGS) sites (http://www.ars-grin.gov/npgs/sitelist.html), which were established for the collection, preservation, and distribution of plant species accessions of national interest. A substantial amount of phenotypic data has been collected in many of these germplasm collections. The soybean repository is located at Urbana, IL USA (https://npgsweb.ars-grin.gov/gringlobal/site.aspx?id=24), and it contains accessions of two annual species – the wild *G. soja* and the cultivated *G. max*, plus accessions of 19 perennial *Glycine* species.

Nearly all of the annual *Glycine* accessions have been characterized by the Collection curation staff for many descriptive traits. Of particular interest to soybean breeders and geneticists are descriptor traits: maturity group; stem termination; flower color; pubescence color, form, and density; pod color; seed coat luster and color; and hilum color. At least two and often several phenotype variants are listed as categories for each trait. The phenotypic category names and codes for each descriptor trait can be found at https://npgsweb.ars-grin.gov/gringlobal/descriptors.aspx (select soybean, then click on any given descriptor name).

Phenotypic variants in most of the above soybean descriptor traits are known to be qualitatively inherited in a monogenic or a digenic (sometimes, in a trigenic or tetragenic) manner. Because inter-genic epistasis plays a role in some cases, the number of phenotypic variants can be fewer than the number expected in its absence. Past soybean inheritance studies involving qualitatively inherited traits have led to the

assignment of gene symbols to the alleles at each of the loci that were inferred to govern the trait. Palmer et al. (2004) listed 251 soybean genes, and also noted that 72 of these were members of 21 classical (i.e., non-molecular) linkage groups (CLGs). Based on molecular marker genotyping of bi-parental mapping populations in which some of those 72 genes were segregating (e.g., Shoemaker and Specht, 1995), 19 of those 21 CLGs (68 of the 72 genes) were assigned to molecular linkage groups (MLGs) that were labeled A1 to O (Cregan et al., 1999). The number of s assigned to MLGs has now increased from 68 to 77 (SoyBase; www.soybase.org, Grant et al., 2010). Obviously, the majority of known soybean genes have yet to be mapped. Moreover, even the genetically mapped genes have low resolution centi-Morgan (cM) map positions, except for a few cloned genes that now have a specified chromosomal base pair (bp) position on the Williams 82 reference genome.

Establishing a chromosomal bp map position for all soybean genes using molecular marker genotyped bi-parental mapping populations would be a laborious and expensive effort. However, two recent publications suggested to us that gene-mapping could be accomplished via an alternative approach. Sonah et al. (2014) used genotyping-by-sequencing (GBS) to generate 47,702 SNPs they used to genotype 304 soybean lines spanning maturity groups (MGs) 000 to II. After performing a population structure analysis, they conducted a genome-wide-association (GWA) analysis on just 139 MG 0 lines they had characterized for five agronomic and seed traits in six field environments. Their primary goal was to discover SNPs associated with those five quantitative traits, but they stated that “to validate our GWAS approach”, they also applied GWA to the flower, pubescence, and hilum color phenotypes that they had also recorded for those 139

lines. These authors also stated that they detected “a towering distribution of many (significant) SNPs” in the chromosomal regions corresponding to four classical genes known to control those three traits. Subsequently, Wen et al. (2015), using 342 land races and 1062 cultivars released during 2007-2012, used the 50K soybean chip to apply GWA to 1402 lines differing in flower color (two phenotypes), pubescence color (two phenotypes), and seed coat color (six phenotypes). In this set of MG I, II, and III genotypes, they detected strong SNP associations for those three traits in the same chromosomal regions as those reported by Sonah et al. (2014).

These two reports indicated that GWA could be used for quickly “mapping” many of the simply inherited classical genes that are known to qualitatively govern traits, and for which extensive phenotypic data exists in many germplasm collections. More importantly, the application of GWA to classical traits can result in immediate high-resolution, chromosomal bp map positions for the controlling genes, which would be useful to researchers interested in cloning any given classical gene of scientific or commercial interest.

To more thoroughly test this thesis, we conducted a GWA analysis using phenotypic category data for ten soybean descriptive traits listed in GRIN for *ca.* 13K *G. max* accessions genotyped with a 50K SNP chip (Song et al., 2013). A GWA on a panel of accessions of this magnitude can offer substantially greater statistical power and mapping resolution compared to the smaller panels used by Sonah et al. (2014) and Wen et al. (2015). Our primary objective was to assess the use of GWA as a tool for chromosomal bp positional mapping of (known and unknown) genes controlling major phenotypic variants associated with each of the ten soybean descriptive traits. Of interest

were three questions: What is the degree of SNP-signal resolution obtainable when a 50K SNP chip is used in a GWA to identify a chromosomal bp position of a gene locus controlling a given pair categorical phenotypic variants *vis-à-vis* a cloned gene bp sequence? Can GWA be used for digenic qualitative gene mapping if there are only three instead of four phenotypes because of epistasis? To what degree can a reduction in accession numbers be tolerated in GWA and yet still provide a GWA signal for a known gene locus. The results generated in this study will likely be of interest to researchers interested in high-resolution GWA mapping of genes governing qualitatively inherited traits in their specific crop species of interest.

3.4. MATERIALS AND METHODS

3.4.1. Plant Materials

The accessions used in this study are maintained in the U.S. Department of Agriculture (USDA) Soybean Germplasm Collection, and were described previously (Bandillo et al., 2015; Song et al., 2015). As of 31 May 2016, this collection contained 22,199 accessions of 21 species in the genus *Glycine* (https://npgsweb.ars-grin.gov/gringlobal/site.aspx?id=24), which included 1,181 wild annual *G. soja* accessions, 19,931 domesticated annual *G. max* accessions, and 1,007 accessions of the 19 perennial species.

3.4.2. Extraction of Genotype and Phenotype Data

Song et al. (2015) used an Illumina Infinium SoySNP50K iSelect Beadchip to genotype 19,648 accessions of the two annual species. Based on a pair-wise genetic similarity analysis of 18,840 *G. max* accessions genotyped with 42,509 SNPs, they discovered that 1682 accessions were 100%, and another 4206 were at least 99.9%, identical to at least one other accession. Relative to the 1168 *G. soja* accessions, the equivalent numbers were 95 and 362. In the *G. max* collection, there also are 600 near-isogenic line (NIL) accessions (not including the recurrent parents). Bandillo et al. (2015) removed the SNP-identical duplicates and also the NILs to conduct a population structure analysis of the two annual *Glycine* species, and then removed the *G. soja* accessions for a subsequent GWA analysis that targeted just two quantitatively inherited traits – soybean seed protein and oil. The step-wise filtering process conducted by Bandillo et al. (2015) resulted in 13,624 *G. max* accessions, and is the same accession set used in the present

study for the GWA mapping of ten descriptive traits. Any SNP with minor allele frequency (MAF) < 0.01 was removed from the genotype dataset for the GWA mapping. The SNP genotype data set is publicly available at http://www.soybase.org/dlpages/index.php.

The phenotypic data used in this study were obtained from the USDA Soybean Germplasm Collection general evaluation trials in which data were collected for morphological, agronomic, and seed quality traits. The trials were grown where the accessions were adapted, and in most cases, there was one replication in each of two successive years. For a comprehensive listing of all of the phenotypic categories and their codes relative to the ten descriptor traits, see the GRIN web-site: https://npgsweb.ars-grin.gov/gringlobal/descriptors.aspx, enter SOYBEAN, then click on these (abbreviated) descriptor names: MatGroup, StemTerm, FlwrColor, PubColor, PubForm, PubDensity, PodColor, SCoatLuster, SCoatColor, HilumColor. The genotyped accessions and their ten-trait phenotypes were filtered (see Fig. 1) to create a final data file of accessions and their phenotype categories by trait (see Table S1). Due to missing phenotype scores for some traits in some accessions, the total number filtered accessions varied by trait.

3.4.3. Genome-Wide Association Analysis

An intensive comparison of various GWA methods conducted by Wang et al. (2012) demonstrated that the mixed linear model (MLM) is the most promising for analyzing either binary/categorical or continuous traits in crops exhibiting population structure. MLM has been utilized in GWA mapping of continuous and binary or categorical traits in model plant species (Atwell et al., 2010), and in crop species such as

rice (Huang et al., 2010), corn (Romay et al., 2013) and barley (Wang et al., 2012). In this study, MLM was used for GWA mapping of either binary or categorical traits to handle the confounding effects caused by strong population structure present in the soybean germplasm collection. For each phenotype, marker-trait associations were tested using the Q+K model y= Xβ + Cγ + Zu + e, where y is a vector of phenotypes; β is a vector of fixed marker effects; γ is a vector of subpopulation effects; u is a vector of polygenic effects caused by relatedness , i.e., $\mathbf{u} \sim MVN(0, \mathbf{K}\sigma_u^2)$; X is a marker matrix; C is an incidence matrix containing membership proportions to each of the five genetic clusters identified by the ADMIXTURE analysis (Bandillo et al., 2015; Alexander et al., 2009); Z is the corresponding design matrix for u; and K is the realized relationship matrix estimated internally in the Factored Spectrally Transformed Linear Mixed Models (FaST-LMM) using the SNP data (Lippert et al., 2011). The above model was implemented using the FaST-LMM algorithm, which is a program designed to accommodate large datasets with reduced computational time. Association analyses were conducted across groups of accessions classified either by MG class or by world region. GWA mapping across all groups was conducted using only MAF>0.01 SNPs, with population structure was accounted for using both γ and u. The qqman R package (Turner, 2014) was used to visualize Quantile-Quantile (qq) plot, and the genomic inflation parameter lambda (Λ), a metric of the degree of inflation of *p*-values (Devlin and Roeder, 1999), was calculated.

We used an error value of -logP=5.17 for the detection of significant SNP associations, which was determined by Bandillo et al. (2015) in 13,624 *G. max* accessions to correspond to an experiment-wise Type I error value of alpha=0.05. Briefly,

the correlation matrix and eigenvalue decomposition among 42,509 SNPs were calculated to determine the effective number of independent tests (M_{eff}) (Li and Ji, 2005). The significance test criteria was then adjusted using the M_{eff}, with the correction (Sidak, 1967) $\alpha_p = 1 - (1- \alpha_e)^{1/Meff}$, where α_p is the computed comparison-wise error rate, whereas α_e is the inputted desired experiment-wise error rate (i.e., 0.05). Multiple-linear regression was used to estimate the proportion of phenotypic variance accounted for by significant SNPs after accounting for population structure effects.

3.4.4. Determining Global Distribution of Allelic Variation

Accessions were grouped into subpopulations defined by world region, which is a major determinant of population structure within the soybean germplasm collection as reported by Bandillo et al. (2015). World region subpopulations consisted of eight major manageable countries of origin: China (36%), North and South Korea (19%), Japan (17%), North and South America (9%), South and Southeast Asia (8%), Europe (5%), Russia (5%), and Others (Bandillo et al, 2015). Based on the results of GWA mapping, the closest SNP that tagged a classical gene locus was used to estimate the frequency of the two alleles at that locus. Allele frequencies were estimated within each subpopulation using the CrossTable function in the gmodels package, implemented in R software version 3.2.1 (R Core Team, 2014). At each SNP locus, a Fisher's exact test was used to test the null hypothesis that frequency of the allele conferring a trait of interest was the same across world regions. The allele frequency output from CrossTable was then used to make plots using the pie function in R.

3.4.5. Candidate Gene Annotations

Gene annotations were extracted using the *G. max* cv. Williams 82 gene models (assembly v1.01, JGI Glyma1.0 and Glyma1.1 annotation) downloaded from Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html). The genomic locations were obtained from the GFF file of *G. max* assembly v1.01, JGI Glyma1.0 and Glyma1.1 annotation, and were displayed using a chromosome visualization tool (Cannon and Cannon, 2011). A 250kb sliding window-approach (125 Kb upstream and 125 Kb downstream from the most significant SNP position) was used to search for functional genes - implemented in BEDTools (Quinlan and Hall, 2010). Candidate genes included (a) soybean genes of known function related to the trait, and/or (b) genes with function-known orthologs in *Arabidopsis*. Annotation data is presented only for non-cloned classical genes and new loci for which a GWA signal was detected in this study (Table S2).

3.5. RESULTS AND DISCUSSION

From a population of ~21K soybean accessions originating from 84 different countries, we extracted a large association panel for mapping genes governing the phenotypes of the ten qualitative descriptive traits. Population size (N in Fig. 1) for the initial GWA conducted on each trait, for which there were multiple phenotypic categories (Fig. 2), ranged from 13,617 to 10,618 accessions (Supplementary Figs. S1-S10). Using the mixed linear model (MLM) that corrects for the effects of population structure and genetic relatedness, GWA mapping identified a total of 739 significant SNPs (-logP > 5. 17) in 57 genomic regions among all ten traits (Table 1). Overall, the GWA Manhattan plots documented significant SNP signals that corresponded to 23 known classical genes – ten of which have been cloned (Supplementary Figs. S1-S10; Fig. 3). Also detected were several strong SNP signals that may correspond to heretofore unknown qualitative genes. The large population size (*ca.* 13K accessions), coupled with the substantial genetic diversity in the soybean germplasm collection, resulted in our GWA analyses providing high mapping resolution relative to pinpointing the chromosomal bp position of the genes controlling the phenotypic variation associated with these qualitatively inherited traits. In addition, the magnitudes of the –logP scores for the SNPs identifying qualitatively inherited genes obtained in this study were substantially higher than any previous GWA or QTL mapping study conducted to date in soybean (www.soybase.org). To leverage the fine bp map resolution obtainable from GWA, we assembled a list of Wm82.a1 version of annotated candidate genes (Supplementary Table S2) located within 250-bp regions centered on each GWA-detected SNP peak signal (but not for cloned gene SNP signals) to assess the plausibility of potential candidate genes for the known

(but not yet cloned) genes, and for the few SNP signals that did not correspond to a known gene locus. We document here that the chromosomal bp positions of the significant SNP signal regions overlapped the coding sequence bp positions of the ten cloned loci of: *E1/e1, E2/e2, E3/e3, Dt1/dt1, Dt2/dt2*, *W1/w1, T/t, Hps*, *I/i,* and *R/r* (see bold-faced bp positions in Table 1), except for the cloned *E4/e4* or the fine-mapped *L1/l1* loci, despite the non-uniform distribution of the chip set of SNPs in many localized regions of soybean genome. Our GWA map findings for each of the ten traits are successively presented in the next ten sections.

3.5.1. Maturity Group

Known Genes: Each accession in the USDA Soybean Germplasm Collection is assigned to one of 13 MG categories (000 to X) that best reflects its adaptation to latitudinal zones. Nine genes are known to control soybean flowering and maturity: *E1/e1* and *E2/e2* (Bernard, 1971), *E3/e3* (Buzzell, 1971), *E4/e4* (Buzzell and Voldeng, 1980), *E5/e5* (McBlain and Bernard, 1987), *E6/e6* (Bonato and Vello, 1999), *E7/e7* (Cober and Voldeng, 2001), *E8/e8* (Cober et al., 2010), and *E9/e9* (Kong et al., 2014; Zhao et al., 2016). The dominant *E* allele always conditions late maturity, except for *E6* and *E9*, which condition early maturity. Soybean leaves, starting with the unifoliolate leaflet pair on young seedlings, perceive the environmental signals of dusk and dawn and track the duration of the night length (Wilkerson et al., 1989). The *E/e* gene loci are presumed to govern the number of hours of cumulative night length needed to trigger floral induction, when existent vegetative meristems are converted to inflorescent meristems, which in turn, produce flowers (Weller and Ortega, 2015). Depending on temperature during this floral evocation stage, flowers appear about 20 to 35 days after floral induction

(Wilkerson et al., 1989; Bastidas et al., 2008). At least two genes, termed the *J/j* loci in the older literature, but recently shown by Cober (2010) to be specifically the genes *J/j* and *E6/e6*, govern the duration of the "juvenility phase", when leaves are not yet competent to perceive the dusk-to-dawn night length, thereby delaying the onset of transmission of the floral stimulus to the vegetative meristems to initiate floral induction (Benlloch et al., 2015). Recently, Zhao et al. (2016) noted that the recessive gene *e9* also delayed flowering and they considered it to be a long juvenility gene. Accessions homozygous for the recessive alleles at the *E6/e6* and *J/j* loci (i.e., *e6e6jj*) have a "long juvenile" period that delays floral induction under short days, making these accessions higher yielding in equatorial latitude soybean production areas (Carpentieri-Pipolo et al., 2000; Cober, 2010; Destro et al., 2001; Harada et al., 2015). Five of the *E/e* loci have been cloned (*E1* by Xia et al., 2012; *E2* by Watanabe et al., 2011; *E3* by Watanabe et al., 2009; *E4* by Liu et al., 2008; *E*9 by Kong et al., 2015), and have respective bp positions on Chrs 6, 10, 19, 20, and 16. Genetic mapping studies have shown that *E7* is linked to *E1* by *ca.* 6 cM (Cober and Voldeng, 2001), and that *E8* resides between two flanking SSR markers on Chr 4 (Cober et al., 2010; Cheng et al., 2011). The map positions of the *E5* maturity gene, and the *E6* and *J* juvenility genes are still unknown. Komatsu et al. (2007) detected a Chr 2 QTL for which the early flowering allele was dominant, and because that allele originated from the late-maturing parent, they speculated that this QTL might be a *J/j* locus or the *E6/e6* locus, but offered no confirmation.

GWA Map Signals: A GWA analysis of all 13,617 accessions spanning 13 MG groups (Fig. 2) generated a Manhattan plot that exhibited five highly significant signals, a moderately significant signal, and two other signals of borderline significance (Fig. S1a).

Notably, three of five major genomic regions had SNP bp ranges that spanned the Chr 6, 10, and 19 bp positions of the cloned *E1/e1*, *E2/e2*, and *E3/e3* loci (Table 1; for a magnified single Chr view of each of these three signals, see Figs. 3a, 3b, 3c, in which the green bars denote the bp position of the cloned gene). Because the *E7/e7* locus is closely linked to the *E1/e1* locus, its GWA signal may be commingled with the latter's signal (Fig. 3a). The other two of the five major signals did not correspond to any cloned or mapped *E/e* genes. However, two highly significant maturity QTLs have been reported that have map positions near these two SNP signals. The Chr 12 QTL detected by Li et al. (2008) and listed in SoyBase as Pod Maturity QTL 26-2, had a large 6-day additive effect on maturity. The Chr 11 QTL detected by Gai et al. (2007) and listed in SoyBase as Pod Maturity QTL 17-2, had a large 7-day additive effect. Lu et al. (2015) recently noted that the two QTL LOD scores peaked at positions near SSR locus Satt442 (Gm12:6361515-6361774), and near SSR locus Satt519 (Gm11:13984414-13984651), which were close to our Chr 12 and 11 GWA SNP positions (Table 1). Because of their large additive effects on maturity, Lu et al. (2015) considered these two QTLs to be important for breeder-manipulated adaptation in China, and stated that they had cloned candidate genes for these maturity QTLs, but were awaiting test results before publishing. The moderately significant single SNP at Gm18:59902680 was nearly identical to the Gm18:59603446) SNP signal detected by Wen et al. (2015), which those authors associated with the SoyBase Pod Maturity QTL 29-8.

Soybean adaptation to latitudes extending away from the equator generally requires that dominant *E* alleles conditioning late maturity at the *E/e* loci be replaced with recessive *e* alleles conditioning early maturity (Tsubokura et al., 2012, 2013; Zhai et al.,

2014; Zhao et al., 2016). This led us to conduct a GWA targeting just the 8,315 high-latitude adapted accessions of MGs 000 to IV (Fig. S1b), and a GWA targeting the 5,302 low-latitude adapted accessions of MGs V to IX (Fig. S1c). The SNP signals for the *E1/e1*, *E2/e2*, and *E3/e3* loci and those on Chr 11 and 12 that had been detected in the GWA of all MGs (Fig. S1a) were again detected in the high-latitude MG set (Fig. S1b), though at diminished –logP values, except for the *E2/e2* signal, whose –logP value signal was strengthened 2-fold. The original three borderline significant SNPs disappeared. In contrast, in the low-latitude MG set (Fig. S1c), maturity class variation attributable to the *E1/e1* and *E3/e3* loci was not detected, and the *E2/e2* signal and Chr 11 signal were not appreciably changed. This result led us to infer that very few, if any accessions in the MG V to X classes were homozygous recessive at the *E1/e1* and *E3/e3* loci. If so, this would imply that the attainment of a finer degree of latitudinal photoperiod adaptation *within* the five southern USA MGs arises solely from the *E2/e2* locus and from the two (yet-to-be cloned and named) *E/e* loci that underlie Chr 11 and 12 QTLs. Bernard (1971) reported that the dominant alleles of *E1* and *E2* delayed maturity/flowering by a respective 18/23 and 14/7 days. Using additional near-isogenic lines and more replications, McBlain et al. (1987) reported that the dominant alleles *E1*, *E2*, and *E3* delayed maturity/flowering by a respective 11/16, 11/7, and 6/6 days. Because the *E2/e2* locus has a smaller allelic effect on flowering date than on maturity date, it offers a distinct advantage over the other two loci when breeders seeking latitudinal photoperiod adaptation want to delay/advance the date of R7 (physiological maturity) without an equal (i.e., *E3/e3*) or larger (i.e., *E1/e1*) delay/advance in the date of soybean stage R1 (first flower). This may explain why maturity variation at the *E2/e2* locus has a stronger signal than the other two loci, not

only *within* the MG 000 to IV classes (Fig. S1b), but also *within* the MG V to X classes (Fig. S1c).

The cloned maturity gene locus *E4/e4* on Chr 20 was not detected in the 13,617-accession GWA of all MGs (Fig. S1a), nor was it detected in the 8,537-accession GWA of MG 000 to IV (Fig. S1b). This may not be surprising, because this locus may not come into play except in those soybean crop production areas that have rapidly developed in ever-higher latitudes, where breeders have been replacing the dominant *E4* allele (late flowering/maturity) with the recessive *e4* allele (early flowering/maturity) to create cultivars with a suitable photoperiod adaptation (Zhai et al., 2014; Zhao et al., 2016). So, we generated a GWA for just 1,199 accessions of MG 00 and 0 (Fig. S1d), and it displayed two significant Chr20 SNP signals, neither of which overlapped the *E4/e4* coding sequence. The more significant SNP max region was located *ca.* 2.3 Mbp upstream of that coding sequence (Table1; Fig. 3d). With only 1,199 accessions (Fig. S1d), this GWA may not have had sufficient statistical power for optimally resolving the *E4* gene position. Yet, a near-doubling of accession numbers, achieved by adding 1,237 MG I and 000 accession set to the 1,199 MG 00-0 set, resulted in the disappearance of the Chr 20 GWA signal (data not shown). The reason may be due to the observation that *E3* is epistatic to *e4* (Saindon et al., 1989a; 1989b).

A long juvenile period, produced in genotypes homozygous for recessive genes of *e6* and *j* (Cober, 2010), and in genotypes homozygous for the recessive gene *e9* located on Chr 16 (Zhao et al., 2016), provides a means for delaying the onset of flowering of genotypes adapted to non-equatorial environments so that these genotypes will then have greater yield potential when grown in near-equatorial short-day latitudes of soybean

production. To determine if we could detect any of these three loci, we conducted a GWA for just the 2,277 accessions of the late MGs VII - X (Fig. S1e); however, we detected only a Chr 12 signal, which (as noted above) likely corresponds to the SoyBase QTL 17-2 detected on Chr 12 by Li et al. (2008). However, that QTL was detected in an RIL population, so no information is available as to whether the early maturity allele for this QTL is dominant, as it would have to be if either our Chr 12 SNP signal or their QTL were to be considered as corresponding to either the *E6/e6* or a *J/j* locus.

3.5.2. Stem Termination Type

Known Genes: Based on the timing and abruptness of the termination of apical stem growth, soybean accessions have been classified into three phenotype categories known as determinate (D) – stem abruptly terminating, indeterminate (N) – stem tapering gradually toward tip, and semi-indeterminate (S) – intermediate between N and D (http://www.ars-grin.gov/npgs/). In D plants, apical stem growth abruptly ceases upon the occurrence of floral induction, which causes all existent meristems to transition from a vegetative state to a reproductive state, leading to formation of inflorescence meristems that eventually produce flowers (Benlloch et al., 2015; Weller and Ortega, 2015). The D plant stem apice thus becomes terminal flower that produces a pod-bearing raceme (Liu et al., 2010; Tian et al., 2010). However, in N plants, the primary apical meristem at the stem tip and those at the branch tips are not receptive to the floral induction signal, which allows stems and branches to continue to elongate before tapering off at the onset of seed-filling (stage R5), when developing seeds become the strongest sink for photosynthetic carbon (Bastidas, et al., 2008; Tian et al., 2010). Flowering in N plants is thus limited to lateral meristems. In S plants, the receptiveness of their apical meristems

to the floral stimulus is partial and gradual, leading to a less abrupt termination of main stem growth *vs.* D (Ping et al., 2014).

GWA Map Signals: Genome-wide association mapping, using all 12,034 accessions that had been classified as having a stem growth habit phenotype of D, S, or N (Fig. 2), resulted in the detection of two major SNP signals on Chr 19 and 18, whose positions corresponded to the cloned genes *Dt1/dt1* (Liu et al., 2010; Tian et al., 2010) and *Dt2/dt2* (Ping et al., 2014) (Fig. S2a; Table 1; Figs. 3e, 3f). A low-level significant GWA signal detected on Chr 19 was close to the *E3/e3* maturity locus – the latter is located *ca.* 2.5 Mbp downstream from the *Dt1/dt1* locus (*ca.* ~25 cM in Fig. 1 of Watanabe et al., 2009). This Chr 19 GWA signal (see right side of Fig. 3e) could have arisen because genetic linkage between the Chr 19 locus and the *Dt1/dt1* locus, but possibly also because of epistasis between those two loci. Bernard (1972) reported that (1) the *Dt1/dt1* gene locus was responsible for the soybean growth habit extremes of determinate (*dt1dt1* genotypes) and indeterminate (*Dt1Dt1* genotypes), (2) the dominant allele at the *Dt2/dt2* locus converted an indeterminate stem growth habit (*Dt1Dt1dt2dt2*) into an semi-determinate stem growth habit (*Dt1Dt1Dt2Dt2*), and (3) the recessive *dt1* gene suppressed the expression of the semi-determinate phenotype in a *dt1dt1Dt2Dt2* genotype, leading to a 9S:3N:4D F2 segregation ratio (i.e., recessive epistasis). To mitigate the impact of the epistatic effect of *dt1dt1* on *Dt2* expression, and to determine how closely the S and N phenotypic classifications (which are based on the presence of a terminal raceme and the degree of stem tapering) correspond to the actual genotype, we restricted our next GWA to the 6,149 accessions scored by the germ-plasm collection staff as having either an S or an N phenotype. This GWA focus on just the S and N

phenotypes was partly successful in strengthening the signal for the *Dt2/dt2* locus (Fig. S2b), but compared to the initial GWA (Fig. S2a), the Chr 19 signal (near the *E/e3* locus) disappeared and a new signal appeared on Chr 6, but again the *Dt1/dt1* signal did not disappear, indicating that a phenotypic-based definition of indeterminate and semi-determinate (as defined above) does not always correspond to the genotypic-based definitions of *Dt1Dt1dt2dt2* (*indeterminate*) and *Dt1Dt1Dt2Dt2 (semi-determinate)*. We next used the most significant SNP nearest the *Dt1* gene as a "tag" to perform discriminant analysis with manual checking, which revealed that 27% (261/951) of the S phenotypes, and 7% (356/5198) of N phenotypes, might actually be *dt1dt1* genotypes. In a new GWA conducted with these 261 S and 365 N accessions omitted (Fig. S2c), the Chr 18 *Dt2* signal was strengthened (Table 1; Fig. 3f), as was the Chr 6 signal, suggesting that the latter may be a "genetic background factor" that influences phenotypic distinction between N and S. Though the Chr 19 *Dt1* signal was further weakened, it was not purged, confirming that the GRIN-listed phenotypes for stem growth habit cannot be assumed to have the corresponding two-locus *Dt* genotypes reported by Bernard (1972), because the genetic background of the accession has a major influence on the stem termination phenotype.

3.5.3. Flower Color

Known Genes: Palmer et al. (2004) listed the six genes known to govern flower color. Yang et al. (2010) later listed the chromosomal positions and cloned candidate genes for five of these six loci, i.e., *W1/w1* (Chr 13), *W2/w2* (Chr 14), *W3/w3* (Chr 14), *W4/w4* (Chr 12), and *Wp/wp* (Chr 2). Yang et al. (2010) noted that *W2/w2* and *W3/w3* were closely linked at the top of Chr 14, and Buzzell et al. (1977) noted that *Wm/wm* is

tightly linked (2.2 cM) to *W1/w1* on Chr 13. Nine phenotypic flower color categories are known, but very few accessions exist for colors other than the common purple and white. Moreover, recessive epistasis plays a significant role in that only in a *W1W1* genotype background is expression of any of the seven variant flower colors governed by the other five loci possible. The *w1w1* genotypes always have white flowers.

GWA Map Signals: Relative to flower color in 12,431 accessions that we used for an initial GWA, five of the nine known phenotypic categories were present (Fig. 2), i.e., blue (*W1W1w2w2*), dark purple (*W1W1W3W3W4W4*), light purple (*W1W1W3W3w4w4*), purple (*W1W1w3w3W4W4*), near-white (*W1W1w3w3w4w4*), and white (*w1w1*) - see Yan et al. (2014) for phenotype-genotype details. However, just four significant regions were detected in this GWA (Fig. S3a), and all four were on Chr 13 and were not far from each other (Table 1; for a magnified view of the most significant SNP signal and green bar denoting the position of the cloned *W1/w1* locus, see Fig. 3g). When we next limited the GWA mapping to the 12,329 accessions that were just P (8,209) or W (4,120), the same four Chr 13 signals were again detected (Fig. S3b), though a new significant SNP signal appeared on Chr 19 at 36603029 bp (Table 1). That location is not consistent with the known chromosomal locations of all other flower color loci, though it is very close to the Chr 19 location of the *L1/l1* gene locus, whose dominant allele gives rise to a black *vs.* brown or tan pod color phenotype (pod color is discussed later). If that Chr 19 flower color signal is not a false positive, then because of our GWA is focused only on the P *vs.* W phenotype categories, this Chr 19 signal could only have been detected if the underlying gene had an epistatic impact on the P *vs.* W phenotypic calls. In any event, the most significant SNP regions identified in the GWA

analyses overlapped the cloned gene (Table 1). The other nearby significant SNP regions in Chr 13 resided about 1-2 Mb upstream of *W1* (Fig. 3g), and might simply arise from the extensive LD in this region. With fewer accessions, Sonah et al. (2014) detected 14 significant SNPs in a single region spanning 2.5 and 4.8 Mb (though their SNP max was 8.1 kb downstream of *W1/w1*), whereas Wen et al. (2015) reported five separate significant SNP signals ranging from 2833623 to 4559799 bp; the latter one was their SNP max, and it was the same SNP max detected in our study (Table 1).

3.5.4. Pubescence Color

Known Genes: Pubescence in soybean consists of the trichomes on stem, leaf, and pod surfaces. Its color is determined by the B-ring mono- *vs.* di-hydroyxlation pattern of the flavonoids deposited in those trichomes (Zabala and Vodkin, 2003). Bernard (1975a) reported that pubescence color was controlled by two loci, *T/t* and *Td/Td*, which interacted in a recessive epistatic manner to produce a F2 segregation ratio of 9 tawny (T): 3 light tawny (Lt): 4 gray (G). The germplasm curation staff used a near-gray (Ng) pubescence color phenotype to characterize accessions with a pubescence color that was indistinguishable from G, but which also had a hilum color that was not consistent with a presumptive *tt* genotype. The curation staff also noted that some accessions with a *TTtdtd* genotype also have an Ng phenotype. Only the *T/t* locus has been cloned (Toda et al., 2002; Zabala and Vodkin,y 2003). A Chr 3 genetic map position for the *Td/td* locus was reported in patent application (see Table 4 in Behm et al., 2011).

GWA Map Signals: Our initial GWA mapping of 12,360 accessions that included all T / Lt / Ng / G phenotypes (i.e., 6,166 / 425 / 85 / 5,684 accessions - see Fig.

2) resulted in the identification of several SNP-significant regions located on Chrs 3, 6, 12, 14, and 20 (Fig. S4a; Table 1). Of the 26 significant SNP signals located on Chr 6 from 17258654 to 19815389 bp (with a max SNP at 17567713 bp), one was at 18252495 bp, thus co-localizing with the position of the cloned *T/t* gene locus (Fig. 3h). For the same reasons noted in the stem termination and flower color sections, we attempted to mitigate epistasis and the low frequency of Lt and Ng phenotypes by conducting a GWA *without* the G pubescence color accessions (Fig. S4b). We expected the *T/t* locus signal to disappear, and the *Td/td* locus signal to be amplified, because the GWA would then be focused solely on pubescence color phenotypic variants inferably arising from just a *TT* *TdTd* *vs. TT* *tdtd* genotypic comparison. In that regard, we were nearly successful, given that the Chr 3 signal (which we infer to be the *Td/td* locus) was amplified 200-fold (Table 1; Fig. 3i), whereas the *T/t* locus signal was nearly extinguished, though not completely so. Our inference that the Chr 3 signal corresponds to the *Td/td* locus is also supported by the findings of Wen et al. (2015), who in their GWA of 1,402 lines for pubescence color, detected not only the *T/t* locus signal at 18118558 bp, but also a significant Chr 3 signal at 47244893 bp (see their Fig. 6A); however, they did not offer any commentary about that signal. Sonah et al. (2014) did not detect a Chr 3 signal in their GWA with 139 accessions, but did detect a large region comprised 68 significant SNPs (i.e., 17313874 to 21182692 bp) associated with the cloned T locus, though their two closest SNPs consisted of one 18.7 kb away, and another 100 kb more distant. In our all-accession GWA (Fig. S4a), we assumed that the borderline significant genomic regions on Chrs 12, 14, and 20 were false positives, given that and signals disappeared in the second GWA analysis (Fig. S4b) and those Chrs are not known to classical genes impacting pubescence

color.

3.5.5. Pubescence Form

Known Genes: Soybean accessions differ in the degree to which trichomes are angled upward from the leaf surface, varying from nearly vertical (i.e., erect) to nearly horizontal (i.e. appressed). Depending on the insect species and location (USA *vs.* Japan or Korea), either the erect or the appressed pubescence form can serve as an herbivory deterrent (Bernard, 1975c; Komatsu et al., 2007; Oki et al., 2012). Pubescence is erect in 99% of the USA-released cultivars in MGs IV and earlier, but is erect in only 52% of USA-released cultivars in MGs V and later. Bernard (1975c) reported that soybean trichome morphology was governed by two genes, *Pa1/pa1* and *Pa2/pa2*, which interacted to produce five phenotypes he called erect (E), near-erect, semi-appressed (Sa), near-appressed, and appressed (A). However, the near-erect and near-appressed classes were not easily distinguishable from the Sa class (nor are these two phenotype categories listed in GRIN). Considering only the E, Sa, and A phenotype classes, Bernard (1975b) surmised that a digenic model with an unusual epistasis pattern could account for his observation of a F2 phenotypic segregation ratio of 4 E : 11 Sa : 1 A; the homozygous genotypes were inferred to be (*Pa1Pa1Pa2Pa2* + *Pa1Pa1pa2pa2*) : *pa1pa1Pa2Pa2* : *pa1pa1pa2pa2*. Lee et al. (1999) postulated that these two genes were duplicates of an ancestral locus, and thus arose from the most recent soybean duplication event, bit because the duplicates were not of equal phenotypic strength to generate a duplicate dominant epistatic F2 ratio of 15 E: 1 A, they postulated that the weaker locus (*Pa2/pa2*) was undergoing evolutionary neo-functionalization. They reported that the *Pa1pa1* locus mapped to Chr 11 and the *Pa2/pa2* locus mapped to Chr 13, but noted that the E / Sa / A

phenotypes were difficult to accurately classify, and stated that this may have influenced their genetic mapping results.

GWA Map Signals: The E / Sa /A accessions in the accession set had frequencies of 7,744 / 1,474 / 2,886 (Fig. 2). In our initial GWA mapping of 12,104 accessions that were E / Sa / A, two high resolution SNP signals were detected that we inferred to correspond to the *Pa1/pa1* and *Pa2/pa2* loci on Chrs 12 and 13, respectively (Fig. S5a). To mitigate the impact of the epistasis, and to better amplify the *Pa2/pa2* signal, just the 4,360 accessions possessing the phenotypes of Sa (inferred to be *pa1pa1Pa2Pa2*) and A (inferred to be *pa1pa1pa2pa2*) were included in the next GWA (Fig. S5b). Though the *Pa1/pa1* signal did not diminish much (probably because of the difficulty encountered by the curation staff in distinguishing among the three phenotypes), the *Pa2/pa2* signal was amplified 4-fold. The mapping resolution for *Pa2/pa2* locus was remarkable, in that GWA SMP signal pinpointed a region of less than 50 kb (i.e., between Gm13:30665757..30708708), whereas the *Pa1* locus mapped to a location between Gm12:37036017..37786243 bp that spanned 750kb (Table 1; for a magnified view of the two signals, see Figs. 3j, 3k). These two loci have not been cloned, so the SNP signals detected here could be useful starting points for researchers interested in doing so. The discrepancy between the Chr 11 map position that Lee et al. (1999) reported for the *Pa1/pa1* locus and the Chr 12 map position for that locus that we report here, may be due to the fact that these two chromosomes are highly homoeologous at the respective map positions (Lee et al. 2001). It is possible that the RFLP markers used by Lee et al. (1999) may not have been homoeologous-specific, leading them to position *Pa1/pa1* on Chr 11, instead of Chr 12, where we mapped it with SNP markers.

3.2.6. Pubescence Density

Known Genes: The presence of trichromes on soybean stem, leaf, and pod surfaces can serve as mechanical barrier to many herbivorous insects (e.g., cutworms; Oki et al., 2012), including insects that vector major yield-reducing viruses (e.g., bean leaf beetles, Lam and Pedigo, 2001). Denser pubescence can delay the timing of insect-mediated pathogen infection (Gunasinghe et al., 1988; Ren et al., 2000; Pfeiffer et al., 2003). Pubescence density is a key insect deterrent trait relative to Japan breeding programs (Komatsu et al., 2007). Palmer et al. (2004) listed five genes that govern pubescence density, *P1/p1* (glabrous/normal), *P2/p2* (normal/puberlent), *Pd1/pd1* and *Pd2/pd2* (dense/normal for both), and the tri-allelic locus *Ps/Ps^s/ps* (sparse/semi-sparse/normal). Bernard and Singh (1969) provided photographs of the range from sparse to dense phenotypes. Genetic map data have revealed that these five loci are respectively located on Chrs 9, 10, 1, 11, and 12 (Palmer et al., 2004; Shoemaker and Specht, 1995; Cregan et al., 1999; Devine, 2003; Song et al., 2004; Komatsu et al., 2007).

GWA Map Signals: GWA mapping of 12,397 accessions that exhibited six phenotypic categories (i.e., dense / glabrous / normal / slightly dense / sparse / semi-sparse – see Fig. 2) resulted in the identification of several SNP signals (Fig. S6a; Table 1; Figs. 3l, 3m), two of which corresponded with the (SoyBase) linkage map positions of *Ps/Ps^s/ps* and *Pd1/pd1* loci that translate into Gm12:34877806 bp and Gm01:55523014 bp), respectively. This finding was consistent with the detection of two significant pubescence density QTLs of 1-2 (SoyBase: Gm12:35314290..37138680) and 1-1 (SoyBase: Gm01:52767178..55838478) reported by Komatsu et al. (2007) in a Japanese mapping population derived from a mating of a densely pubescent, insect-resistant

cultivar with a sparsely pubescent, insect-susceptible cultivar. A strongly significant GWA signal was also detected at the top of Chr 14, along with a borderline significant signal on Chr 7 (Table 1), but neither one can be the *Pd2/pd2* locus, which is known to map to Chr 11 (Devine, 2003; Seversike et al., 2008). Limiting the GWA analysis to just the 12,301 accessions possessing the three phenotypes of Sp / Ssp / N did not appreciably change the GWA results (Fig. S6b) – the *Pd1/ pd1* and *Ps/Pss/ps* signals and the Chr 14 signal remained, though the borderline significant Chr 7 signal did disappear. Further research will be needed to determine if the highly significant Chr 14 signal corresponds to a heretofore undiscovered genetic locus governing pubescence density. To sharpen the GWA focus on the *P1/p1* locus detected in the first GWA, we conducted a GWA using only glabrous and normal accessions (N = 7,660), and identified two separate, but closely located, significant regions of eleven and eight SNPs corresponding to the *P1/p1* locus, one with a SNP max of Gm09:4424863), and the other with a SNP max in the other region (Gm09:46139114) (Fig. 6c; Table 1); for a magnified view of those two adjacent positions, see Fig. 3n). Because of a very low frequency (just 35) of glabrous accessions (Fig. 2), this GWA exhibited substantial noise (i.e., many signals in the genome), but the Chr 9 *P1/p1* locus signals still stood out from that noise in terms of strong –logP values (Fig. S6c). Hunt et al. (2011) conducted transcriptional profiling of the two NILs Clark-*p1p1* and Clark-*P1P1* but mainly focused on Glyma04g35130 (BURP), which was overexpressed in Clark-*p1p1*. They offered no commentary about Glyma09g38410 (calreticulin-3 precursor), which was also listed in their Table 1, as being overexpressed in Clark-*P1P1*. It has Chr 9 bp position (i.e., 43780130 - 43785822) that falls within our Chr 9 SNP signal region (43686430 – 4485534), thus making Glyma09g38410 a

plausible candidate gene for *P1/p1* (Table S2). Our GWA mapping results will be useful to those wishing to clone the *Ps/Ps^s/ps*, *Pd1/pd1*, *P1/p1* loci, and the unknown gene locus corresponding to the strong Chr 14 SNP signal, particularly given that SoyBase lists no pubescence density QTLs on that Chr 14.

3.5.7. Pod Color

Known Genes: Pod walls of mature soybean plants exhibit a characteristic color that is frequently used in variety description and identification (Bernard, 1967). There are three distinguishable classes of pod color, i.e., black (Bl), brown (Br), and tan (Tn). The color of immature pods is green (i.e., chlorophyll), but shortly after the R7 stage of physiological maturity, the chlorophyll disappears to reveal the mature dry pod wall color. Soybean pod color is controlled by two genes, symbolized as *L1/l1* and *L2/l2* (Woodworth and Veatch 1929; Bernard, 1967; Kiang, 1990). The two loci interact in a dominant epistatic manner to produce a F2 phenotypic segregation ratio of 12 Bl : 3 Br : 1 Tn. The corresponding homozygous genotypes are (*L1L1 L2L2* + *L1L1 l2l2*) : *l1l1 L2L2* : *1l1 l2l2*. The *L1/l1* locus is loosely linked to the *Dt1/dt1* locus on Chr 19, and the *L2/l2* locus has been located at the top of Chr 3 (Song et al., 2004). To date, the *L1* gene has been fine-mapped to a 184.43 kb region on Chr 19 that spans 13 potential candidate genes, but the most likely one has not cloned *per se* (He et al., 2015).

GWA Map Signals: A GWA using the 12,365 accessions that exhibited five pod color phenotypic variants (Bl / Dbr / Br / Lbr / Tn) (Fig. 2) produced two highly significant SNP signals on Chr 19 and 3, plus a significant SNP signal on Chr 1 (Fig. S7a; Table 1; Figs. 3o, 3p). The Chr 3 SNP signal corresponding to the *L2/l2* locus

(Br/Tn) spanned a 1,091-kb region (i.e., Gm03:246658 to 1338018 bp). The Chr 19 SNP signal corresponding to the *L1/l1* locus (Bl / Br + Tn) spanned a 618-kb region (i.e., Gm19:37503524 to 38121212 bp). Interestingly, the Chr 1 SNP signal had a bp position nearly identical with the SNP signal detected for a gene locus governing green *vs.* yellow seed coat color (discussed later). To determine if we could improve the resolution of each of the two main signals, we conducted a GWA with the 12,064 accessions exhibiting just the Bl / Br / Tn pod colors, which have a respective frequency counts of 604 / 8,258 / 3,202 (Fig. 2). This GWA improved the strength of the *L2/l2* signal, and the Chr 1 signal disappeared, but surprisingly, so did the *L1/l1* signal (Fig. S7b), suggesting that the Dbr and Lbr phenotypes, which were included in the prior GWA, but were omitted in this GW, were variants more associated with *L1/l1* locus (and/or the Chr 1 signal) than with the *L2/l2* locus for Br/Tn color. Our final GWA targeted only the 8,862 accessions that had Bl / Br pod colors (i.e., the respective genotypes of *L1L1* -- -- / *l1l1 L2L2*) (Fig. S7c), and it resulted in the detection of only the *L1/l1* signal at a high significance level Table 1; Fig. 3o). He et al. (2015) inferred that, of the 13 gene candidates located in their fine-mapped Chr 19 *L2/l2* region, Glyma19g27460 was the most likely candidate; however, that candidate gene has a bp position located *ca.* 2.75 Mbp downstream from our region of 48 significant SNPs (i.e.,Gm19:36397778 to 38521183). The reason for this substantive localization difference between our study and their study is not clear.

3.5.8. Seed Coat Luster

Known Genes: Woodworth (1932) reported that the presence (+) / absence (-) of bloom on the soybean seed coat (the only two phenotypic categories he observed) was controlled in a triplicate recessive epistatic manner in which a dominant allele at *each* of

three loci had to be present for a bloom+ phenotype. He symbolized these three genes as *B1/b1*, *B2/b2*, and *B3/b3*, and observed a 27 bloom+ : 37 bloom- F2 ratio in one mating that supported his hypothesis that inbred cv Sooty was a *B1B1B2B2B3B3* homozygote), but in another mating, he reported just a monogenic 3:1 F2 ratio. Tang and Tai (1962) could not confirm this, and observed only a digenic 9 bloom+ : 7 bloom- F2 ratio. Lorenzen et al. (1989) used pedigree analysis to show that a RFLP marker on Chr 15 (LG-E) was completely associated (by co-descent) with a dull luster phenotype governed by a what they called "*B* gene locus" (without designating a numerical locus number). Chen and Shoemaker (1998) reported that they had mapped the *B1/b1* gene to Chr 13 (LG-F). Gijzen et al. (1999; 2003b; 2006) documented that the amount of endocarp adhering to the seed surface was the primary determinant of seed coat luster. A bloom+ phenotype is produced by dense or contiguous covering of honeycomb-like endocarp tissue, whereas a dull phenotype has a fragmented or patchy covering of endocarp, but a shiny phenotype lacks any endocarp deposit. A hydrophobic protein (HPS) is synthesized in the endocarp of the inner ovary (pod) wall, and though HPS is abundant on the surface of dull seed coats, it is present in only trace amounts on shiny seed coats. Gijzen et al. (1999) noted that this relationship was not absolute, and observed that the cv Sooty had a bloom+ phenotype (i.e., due to a heavy coating of endocarp tissue), but only has a trace amount of surface HPS, in contrast with cv Williams 82, which had a shiny phenotype (i.e., due to an absence of endocarp tissue), though it too had only a trace amount of surface HPS. However, when the *B1* gene was backcross-introgressed from cv Sooty (bloom+, HPS-) into the Clark-*ii* NIL (dull, HPS+), it resulted in a Clark-*iiB1B1* NIL (bloom+, HPS+), revealing that the latter NIL received from its Sooty parent a heavy

coating of endocarp (i.e., bloom), and from its Clark-ii NIL parent an abundance of surface HPS. Yet, when Gijzen et al. (2003b) mated soybean line OX281 (dull, HPS+) with Mukden (shiny, HPS-), only the two parental phenotypic combinations observed in the 82 F2 plants (and in a respective 3:1 ratio). Thus, the phased dominant/recessive phenotypes of dull/shiny and HPS+/HPS- seemed to be controlled by either one pleiotropic locus, or else by two tightly linked loci, that the authors mapped to a single "*B/b*" gene location on Chr 15 (LG-E). Gijzen et al. (2006) cloned the core *Hps* locus and showed that it consisted of a tandem array of reiterated units, with each 8.6 kb unit containing a single HPS open reading frame. The HPS protein is a critical allergen (Gijzen et al., 2003a) (http://www.allergen.org/viewallergen.php?aid=342) that causes asthma in persons allergic to soybean dust generated during fall field harvest and at seaport seed loading facilities.

GWA Map Signals: In GRIN, six phenotypic categories for this trait are listed: dense bloom / bloom / light bloom / dull / intermediate / shiny; however, most accessions belong to the D, I, or S categories (Fig. 2). We conducted an initial GWA using 12,278 accessions exhibiting five of those six categories (Fig. S8a), and detected two separate significant signals on Chr 15 suggesting that *B/b* and *Hps/hps* are two tightly linked loci (not one locus), and a strong signal at the top of Chr 9 that we symbolized as a non-numbered *B?/b?* locus to distinguish it from the mapped *B1/b1* locus that corresponds to the weaker signal on Chr 13. The intermediate (I) seed coat luster phenotype accounts for than half (i.e., 7,280) of the total accession set (Fig. 2), but the I accessions tend not to be consistent in luster phenotype when grown in different environments, in contrast to the accessions characterized as having a shiny or dull luster phenotypes that tend to be more

reliably consistent calls across environments. For that reason, we conducted another GWA omitting the I accessions, thereby using just the 4,998 accessions that were classified as B / Lb / D / S phenotypes (Fig. S8b), and it resulted in a 2-fold reduction in the Chr 15 and Chr 9 signals (likely because of the loss of statistical power when going from about 12K to 5K accessions). Interestingly, the Chr 13 (*B1/b1*) signal disappeared, whereas a new signal appeared on Chr 8 corresponding to gene locus *I/i*. The reason for the disappearance of the former is not clear, given the Chen and Shoemaker (1998) mapped *B1/b1* based on segregation of D *vs.* S. The appearance of the *I/i* signal may be related to the fact that the seed luster phenotype call can be influenced by whether the seed coats are fully pigmented (as in case of *ii* genotypes) or are yellow (as in *II* genotypes). A borderline significant signal also appeared on Chr 11. We conducted a final GWA using just the 4,868 accessions that exhibited just the D *vs.* S phenotypes (Fig. S8c). The Chr 15 (*B/b*) and Chr 9 (*B?/b?*) signals were re-strengthened by this targeting (Table 1). Note that the *Hps* gene has a signal just upstream from the *B/b* gene (Fig. 3q), which suggests two linked loci, rather than one pleiotropic locus (as noted above). However, the detection of two independent GWA SNP signals (i.e., Chr 15 and Chr 9) is not consistent with a postulated monogenic model of inheritance for the dominant D *vs.* the recessive S. But, the alternative digenic model is also not plausible, unless digenic epistasis results in the reduction of phenotype categories from an expected digenic four to just an observed epistatic number of two.

3.5.9. Seed Coat Color / Hilum Color

Known Genes: Considerable variation exists in soybean seed coat color and hilum color. The predominant seed coat color is yellow (Fig. 2). The colors of green and

black are less frequent, but are more common than the other colors. The inheritance of seed coat and hilum color was examined in detail by Bhatt and Torrie (1968) and later clearly summarized by Palmer et al. (2004). Given that all (but three of the) *G. soja* accessions exhibit black seed coats, whereas the majority of *G. max* accessions have yellow seed coats, the preference of yellow over pigmented seed coats was clearly domestication-related. Both seed coat and hilum color are governed by four independent major loci that interact in an epistatic manner. The four loci include: (1) the 4-allele inhibitor locus ($I/i^k,i^i/i$) whose alleles produce differing degrees of pigmentation intensity and pigmentation spread (i.e., the I allele attenuates the black and Imperfect black seed coat and hilum colors to a grey color, and converts the brown and buff colors to a yellow color (Bhatt and Torrie, 1968), (2) the pubescence color *T/t* locus that pleiotropically governs a +/- *di*-hydroxylation of the B-ring of the flavonoid pigments in the seed coat and hilum, (3) the 3-allele $R/r^m/r$ locus that governs the production of proanthocyanin pigments in the seed coat and hilum (Zabala and Vodkin, 2014), and (4) the flower color *W1/w1* locus that pleiotropically governs the production of +/- *tri*-hydroxylation of the B-ring pro-anthocyanin pigments in seed coat and hilum in *RRttW1W1* genotypes. All four loci have been cloned. The yet to be cloned *O/o* locus controls the phenotypic change from brown (*OO*) to red brown (*oo*) in the seed coat and hilum. There are three seed coat saddle pattern loci (i.e., *Kn/kn* where n = 1, 2, or 3), but *K/k* saddle pattern accessions are too rare for a useful GWA analysis.

GWA Map Signals: Only six seed coat color, and only six hilum color, variants had a phenotypic frequency >0.01 (Fig. 2). One GWA was focused on just the seed coat color variants (Fig. S9a; N = 12,174), and another GWA was focused on just the hilum

color variants (Fig. S10a; N = 10, 292), with both producing expected signals corresponding to the respective *T/t* and *I/i*^{k}/*i*^{i}/*i* loci on Chr 6 (18766611 bp) and Chr 8 (8396392 bp). Green seed coat color arises when the chlorophyll present therein does not degrade at seed maturity as it does in yellow seed coats (Woodworth, 1921). The Gn and Y phenotypes are controlled by the single gene *G/g*, which has been mapped to Chr 1 (Cregan et al. 1999). To improve the resolution of the seed coat color Chr 1 signal (Fig. S9a), we conducted GWA restricted to just the 10,134 accessions exhibiting Y and Gn seed coats (Fig. S9b). As expected, the *T/t* and *I/i* signals disappeared, and there was a 3-fold amplification of the –logP value of the Chr 1 signal. That signal corresponded to the *G/g* locus, and spanned a very small 25-kb region near the SNP max located at Gm01:52253980 bp (Table 1; Fig. 3r). Relative to the cloned *I/i* locus (Todd and Vodkin, 1996; Tuteja et el., 2009), the seed coat color GWA (Fig. S9a; Table 1) produced a signal that was weaker than the high resolution stronger signal generated in hilum color GWA (Fig. S10; Table 1; Fig. 3s), primarily because of the number of yellow accessions were far greater than the number of non-yellow phenotypes. For hilum color, Sonah et al. (2014) reported ten SNPs associated with the *I/i* locus on Chr 8, with SNP max at 84803396 bp, which was not far from our SNP max at 8396392 bp. Note that a signal for the *R/r* locus was not detected in the initial GWA of either trait, so we conducted a GWA using only the 4,158 accessions with Bl / Br hilum color phenotypes (Fig. 2) whose respective inferred genotypes are *RR* /*rr*. A highly significant signal was detected on Chr 9 (Fig. S10b; Table 1; Fig. 3t), and the SNP region (i.e., Gm09:41660046..43669720) brackets the cloned *R/r* locus bp position. A moderately significant signal detected on Chr 12 may represent an unknown gene locus that somehow impacts the Bl *vs.* Br

classification. Finally, to locate the gene locus *O/o*, a final GWA was performed on just the 3,037 accessions with Br / Rbr phenotypes (Fig. S10c). The *I/i* and *O/o* loci are known to be linked (about 18 cM) on Chr 8 (Palmer et al., 2004), and indeed the GWA signal for *O/o* was identified at a comparable bp positon (i.e., 312 kb; Gm08:4800584..5113384) just upstream from *I/i* (Table 1; Fig. 3u).

3.6. Phenotypic variance and distribution of the mapped genes

For each trait, the largest effect SNP signal explained the largest fraction of the total phenotypic variance, ranging from 11 to 59% (Fig. 4a), with the highest percentages observed for flower color (59%), pubescence color (52%) and hilum color (33%)). For hilum and seed coat color, the cumulative effects of five genes (e.g., G, *I, O, R and T)* explained up to 79% and 77% of total phenotypic variance, respectively. Overall, the cumulative contributions of all significant SNP signals to phenotypic variance explained about 48% on average, though it varied from 11% to 83% depending on the trait, which is comparable to SNP associations identified in *Arabidopsis thaliana* (Atwell et al., 2010), rice (Huang et al., 2011), and corn (Romay et al., 2014). In our study, the identified SNPs conferring new loci explained additional variation that ranged 4 – 15 %. For MG, the *E* genes (e.g., *E1, E2, E3, and E4*) explained 16% of phenotypic variance, whereas the non-*E* genes controlled an additional 7% of phenotypic variance. Population structure also explained some portion of total phenotypic variance, which ranged from 4 – 50 %, with the highest proportion observed for MG (50%), likely because population structure is closely related to the latitudinal photoperiod sensitivity of soybean (Bandillo et al., 2015). It is possible, of course, that some portion of phenotypic variance may have arose from

imperfect LD, imperfect phenotyping, digenic epistasis, and/or (undetected) small-effect modifier loci.

The global distributions of narrowed loci revealed an essential pattern of allelic variation by gene locus that likely reflected a geographical-based differential in the history of soybean breeding (Fig. S11). Several traits were found to be correlated with world region than MG, indicating that *G. max* subpopulations are structured more by geography than by MG class (see Fig. 4b). The overall distribution pattern of allelic frequency at the various gene loci illustrate how accessions originating from China and North America (i.e., USA and Canada) diverged from accessions originating from Japan and Korea. Between Japan and Korea, however, allelic frequency spectrum across were almost the same except for loci associated with maturity *(E3/e3*), flower color (*W1/w1*), pubescence (*Pa1/pa1, Pa2/pa2,* and *Ps/ps)* and hilum color (*O/o, R/r*). Similarly, China and North America differ in allele frequency spectrum of loci associated with breeding and genetic improvement such as maturity (*E1/e1, E2/e2,* and *E3/e3*), stem growth habit (*Dt1* and *Dt2*) and pubescence (*Pa1/pa1*, *Pa2/pa2, Ps/ps* and *Pa1/pa1)* and seed coat/hilum color (*I/i* and *O/o*).

Breeding objectives also are factors contributing to the observed substantial allelic variation. For example, the degree of trichome density and its orientation on the epidermal surfaces of soybean plants has been used in breeding aimed at deterring insect feeding or impairing the viability of insect larva (Hulburt et al., 2004; Kanno, 1996; Lambert et al., 1992). Based on global distributions of *Ps/ps,* accessions in Japan and Korea are predominantly *Ps,* while America and China had predominantly *ps* (frequency > 0.85). This is not surprising given the fact that Japan has used sparse pubescence in

their breeding programs as a key insect control strategy (Komatsu et al., 2007; Oki et al., 2012), while in America erect and normal pubescence are needed to deter feeding by the potato leaf hopper (Broersma, et al., 1972), which migrates in early summer northward from the Gulf Coast states where it over-winters each year (http://extension.cropsciences.illinois.edu/fieldcrops/alfalfa/potato_leafhopper/). Substantial allelic variation at some loci also might reflect cultural preferences and farming practices.

3.7. SUMMARY

Genome-wide association analysis is nominally treated as a tool to be used mainly for dissecting the genetic architecture of quantitatively inherited traits. However, as documented here, GWA can also serve as a highly useful tool for detecting major qualitative genes governing categorically defined phenotype variants that exist for given traits in a germplasm collection. Indeed, we used GWA to identify the chromosomal bp positions of 23 classical genes governing the phenotypic variants listed for ten key soybean descriptor traits. Because some classical genes had been cloned, we were able to show that the SNP signal regions we detected for their phenotypic variants had chromosomal bp positions that, but one exception, bracketed the cloned gene bp positions. Of particular interest was our detection of strong SNP signals that possibly tag heretofore unknown genes controlling some of these classical soybean descriptive traits.

This demonstration that GWA mapping aimed at qualitatively inherited traits can be used to quickly generate high-resolution positions for the controlling genes on a genome sequence map is likely to be of interest to researchers in other crop species that have

germplasm collections for which extensive data also exists for qualitatively inherited traits. We are now applying the GWA qualitative gene mapping protocol to all other qualitatively inherited soybean descriptor traits, and the results, when complete, will be documented in a forthcoming publication.

3.8. SUPPLEMENTAL MATERIAL

Table S1 contains a list of the 13,624 *G. max* accessions and corresponding phenotypic codes for each of ten descriptor traits; Table S2 contains a list of (Glyma) candidate genes in a 250 Kb window centered on each detected significant GWA SNP signal (excluding those for cloned genes); the Supplemental Figure contains Figures S1 to S11.

3.9. CONFLICT OF INTEREST

Authors declare that they have no conflicts of interest.

3.10. REFERENCES

Alexander, D.H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655-1664.

Atwell S., Huang Y. S., Vilhjalmsson, B. J., Willems, G., Horton M., Y. Li, et al. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631.

Bandillo, N., D. Jarquin, Q. Song, R. Nelson, P. Cregan, J. Specht, and A. Lorenz. 2015. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Gen. 8:1-13.

Bastidas, A. M., T. D. Setiyono, A. Dobermann, K. G. Cassman, R. W. Elmore, G. L. Graef, and J. E. Specht. 2008. Soybean Sowing Date: The Vegetative, Reproductive, and Agronomic Impacts. Crop Sci. 48:727-740.

Behm J., L. Cerny, T. Floyd, J. Hall, and D. Wooten. 2011. Methods and compositions for increased yield. Patent Application Publication. US 2011/0010793 A1.

Benlloch, R., A. Berbel, L. Ali, G. Gohari, T. Millán, and F. Madueño. 2015. Genetic control of inflorescence architecture in legumes. Front. Plant Sci. 6:543.

Bernard, R.L. 1967. The inheritance of pod color in soybeans. J. Hered. 58:16-168.

Bernard, R.L. 1971. Two major genes for time of flowering and maturity in soybeans. Crop Sci. 11:242-244.

Bernard, R.L. 1972. Two genes affecting stem termination in soybeans. Crop Sci. 12:335-239.

Bernard R.L. 1975a. The inheritance of near-gray pubescence color. Soybean Genet. Newsl. 2:31–33.

Bernard, R.L. 1975b. The inheritance of semi-sparse pubescence. Soybean Genet. Newsl. 2:33-34

Bernard, R.L. 1975c. The inheritance of appressed pubescence. Soybean Genet. Newsl. 2:34–36.

Bernard, R.L. and B.B. Singh. 1969. Inheritance of Pubescence Type in Soybeans: Glabrous, Curly, Dense, Sparse, and Puberulent. Crop Sci. 9:192-197.

Bhatt, G.M. and J.H. Torrie. 1968. Inheritance of pigment color in the soybean. Crop Sci. 8:617-619.

Bonato, E.R. and N.N. Vello. 1999. E6 a dominant gene conditioning early flowering and maturity in soybeans. Genet. Mol. Biol. 22:229-232.

Broersma, D.B., R.L. Bernard, and W.H. Luckmann. 1972. Some Effects of Soybean Pubescence on Populations of the Potato Leafhopper. Journal of Economic Entomology 65:78-82.

Buzzell R.I., Buttery B.R., and Bernard R.L. 1977. Inheritance and linkage of magenta flower gene in soybeans. Can J Genet Cytol. 19:749–751.

Buzzell, R.I. 1971. Inheritance of a soybean flowering response to fluorescent-daylength conditions. Can. J. Genet. Cytol. 13:703-707.

Buzzell, R.I. and H.D. Voldeng. 1980. Inheritance of insensitivity to long daylength. Soybean Genet. News. 7:26-29.

Cannon, E.K. and S.B. Cannon 2011. CViT: "Chromosome Visualization Tool" – A whole-genome viewer. International Journal of Plant Genomics. DOI:10.1155/2011/373875.

Carpentieri-Pípolo V., L.A. Almeida, R.A.S. Kiihl, and C.A. Rosolem. 2000. Inheritance of long juvenile period under short day conditions for the BR80-6778 soybean (*Glycine max* (L.) Merrill) line. Euphytica 112:203-209.

Chen, Z. and R.C. Shoemaker. 1998. Four genes affecting seed traits in soybeans map to linkage group F. J. Heredity 89:211-215.

Cheng, L., Y. Wang, C. Zhang, C. Wu, J. Xu, H. Zhu, J., et al. 2011. Genetic analysis and QTL detection of reproduction period and post-flowering reproductive period responses in soybeans. Theor. Appl. Genet. 123:421-429.

Cober, E.R. 2010. Long juvenile soybean flowering responses under very short photoperiods. Crop Sci.51:140-145.

Cober, E.R. and H.D. Voldeng. 2001. A new soybean maturity and photoperiod-sensitivity locus linked to *E1* and *T*. Crop Sci. 41:698-701.

Cober, E.R., J.W. Tanner, and H.D. Voldeng. 1996. Genetic control of photoperiod response in early-maturing, near-isogenic soybean lines. Crop Sci. 36:601-601.

Cober, E.R., S.J. Molnar, M. Harette, and H.D. Voldeng. 2010. A new locus for early maturity in soybean. Crop Sci. 50:524-527.

Cregan P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, et al. 1999. An integrated genetic linkage map of the soybean. Crop Sci. 39:1464–1490

Destro, D., V. Carpentieri-Pipolo, R.A.S. Kiihl, and L.A. Almeida. 2001. Photoperiodism and genetic control of the long juvenile period in soybean: a review. Crop Breeding and Appl. Biotechnol. 1:72-92.

Devine, T.E. 2003. The Pd2 and Lf2 loci define soybean linkage group 16. Crop Sci. 43:2028-2023.

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

Gai, J., Y. Wang, X. Wu, and S. Chen. 2007. A comparative study on segregation analysis and QTL mapping of quantitative traits in plants – with a case in soybean. Front. Ag. China 1:1-7.

Gijzen, M., Miller, S.S., Kuflu, K., Buzzell, R.I., and Miki, B.L.A. 1999. Hydrophobic protein synthesized in the pod endocarp adheres to the seed surface. Plant Physiol. 120:951–959.

Gijzen, M., R. Gonzales, D. Barber, and F. Polo. 2003a. Level of airborne Gly m 1 in regions of soybean cultivation. J. Allergy Clin. Immunol. 12:803-804.

Gijzen, M., C. Weng, K. Kuflu, L. Woodrow, K. Yu, and V. Poysa. 2003b. Soybean seed lustre phenotype and surface protein cosegregate and map to linkage group E. Genome 46:659-664.

Gijzen, M., K. Kuflu, and P. Moy. 2006. Gene amplification of the *Hps* locus in *Glycine ma*x. BMC Plant Biology **6**:6.

Gillman, J.D., A. Tetlow, J.-D. Lee, J.G. Shannon, and K. Bilyeu. 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.

Grant, D., R. T. Nelson, S. B. Cannon, and R. C. Shoemaker. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38:D843–D846.

Gunasinghe, U.B., M.E. Irwin, and G.E. Campmeier. 1988. Soybean leaf pubescence affects aphid vector transmission and field spread of soybean mosaic virus. Annals. Appl. Biol. 112(2).259-272.

Harada, A., L.S.A. Gonçalves, R.A.S. Kiihl, and D. Destro. 2015. Flowering under short days: Juvenile period and inductive phase estimates in soybean genotypes. Agronomy Science and Biotechnology. 1:10-16.

He, Q., H. Yang, S. Xiang, D. Tian, W. Wang, T. Ahao, and J. Gai. 2015. Fine mapping of the genetic locus L1 conferring black pods using a chromosome segment substitution line population of soybean. Plant Breeding 134:437-445.

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, et al. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961-967.

Hulburt, D.J., H.R. Boerma and J.N. All. 2004. Effect of pubescence tip on soybean resistance to lepidopteran insects. J. Econ. Entomol. 97:621-627.

Hunt, M., N. Kaur, M. Stomvik, and L. Vodkin. 2011. Transcript profiling reveals expression differences in wild-type and glabrous soybean lines. BMC Plant Biology 11:145.

Kanno, H. (1996) Role of leaf pubescence in soybean resistance to the false melon beetle, *Atrachya menetriesi* Faldermann (Coleoptera: Chrysomelidae). Appl. Entomol. Zool. 31:597–603.

Kiang, Y. T. 1990. Linkage analysis of Pgd1, Pgi1, pod color (*L1*), and determinate stem (*dt1*) loci on soybean linkage group 5. J. Hered. 81:402-404.

Komatsu, K., S. Okuda, M. Takahashi, R. Matsunaga, and Y. Nakazawa. 2007. Quantitative trait loci mapping of pubescence density and flowering time of insect-resistant soybean (Glycine max L. Merr.). Genet. Mol. Biol. 30:635-639.

Kong, F. H. Nan, D Cao, Y. Li, F. Wu, J. Wang, et al. 2014. A new dominant gene *E9* conditions early flowering and maturity in soybean. Crop Sci. 54:2529-2535.

Lam, W. F., and L.P. Pedigo. 2001. Effect of trichome density on soybean pod feeding by adult bean leaf beetles (Coleoptera: Chrysomelidae). J. Economic Entomology 94:1959.

Lambert, L., R.M. Beach, T.C. Kilen, and J.W. Todd. 1992. Soybean pubescence and its influence on larval development and oviposition preference of lepidopterous insects. Crop Sci. 32:463-466.

Lee, J.M., A.L. Bush, J.C. Specht, and R.C. Shoemaker. 1999. Mapping of duplicate genes in soybeans. Genome 42:829-836.

Lee, J.M., D. Grant, C.E. Valejos, and R.C. Shoemaker. 2001. Genome organization in dicots. II. Abravidopsis as a 'bridging species' to resolve genome evolution events among legumes.

Li, J. and L. Ji. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb) 95:221-227.

Li, W., D. Zheng, K. Van, and S. Lee. 2008. QTL mapping for major agronomic traits across two years in soybean (Glycine max L. Merr.). J. Crop Sci, Biotech. 11:171-190.

Lippert, C., J. Listgarten, Y. Liu, C.M. Kadie, R.I. Davidson, and D. Heckerman. 2011. FaST linear mixed models for genome-wide association studies. Nature Methods 8:833-U94.

Liu, B., A. Kanazawa, H. Hatsumura, R. Takashashi, K. Harada, and J. Abe. 2008. Genetic redundancy in soybean photoresponses associated with duplication of phytochrome A gene. Genetics 180:995-1007.

Liu, B., S. Watanabe, T. Uchiyama, F. Kong, A. Kanazawa, Z. Xia, et al. 2010. The soybean stem growth habit gene *Dt1* is an ortholog of Arabidopsis TERMINALFLOWER1. Plant Physiol. 153:198–210.

Lorenzen, L.L., Boutin, S., Young, N., Specht, J.E., and Shoemaker, R.C. 1995. Soybean pedigree analysis using map-based molecular markers: I. Tracking RFLP markers in cultivars. Crop Sci. 35:1326–1336.

Lu, S. Y. Li, J. Wang, H. Nan, D. Cao, X. Li, et al. 2015. Identification of additional QTLs for flowering time by removing the effect of maturity gene E1 in soybean. Journal of Integrative Agriculture. 15:60345.

McBlain, B.A and R.L. Bernard. 1987 A new gene affecting the time of flowering and maturity in soybeans. J. Hered. 78:160-162.

McBlain, B.A, J.D. Hesketh, and R.L. Bernard. 1987. Genetic effects on reproductive phenology in soybean isolines differing in maturity genes. Can. J. Plant Sci. 67:105-116.

Oki, N., K. Komatsu, T. Sayama, M. Ishimoto, M. Takahashi, and M. M Takahashi. 2012. Genetic analysis of antixenosis resistance to the common cutworm (*Spodoptera litura* Fabricius) and it relationship with pubescence characteristics in soybean (Glycine max (L.) Merr.). Breeding Science 61:608-617.

Palmer R.G., T.W. Pfeiffer, TW, G.R. Buss, and T.C. Kilen. 2004. Qualitative genetics. In: H.R. Boerma and J.E. Specht, editors. Soybeans: improvement, production, and uses. 3rd ed. Agron. Monogr. 16. ASA, CSSA, and SSSA, Madison, WI. p. 137–233.

Pfeiffer T. W., R. Peyyala, Q. Ren, and S. A. Ghabrial. 2003. Increased soybean pubescence density. yield and soybean mosaic virus resistance effects. Crop Sci. 43:2071–2076

Ping, J., Y. Liu, L. Sun, M. Zhao, Y. Li, M. She, et al. 2014. *Dt2* is a gain-of-function MADS-domain Plant Cell 26:2831-2842 factor gene that specifies semideterminacy in soybean. Plant Cell 26:2831-2842.

Quinlan A. R. and I.M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26(6):841-842. doi: 10.1093/bioinformatics/btq033

R Development Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ren, Q., T.W. Pfeiffer, and S.A. Ghabrial. 2000. Relationship between soybean pubescence density and soybean mosaic virus field spread. Euphytica 111:191–198.

Romay, M.C., M.J. Millard, J.C. Glaubitz, J.A. Peiffer, K.L. Swarts, T.M. Casstevens, et al. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14:R55-2013-14-6-r55.

Saindon, G., W.E. Beversdorf, and H.D. Voldeng. 1989a. Adjustment of the soybean phenology using the E4 locus. Crop Sci. 29:1361-1365.

Saindon, G., H.D. Voldeng, W.E. Beversdorf, and R.I. Buzzell. 1989b. Genetic Control of Long Daylength Response in Soybean. Crop Sci. 29:1436-14.

Seversike, T.M., J.D. Ray, J.L. Shultz, and L.C. Purcell. 2008. Soybean molecular linkage group B1 corresponds to classical linkage group 16 based on map location of *lf2* gene. Theor. Appl. Genet. 117:143-147.

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

Sidak, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc. 62:626–633.

Sonah, H., L. O'Donoughue, E. Cober, I. Rajcan, and F. Belzile. 2015. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant. Biotechnol. J. 13:211-221.

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

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, and P.B. Cregan. 2013. Development and evaluation of SoySNP50K, a highdensity genotyping array for soybean. PLoS ONE 8:e54985.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Ficus, R.L. Nelson, and P.B. Cregan. 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3 5:1999-2006.

Tang W. T. and Tai G. 1962. Studies on the qualitative and quantitative inheritance of an interspecific cross of soybean, *Glycine max* X *G. formosana*. Bot. Bull. Acad. Sin. 3:39-54.

Tian, Z., X. Wang, R. Lee, Y. Li, J.E. Specht, R.L. Nelson, P.E. McClean, L. Qiu, and J. Ma. 2010. Artificial selection for determinate growth habit in soybean. Proc. Natl. Acad. Sci. USA 107:8563–8568.

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

Todd, J.J. and L.O. Vodkin. 1996. Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. Plant Cell 8:687-699.

Tsubokura, Y, H., Matsumura, M. Xu, B. Liu, H. Nakashima, T. Anai, et al. 2012. Genetic variation in the maturity locus *E4* is involved in adaptation to high latitudes in soybean. Agronomy 3:117-134.

Tsubokura, Y., S. Watanabe, Z. Xia, H. Kanamori, H. Yamagata, A. Kaga, et al. 2013. Natural variation in the genes responsible for maturity loci *E1, E2, E3,* and *E4* in soybean. Annals of Botany 113:429-441.

Turner, S.D. 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *biorXiv*.doi:10.1101/005165.

Tuteja, J.H., G. Zabata, K. Varala, M. Hudson, and L.O. Vodkin. 2009. Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in *Glycine max* seed coats. Plant Cell 21:3063-3071.

Wang, M., N. Jiang, T. Jia, L. Leach, J. Cockram, R. Waugh, et al. 2012. Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theor. Appl. Genet. 124:233-246.

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

Watanabe, S, K.Z.Xia, R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, Y. Nakamoto, et al. 2011. A map-based cloning strategy employing a residual heterozygous line reveals that the *GIGANTEA* gene is involved in soybean maturity and flowering. Genetics 188:395-407.

Weller and Ortega. 2015. Genetic control of flowering time in legumes. Front Plant Sci. 6:207.

Wen, Z., J.F. Boyse, Q. Song, P.B. Cregan, and D. Wang. 2015. Genomic consequences of selection and genome-wide association mapping in soybean. BMC Genomics 16:671.

Wilkerson, G.G., J.W. Jones, K.J. Boote, and G.S. Buol. 1989. Photo-periodically sensitive interval in time to flower of soybean. Crop Sci. 29:721–726.

Woodworth, C. M., 1921 Inheritance of cotyledon, seed-coat, hilum, and pubescence colors in soybeans. Genetics 6:487–553.

Woodworth, C.M. 1932. Genetics and breeding in the improvement of the soybean. Bull. Agric. Exp. Stn. (Illinois) 384:297-404.

Woodworth, C.M. and C. Veatch. 1929. Inheritance of Pubescence in Soy Beans and Its Relation to Pod Color. Genetics. 14(5):512-518.

Xia, Z., S. Watanabe, T. Yamada, Y. Tsubokura, H. Nakashima, H. Zhai, et al. 2012. Positional cloning and characterization reveal the molecular basis for soy bean maturity locus E1 that regulates photoperiodic flowering. PNAS 109:E2155-E2164.

Yan F., S. Di, F.R. Rodas, T.R. Torrico, Y. Murai, T. Iwashina, et al. 2014. Allelic variation of soybean flower color gene *W4* encoding dihdroflavonol 4-reductase 2. BMC Plant Biology 14:58.

Yang, K, N. Jeong, J.-K. Moon, Y.-H. Lee, S,-H. Lee, H.M. Kim, et al. 2010. Genetic analysis of genes controlling natural variation of seed coat and flower colors in soybean. J. Hered. 101:757-768.

Zabala G. and L. O. Vodkin. 2003. Cloning of the pleiotropic *T* locus in soybean and two recessive alleles that differentially affect structure and expression of the encoded flavonoid 3’ hydroxylase. Genetics 163:295–309.

Zabala, G. and L.O. Vodkin. 2005. The *Wp* mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. Plant Cell 17:2619–2632.

Zabala G. and L.O. Vodkin. 2007. A rearrangement resulting in small tandem repeats in the F3’5’H genes of white flower genotypes is associated with the soybean *W1* locus. Plant Gen 2:113–124.

Zabala G. and L.O. Vodkin. 2014. Methylation affects transposition and splicing of a large CACTA transposon from a MYB transcription factor regulating anthocyanin synthase genes in soybean seed coats. PLoS One 9:e111959.

Zhai, H., S. Lu, Y. Wang, X. Chen, H. Ren, J. Yang, et al. 2014. Allelic variation at four major maturity *E* genes and transcriptional abundance of the E1 gene are associated with flowering time and maturity of soybean cultivars. PLoS One 9:e97636.

Zhao, C., R. Takeshima, J. Zhu, M. Xu, M. Sato, S. Watanabe, et al. 2016. A recessive allele for delayed flowering at the soybean maturity locus E9 is a leaky allele of FT2a, a FLOWERING LOCUS T ortholog. BMC Plant Biology 16:20.

3.11. FIGURES

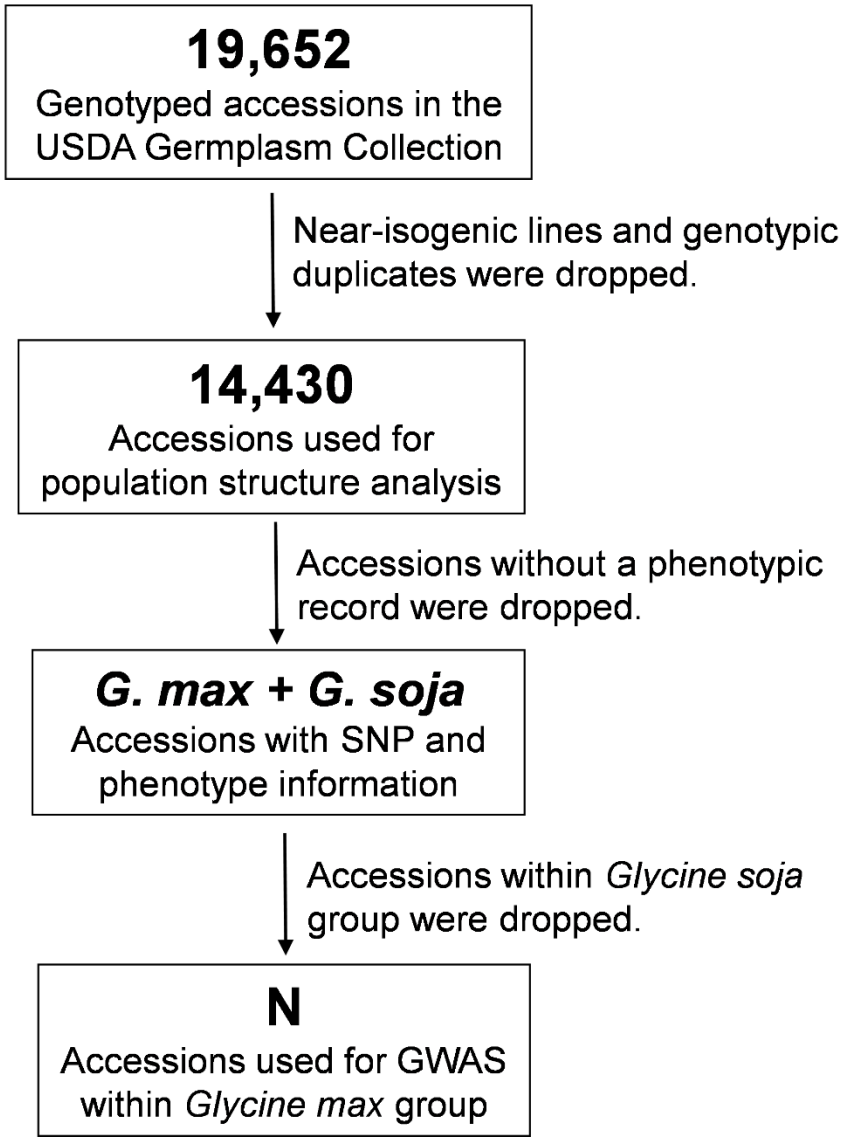


Fig. 2. The stepwise filtering of *G. max* accessions held in the USDA Germplasm Collection that was conducted prior to the genome-wide association mapping of ten descriptor traits.

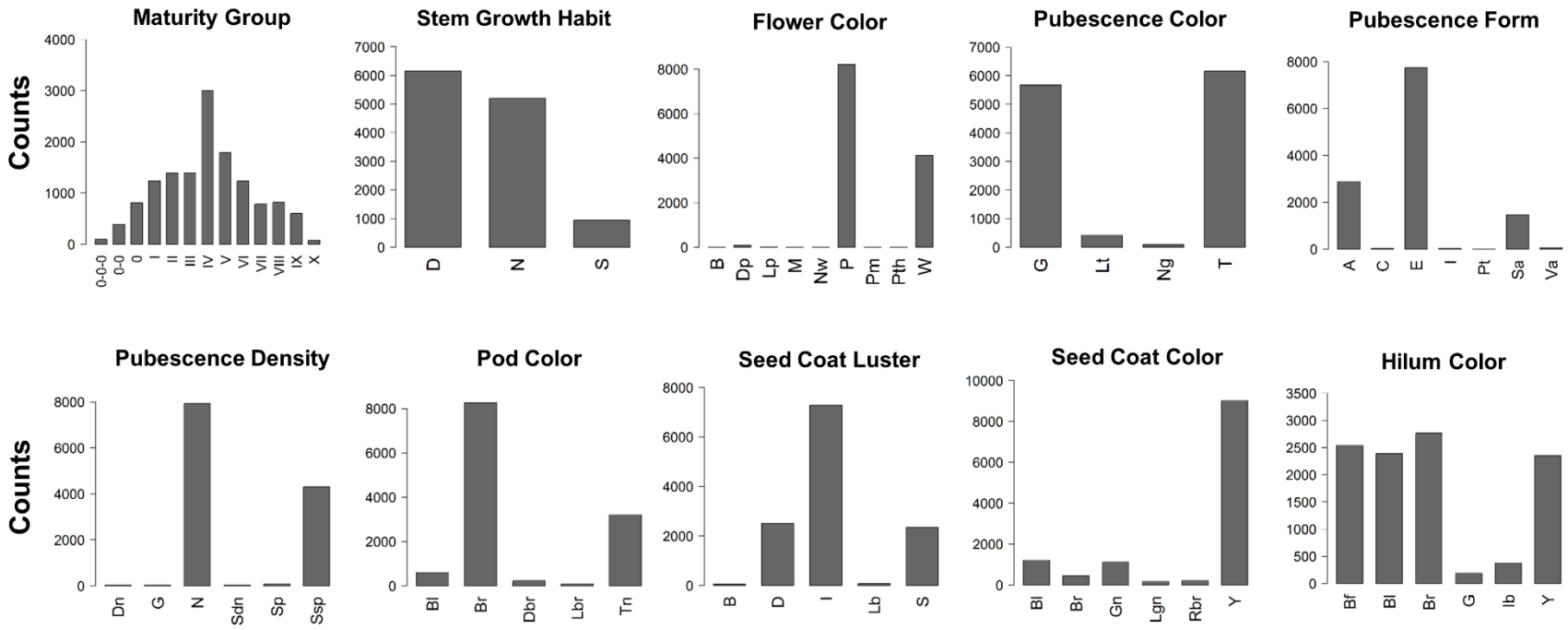


Fig. 2 Frequency distribution of the soybean accessions for the various phenotypic variants available in each of the ten descriptive traits that were used in the initial genome-wide analysis conducted on each trait (i.e., **N** value in **Fig. 1** for the *G. max* accession).

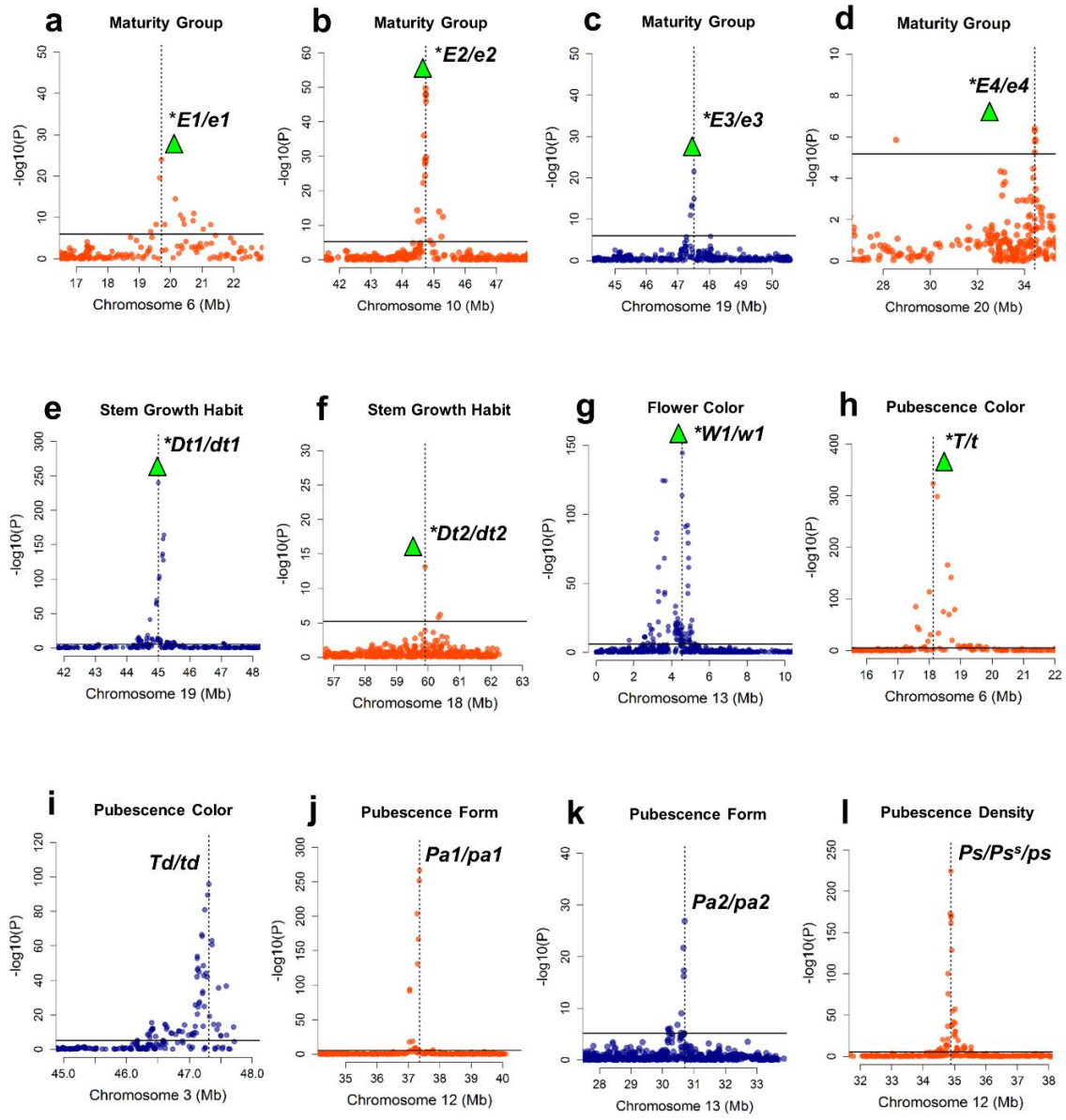


Fig. 3 SNP association signal $-\log_{10}(P)$ values corresponding with trait-controlling classical genes are plotted against the physical position (bp) on the specific chromosomes in these panels: **a-d** Maturity Group, **e–f** Stem Term Type, **g** Flower Color, **h–i** Pubescence Color, **j–k** Pubescence Form, **l–n** Pubescence Density, **o–p** Pod Color, **q** Seed Coat Luster, **r** Seed Coat Color, and **s–u** Hilum Color. An asterisk identifies the cloned classical gene loci. Green bars depict the (Table 1) bp positions of the Glyma gene-coding sequences (with a red bar depicting the Glyma gene that is inferred to be the *L1/l1* locus). The solid horizontal line denotes the calculated threshold value ($-\log_{10} P > 5.17$) for declaring a significant association. The dashed vertical line denotes the bp position of the SNP that had the maximum $-\log_{10}P$ value for the given classical locus.

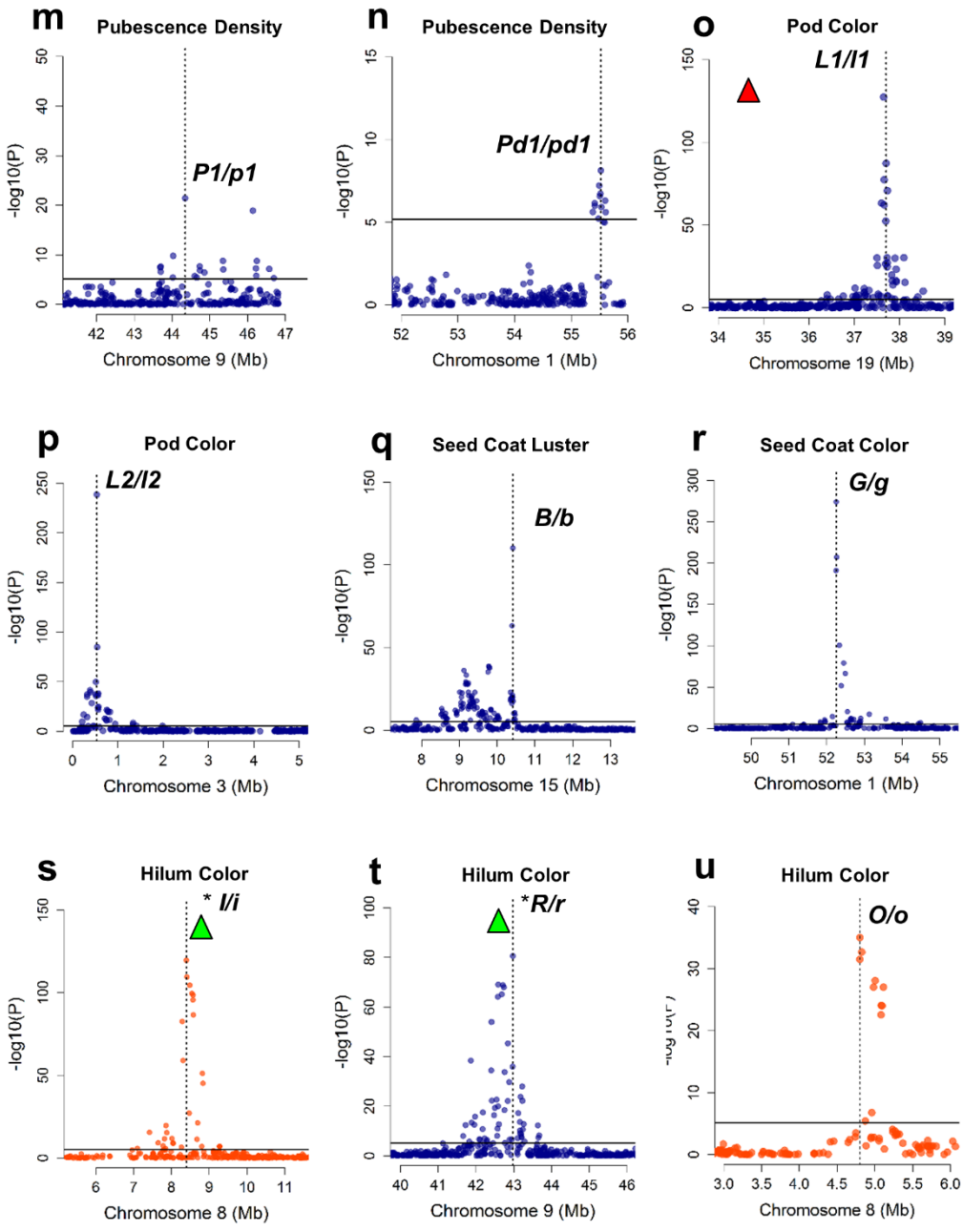


Fig. 3.3 (*Continued*).

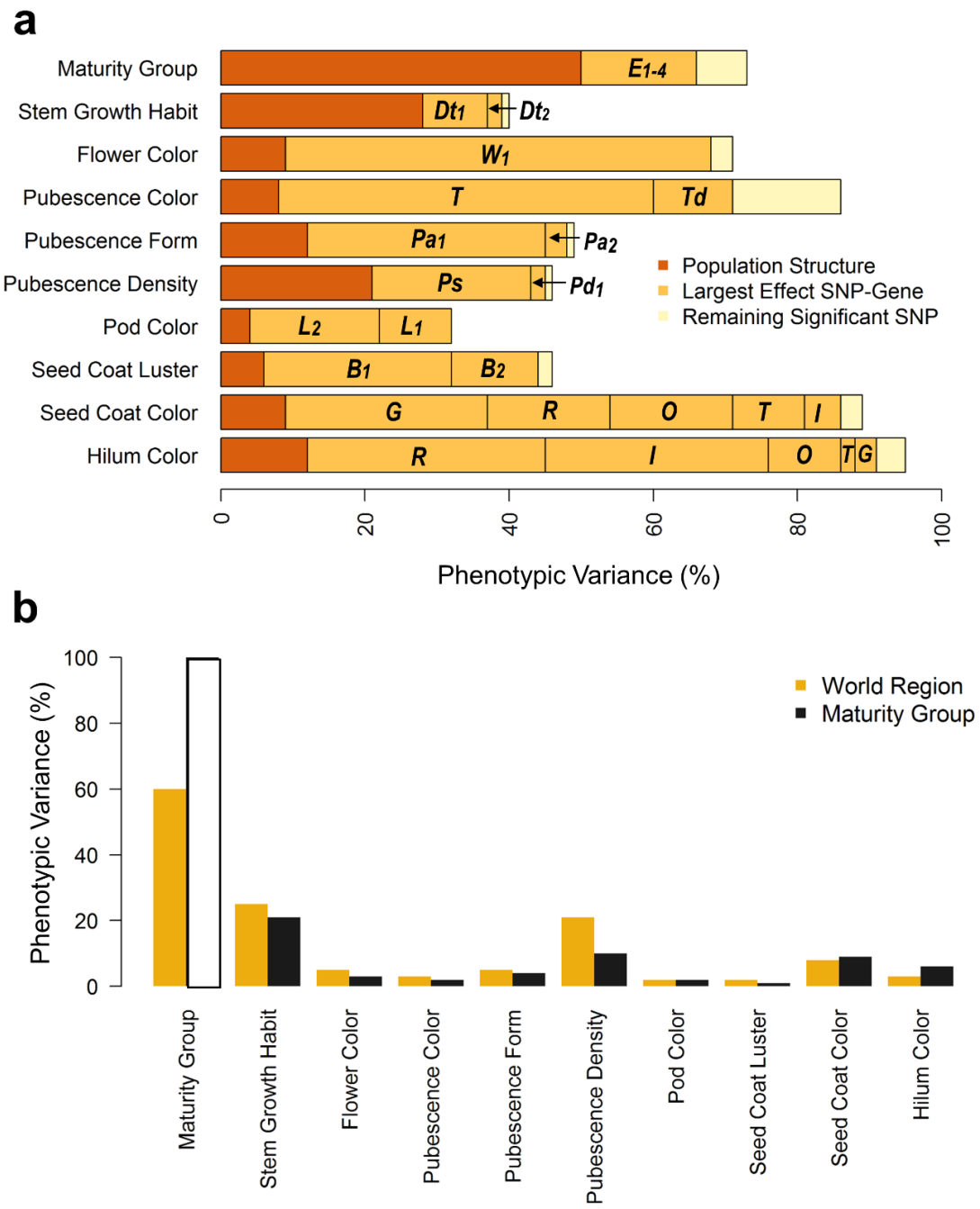


Fig. 4 (a) Contributions of significant SNPs and population structure (defined by ADMIXTURE K=5) to phenotypic variance of each of the ten descriptive traits. The proportion of phenotypic variance accounted for by significant SNPs was partitioned into largest effect SNPs (tagging known/candidate genes) and small effect SNPs and calculated after accounting for population structure effects, **(b)** Contribution of world region and maturity group, which are the major determinant of population of structure within the collection, to trait phenotypic varia

Table 1. Summary of SNP-association signals that exceeded an experiment-wise significance criterion of -logP > 5.17 in a genome-wide association analysis (GWA) using the Q+K model performed on 13,624 G. max accessions for each of the below-listed ten soybean descriptor traits. The significant SNP-associations detected in this study are ordered by trait, then by chromosome, and thence within each chromosome according to the base-pair regions of the significant SNPs. Data in each table row originates from the underscored supplemental figure (i.e., GWA analysis Manhattan plot).

				Significant SNP - Trait Associations‡						Cloned Gene Information§					
Descriptor Trait	**Supp. Fig. No.†**	**Chr**	**LG**	**SNPs**	**First**	**Last**	**Max**	**-logP**	**Gene Locus¶**	**Glyma Name**	**Other name**	**Start**	**End**	**Size**	**References for the cloned genes, or for-non cloned genes**
				- no-	---------- bp ----------							---------- bp ----------			
Maturity group	S1a, S1b	6	C2	16	**18724519**	**21426047**	21273525	23.97	***E1/e1*** ¶	Glyma06g23026	RPR	**20006973**	**20007810**	838	Xia et al. (2012)
	S1a, S1b	8	A2	1	3642671	3642671	3642671	5.19							
	S1a, S1b, S1c	10	O	20	**44476584**	**45294441**	44743315	49.73	***E2/e2*** ¶	Glyma10g36600	GmGIa	**44716720**	**44738268**	2154	Watanabe et al. (2011)
	S1a, S1b, S1c	11	B1	4	10721006	11572077	11269310	18.83						9	SoyBase Pod Maturity QTL 17-2 (Lu et al., 2015)
	S1a, S1b, S1c, S1e	12	H	6	5491240	5786241	5491240	14.56							SoyBase Pod Maturity QTL 26-2 (Lu et al., 2015)
	S1a, S1b	13	F	1	36616135	36616135	36616135	5.69							
	S1b	18	G	1	59902680	59902680	59902680	7.02							SoyBase Pod Maturity QTL29-8 (Wen et al. 2015)
	S1a, S1b, S1c	19	L	7	**47270486**	**48037479**	47510130	21.54	***E3/e3*** ¶	Glyma19g41210	GmPhyA3	**47511246**	**47519957**	8712	Watanabe et al. (2009)
	S1d	20	I	2	3145294	3150963	3150963	5.36							
	S1d	20	I	1	28550287	28550287	28550287	5.85							
	S1d	20	I	5	34434402	34462359	34437459	6.39	***E4/e4*** ¶	Glyma20g22160	GmPhyA2	32087580	32093266	5687	Liu et al. (2008)
Stem term type	S2b, S2c	6	C2	1	38948190	38948190	38948190	6.80							
	S2a, S2b	18	G	4	**59902680**	**60380782**	59902680	17.27	***Dt2/dt2*** ¶	Glyma18g50910	Loc100788956	**59918841**	**59927027**	8187	Ping et al. (2014)
	S2a	19	L	1	43080829	43080829	43080829	5.26							
	S2a, S2b, S2c	19	L	55	**44329464**	**45525374**	45000827	238.79	***Dt1/dt1*** ¶	Glyma19g37890	PEPB; TFL1	**44979743**	**44981385**	1643	Liu et al. (2010); Tian et al. (2010)

	S2a	19	L	2	47069443	47076497	47069443	6.22	***E3/e3*** ¶	Glyma19g41210	GmPhyA3	47511246	47519957	8712	Watanabe et al. (2009)
Flower color	S3a, S3b	13	F	2	1629730	1756025	1756025	8.05							
	S3a, S3b	13	F	20	2493212	3044754	2833623	24.25							
	S3a, S3b	13	F	23	3175654	3825644	3657853	152.79							
	S3a, S3b	13	F	29	**4173955**	**5480962**	4559799	169.79	***W1/w1*** ¶	Glyma13g04210	*W1*	**4552711**	**4557371**	4661	Zabala and Vodkin (2007); Sonah et al. (2014); Wen et al. (2015)
	S3b	19	L	1	36603029	36603029	36603029	14.81	***L1/l1***	Glyma19g27460		34750891	34752190	1300	Bernard (1967); He et al. (2015)¶
Pubescence color	S4a, S4b	3	N	65	46332845	47702654	47307916	95.83	***Td/td***						Bernard (1975a); Behm et al. (2013); Wen et al. (2015)
	S4a, S4b	6	C2	26	**17258654**	**19815389**	17567713	298.24	***T/t*** ¶	Glyma06g21920	*T*	**18534619**	**18541464**	6846	Zabala & Vodkin (2002); Sonah et al. (2014); Wen et al. (2015)
	S4a	6	C2	1	25762003	25762003	25762003	7.45							
	S4a	6	C2	2	30106667	30163816	30163816	7.08							
	S4b	6	C2	1	38948190	38948190	38948190	5.23							
	S4a	12	H	1	18537212	18537212	18537212	8.24							
	S4a	12	H	1	21317830	21317830	21317830	9.06							
	S4a, S4b	14	B2	1	46968410	46968410	46968410	5.40							
	S4a	20	I	1	13081929	13081929	13081929	8.45							
Pubescence form	S5a, S5b	12	H	16	37036017	37786243	37356120	251.32	***Pa1/pa1***						Bernard (1975b); Lee et al. (1999)
	S5a, S5b	13	F	11	30181642	30708708	30708708	26.87	***Pa2/pa2***						Bernard (1975b); Lee et al. (1999)
Pubescence density	S6a, S6b, S1c	1	D1a	4	55493281	55523014	55523014	8.14	***Pd1/pd1***						Bernard and Singh (1969); PDens. QTL 1-1 (Komatsu et al., 2007)
	S6a	7	M	1	17153201	17153201	17153201	5.31							
	S6a, S6b, S1c	12	H	53	34477297	35525603	34877806	224.28	***Ps/Ps^s/ps***						Bernard (1975d); Pub. Density QTL 1-2 (Komatsu et al., 2007)
	S6a, S6b	14	B2	1	4934894	4934894	4934894	69.92							
	S6c	9	K	11	43686430	44855340	44348623	21.41	***P1/p1***						Bernard and Singh (1969); Zabala and Vodkin (2014)
	S6c	9	K	8	45344827	46694731	46139114	18.93							
Pod color	S7a	1	D1a	2	52253980	52263952	52253980	6.38							

	S7a, S7b	3	N	27	246658	1338018	537774	238.55	***L2/l2***						Bernard (1967)
	S7a, S7c	19	L	48	36397778	38521183	37649854	127.49	***L1/l1***	Glyma19g27460		34750891	34752190	1300	Bernard (1967); He et al. (2015)
Seed coat luster	S8b	8	A2	1	6897932	6897932	6897932	6.25							
	S8b	8	A2	7	**8272057**	**8656325**	8462762	12.78	***I/i^{k}/i^{i}/i*** ¶	Inverted Repeats #	*CHS1-3-4* #	**8462596**	**8469679**	7084	Todd and Vodkin (1996); Tuteja et al. (2009#**)**
	S8a, S8b, S8c	9	K	1	1456482	1456482	1456482	39.11	***B?/b?***						Woodworth (1932); proposed the locus symbols (present study)
	S8b	11	B1	1	15920433	15920433	15920433	6.80							
	S8a	13	F	1	34089340	34095060	34095060	10.51	***B1/b1***						Woodworth (1932); Tang and Tai (1962); Chen & Shoe (1998)
	S8a, S8b, S8c	15	E	17	**7861342**	**8941824**	8893988	9.26	***Hps*** ††	Glyma15g11970	*Hps*††	**8868741**	**8875714**	6974	Gijzen et al. (1999, 2003, 2006††)
	S8a, S8b, S8c	15	E	62	9516289	10475708	10416352	110.19	***B/b***						Lorenzen et al. (1995); Gijzen et al. (2003)
Seed coat color	S9a, S9b	1	D1a	30	52253980	52493627	52253980	274.08	***G/g***						Woodworth (1921)
	S9a	6	C2	5	**18033759**	**18810733**	18766611	11.12	***T/t*** ¶	Glyma06g21920	*T*	**18534619**	**18541464**	6846	Zabala and Vodkin (2002)
	S9a	8	A2	7	4802080	5113384	5003648	8.64	***O/o***						
	S9a	8	A2	16	**8013021**	**9120830**	8462762	37.47	***I/i^{k}/i^{i}/i*** ¶	Inverted Repeats #	*CHS1-3-4* #	**8462596**	**8469679**	7084	Todd and Vodkin (1996); Tuteja et al. (2009#)
	S9a	9	K	1	42974503	42974503	42974503	5.80	***R/r^{m}/r*** ¶	Glyma09g36983	*R2R3 MYB*	42562649	42564660	2012	Gillman et al. (2011); Zabala and Vodkin (2014§)
Hilum color	S10a	6	C2	17	**17567713**	**19540686**	18766611	73.52	***T/t*** ¶	Glyma06g21920	*T*	**18534619**	**18541464**	6846	Zabala and Vodkin (2002)
	S10c	8	A2	11	4800584	5113384	4802080	35.00	***O/o***						
	S10a, S10b, S10c	8	A2	20	**7103134**	**8836971**	8396392	119.55	***I/i^{k}/i^{i}/i*** ¶	Inverted Repeats #	*CHS1-3-4* #	**8462596**	**8469679**	7084	Todd and Vodkin (1996); Tuteja et al. (2009#)
	S10a, S10c	8	A2	3	9271761	9274750	9274750	7.16							
	S10b	9	K	55	**41660046**	**43669720**	42974503	80.58	***R/r^{m}/r*** ¶	Glyma09g36983	*R2R3 MYB*	**42562649**	**42564660**	2012	Gillman et al. (2011); Zabala and Vodkin (2014)
	S10b	12	H	1	487052	487052	487052	7.57							

† Listed in this column are the supplemental figure numbers and panel labels (i.e., a, b, c, etc.) displaying the GWA Manhattan plots generated for each trait. Underscoring denotes the supplemental figure (and the panel therein) of the GWA data presented in the adjacent table row.

‡ Listed in these table columns are the detected number of significant SNPs within a contiguous region, the bp positions of first and last SNP spanning that region, and the bp position of the SNP in that region that exhibited maximum -logP value and if underscored, corresponding table row data were used for the magnified single-chromosome GWA Manhattan plot presented in the Fig. 3 panels.

§ Listed in these table columns is information for genes that have been cloned to date, and includes the cloned gene Glyma name, plus any other name, the start and stop bp positions of the coding sequence, and the gene bp length. If the cloned gene start-stop bp positions were bracketed by the first-last bp positions of GWA SNPs, the bp regions of the former and latter were **bold-faced**.

¶ Just these classical loci have been cloned to date. The *L1/l1* locus was not footnoted in this table because it was fine-mapped by He et al. (2015) to a Chr 19 region of 13 potential candidate genes. Of those 13, Glyma gene listed in this table for *L1/l1* was *inferred* by He et al. (2015) as the causal gene, but that inference has not yet been *experimentally* verified.

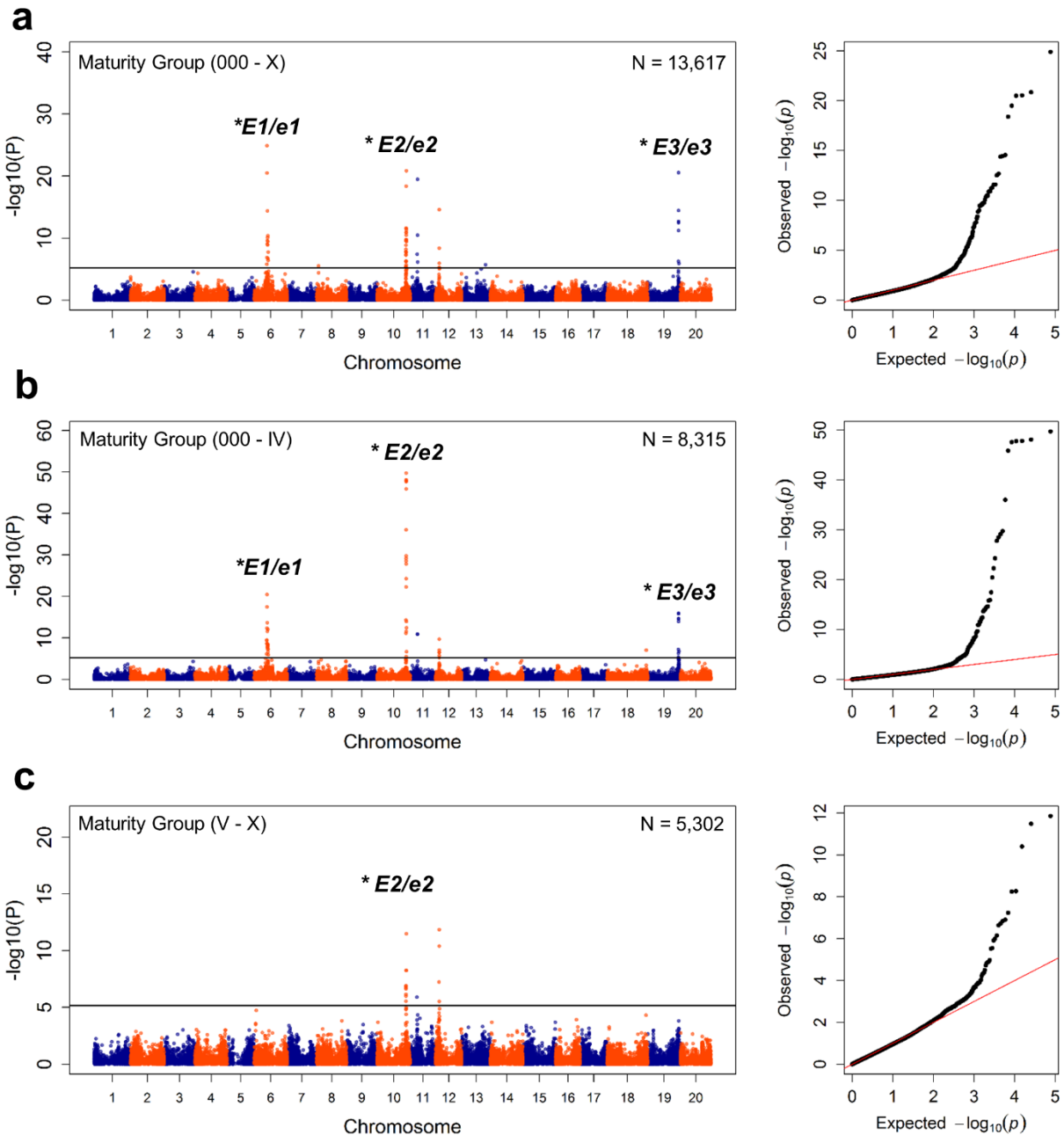
The dominant alleles at the $I/i^{k}/i^{i}/i$ locus consist of two inverted repeats of the three contiguously arranged chalcone synthetase genes of CHS1, CHS2, and CH3, so a single Glyma number is not available for this locus. For specific details, see the Tuteja et al. (2009) reference.

†† The *Hps* gene consists of a tandem array of a reiterated 8.6 kb coding units - each unit is a single open reading frame for the HPS protein. A null *Hps* allele has not been identified, but the copy number variants constitute multiple alleles (i.e., many copies in dull seed coat genotypes, fewer copies in shiny seed coat genotypes). For specific details, see the Gijzen et al. (2006) reference.

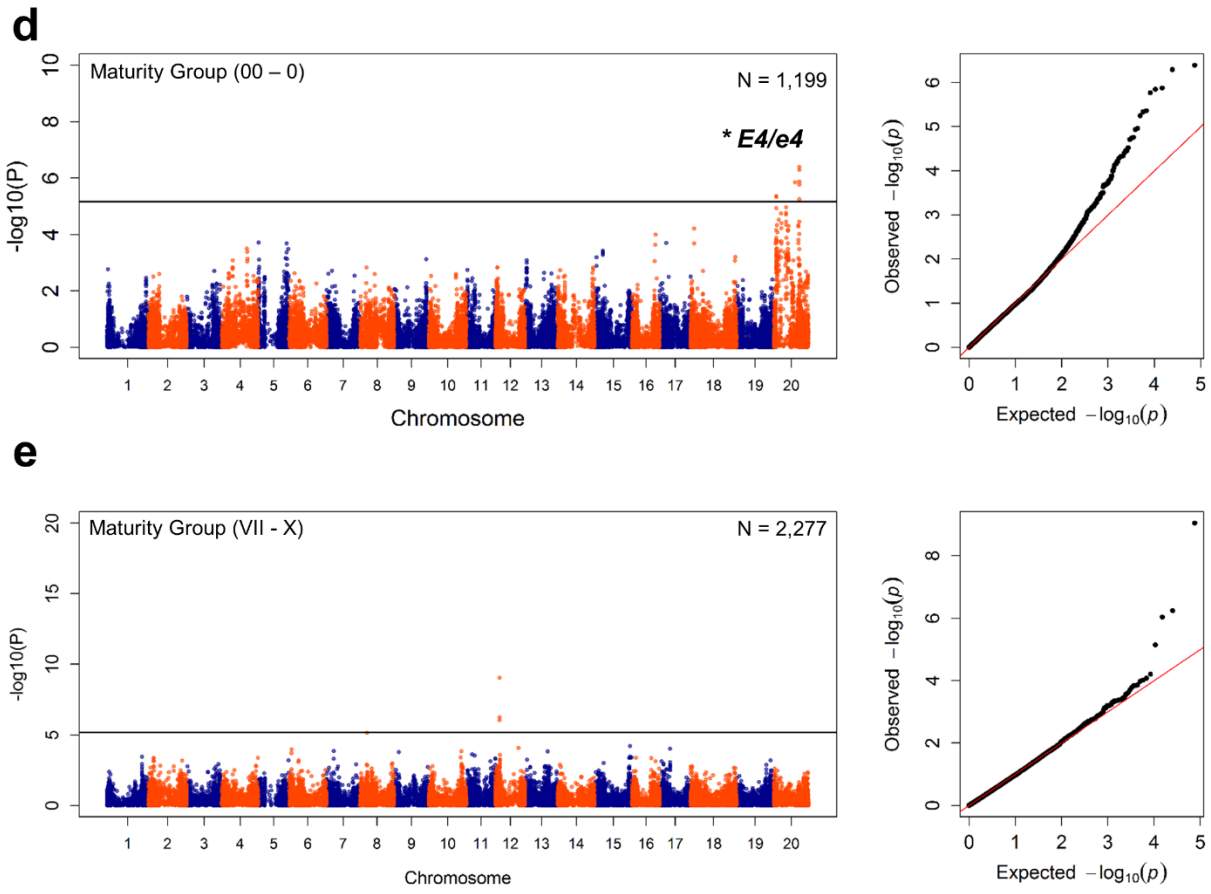
3.12. APPENDIX

The below texts applies to the figure captions for S1 – S10.

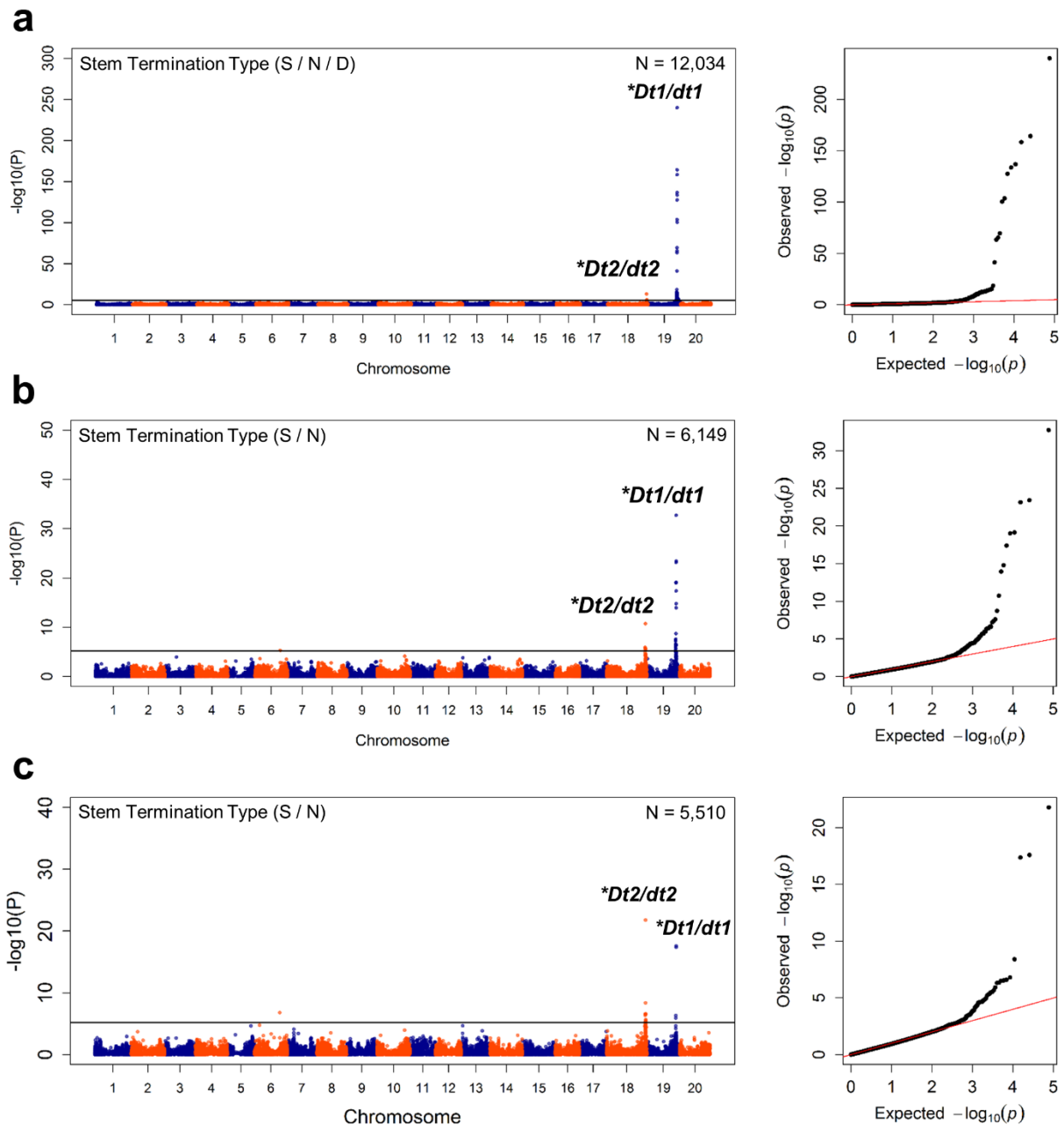
Markers are plotted on the x-axis according to their chromosomal physical position. The solid horizontal line indicates the calculated threshold value for declaring a significant association. Each peak significant SNP association is denoted by dashed vertical line. Qualitative genes (i.e., dominant/recessive alleles, asterisked if cloned) known to control the descriptor trait are depicted at genomic regions showing strong association signals.



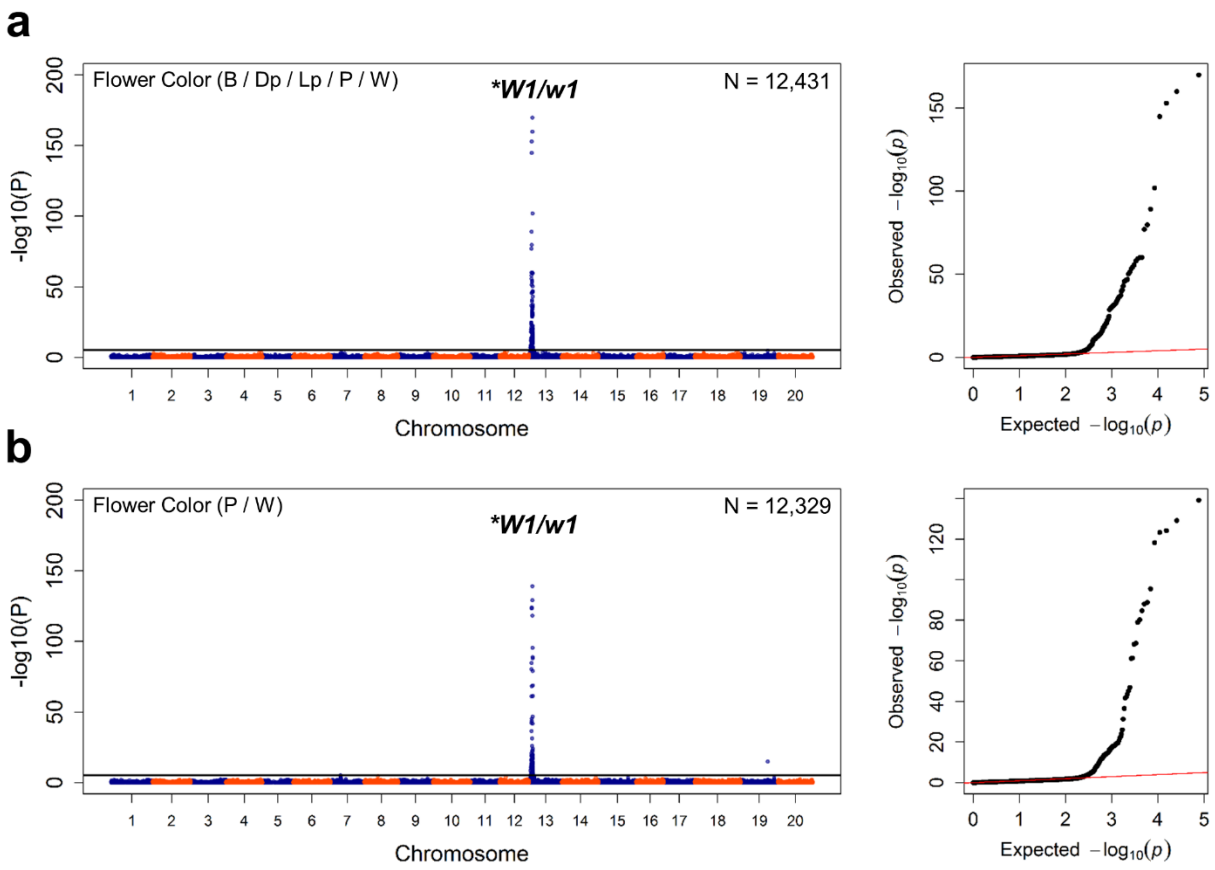
Supplementary Figure S1. Genome-wide association mapping for Descriptor 1 – Maturity Group. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$ (P value) (right) is shown for **(a)** all 13 MG categories (i.e., 000 - X), **(b)** only the northern seven MG categories (000 - IV), and **(c)** only the southern MG categories (V - X).



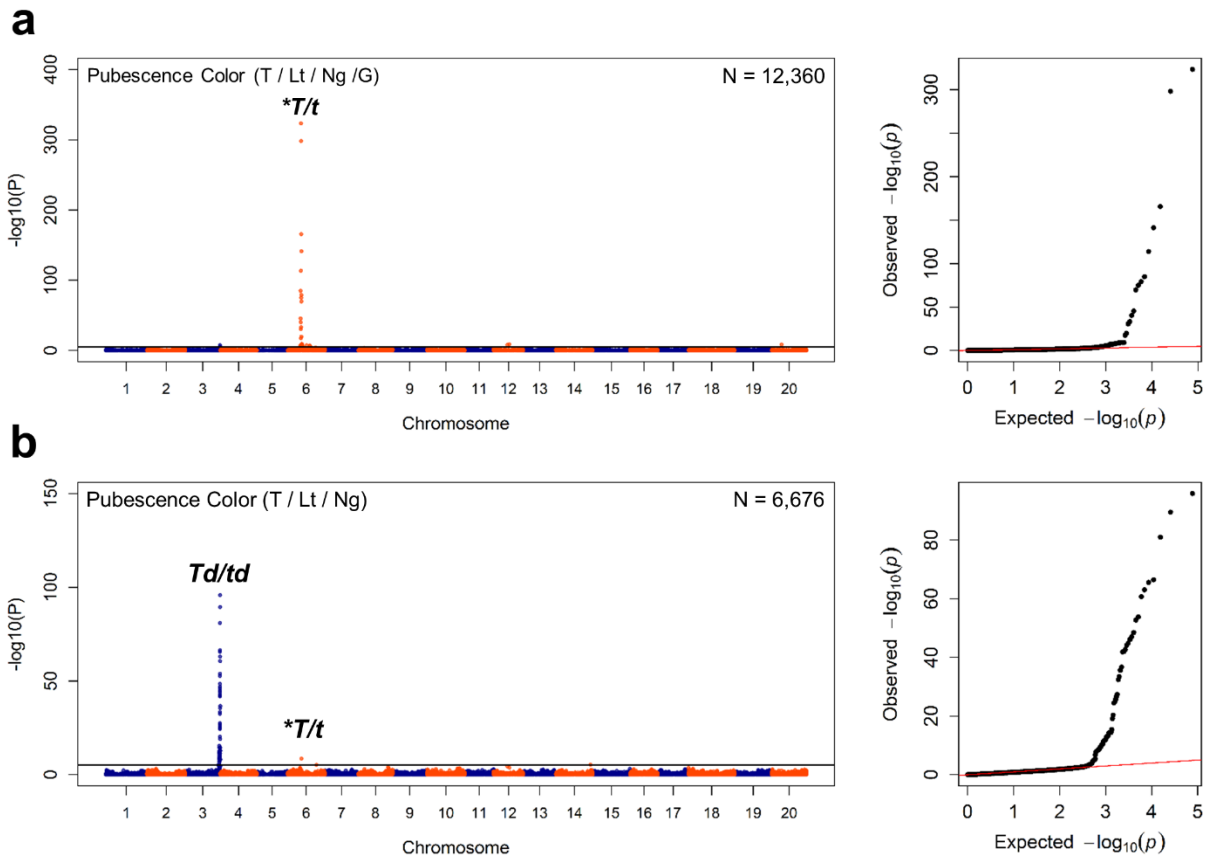
Supplementary Figure S1 (*Continued*). Genome-wide association mapping for Descriptor 1 - Maturity Group. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$(P value) (right) is shown for **(d)** early MG categories (000, 00, 0) and **(e)** late MG categories (VII – X).



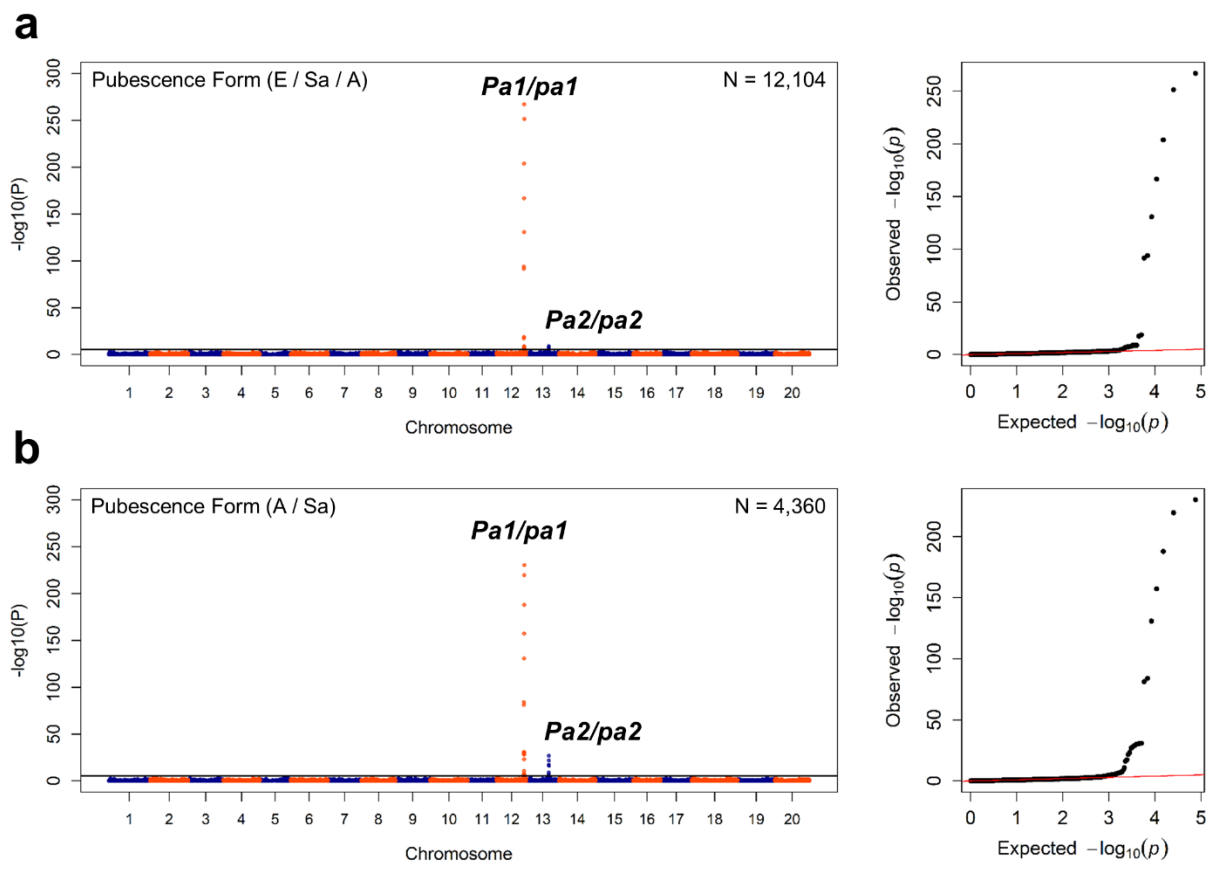
Supplementary Figure S2. Genome-wide association mapping for Descriptor 2 - Stem Termination Type. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$ (P value) (right) is shown for **(a)** all phenotype categories of semi-determinate (S) / indeterminate (N) / determinate (D), **(b)** only the two categories of S / N, and **(c)** with the two categories of S / N corrected for classification error.



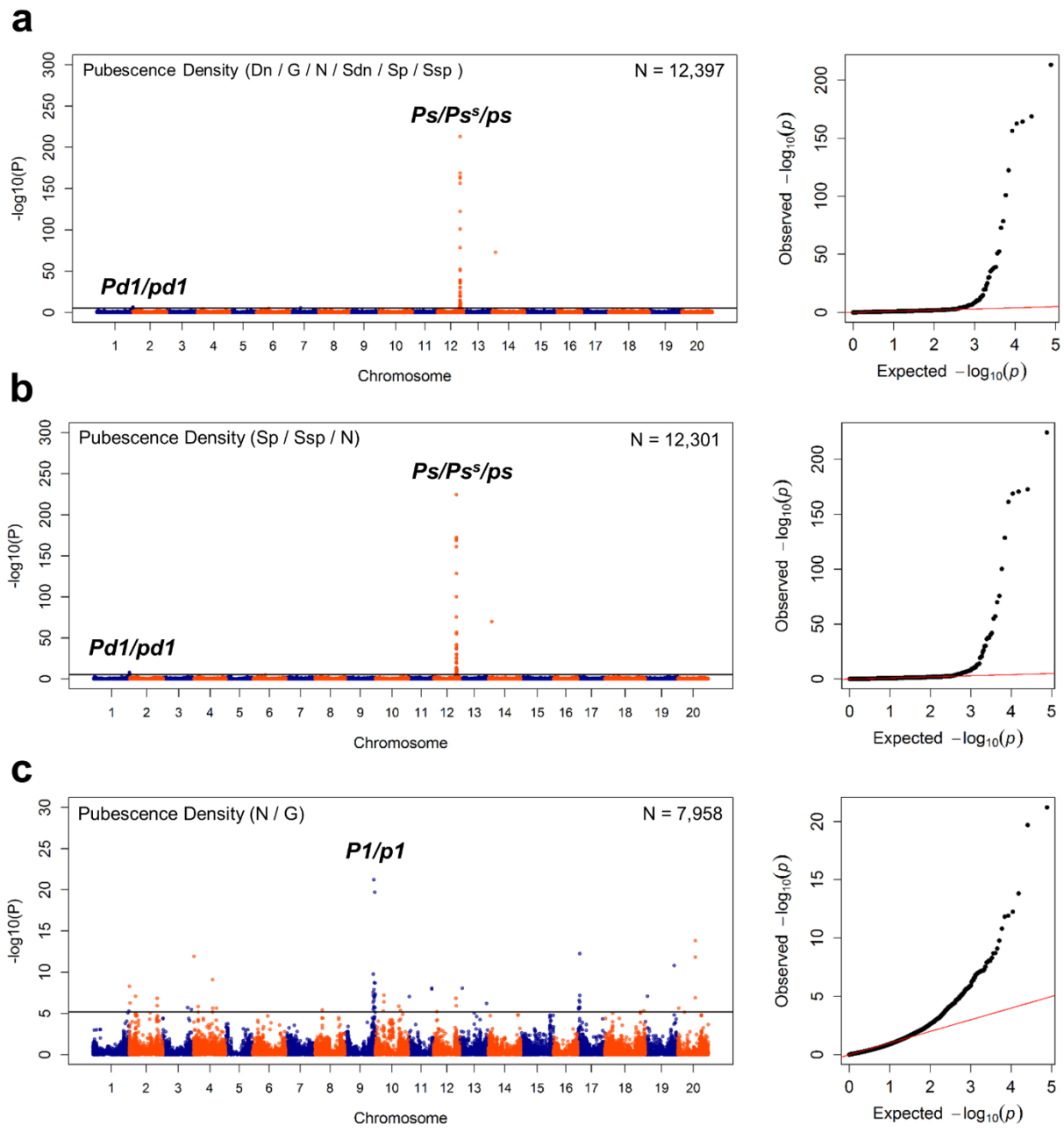
Supplementary Figure S3. Genome-wide association mapping for Descriptor 3 - Flower Color. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$ (P value) (right) is shown for **(a)** all phenotypic categories of purple (P) / white (W) / others, and **(b)** only the two categories of P / W.



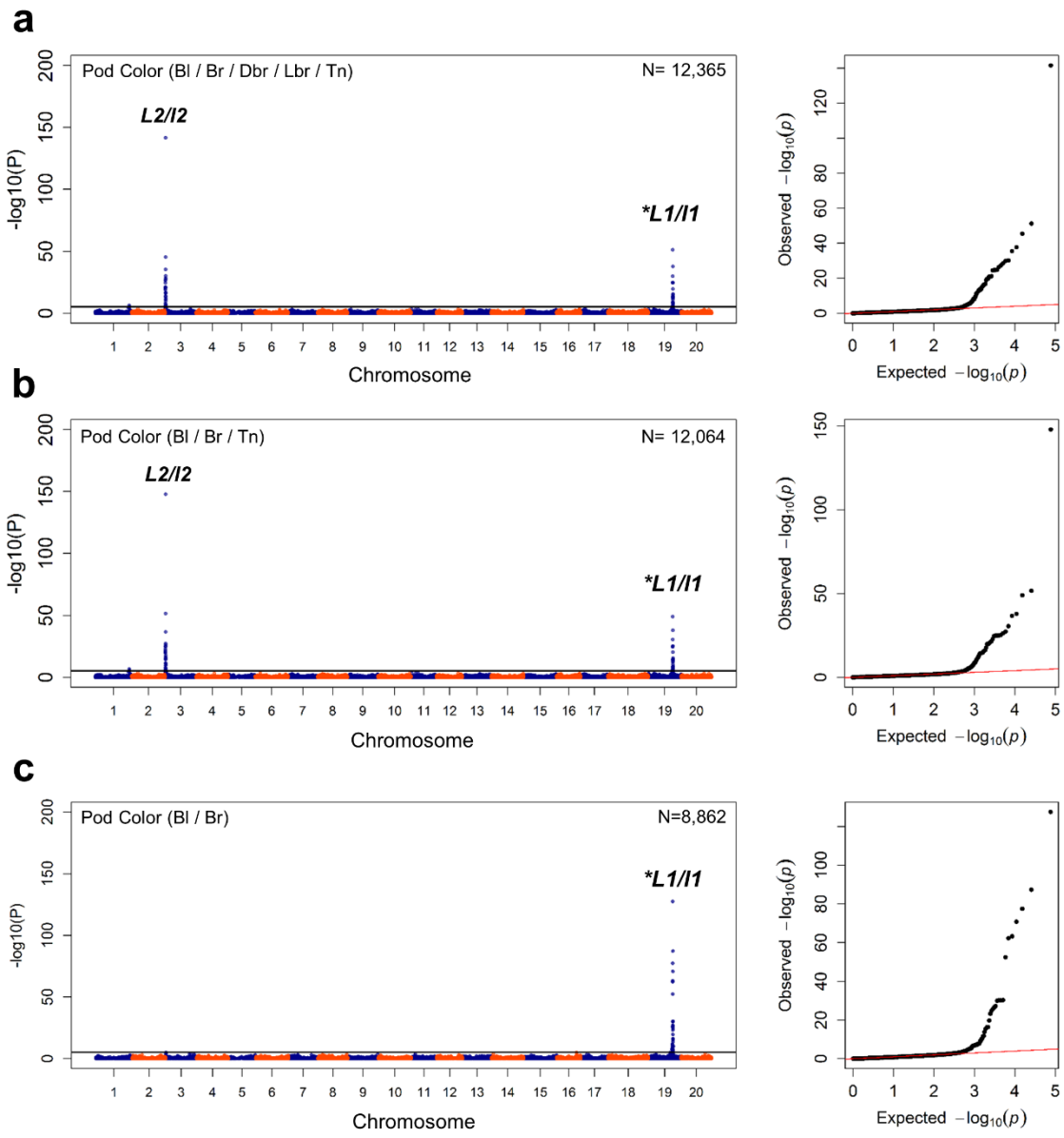
Supplementary Figure S4. Genome-wide association study for Descriptor 4 - Pubescence Color. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$ (P value) (right) is shown for **(a)** all phenotypic categories of tawny (T) / light tawny (Lt) / near grey (Ng) / grey (G), and **(b)** only the categories of T / Lt / Ng.



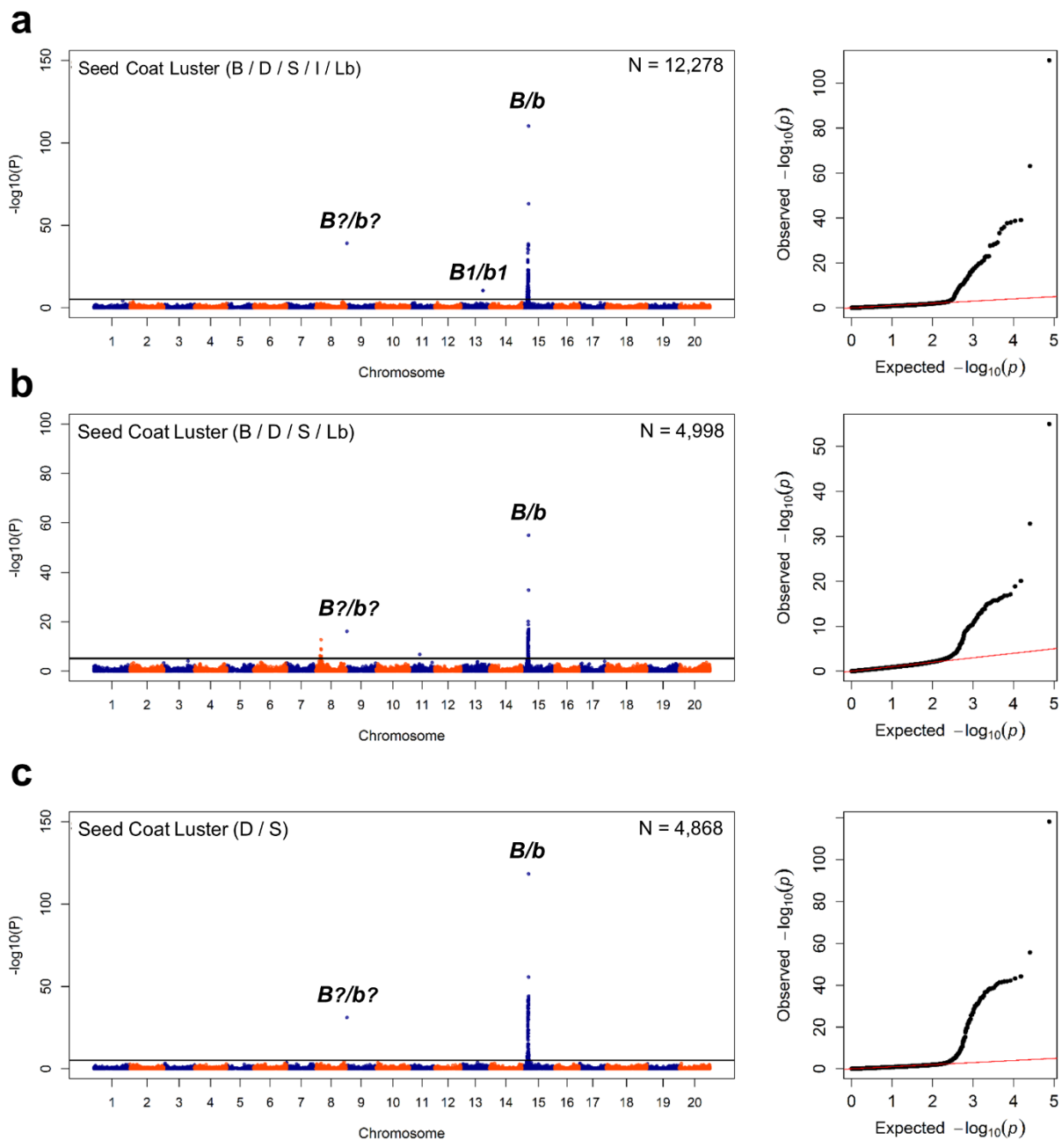
Supplementary Figure S5. Genome-wide association study for Descriptor 5 - Pubescence Form. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$ (P value) (right) is shown for **(a)** all phenotypic categories of erect (E) / semi-appressed (S) / appressed (A) and **(b)** only the categories of Sa / A.



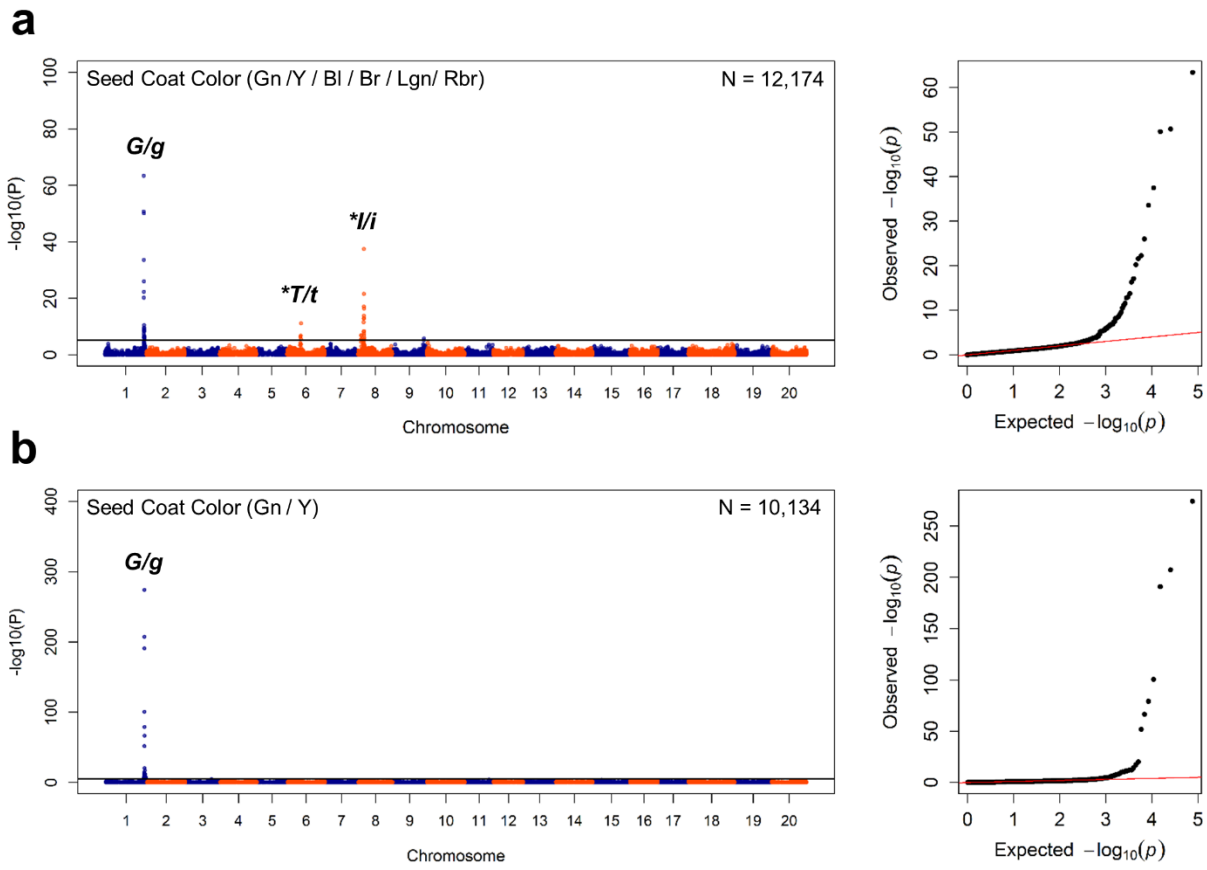
Supplementary Figure S6. Genome-wide association mapping for Descriptor 6 - Pubescence Density. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$ (P value) (right) is shown for **(a)** all phenotypic categories of dense (Dn) / glabrous (G) / normal (N) / semi-dense (Sdn) / sparse (Sp) / semi-sparse (Ssp) , **(b)** only the categories of Sp / Ssp / N, and **(c)** only the categories of N / G.



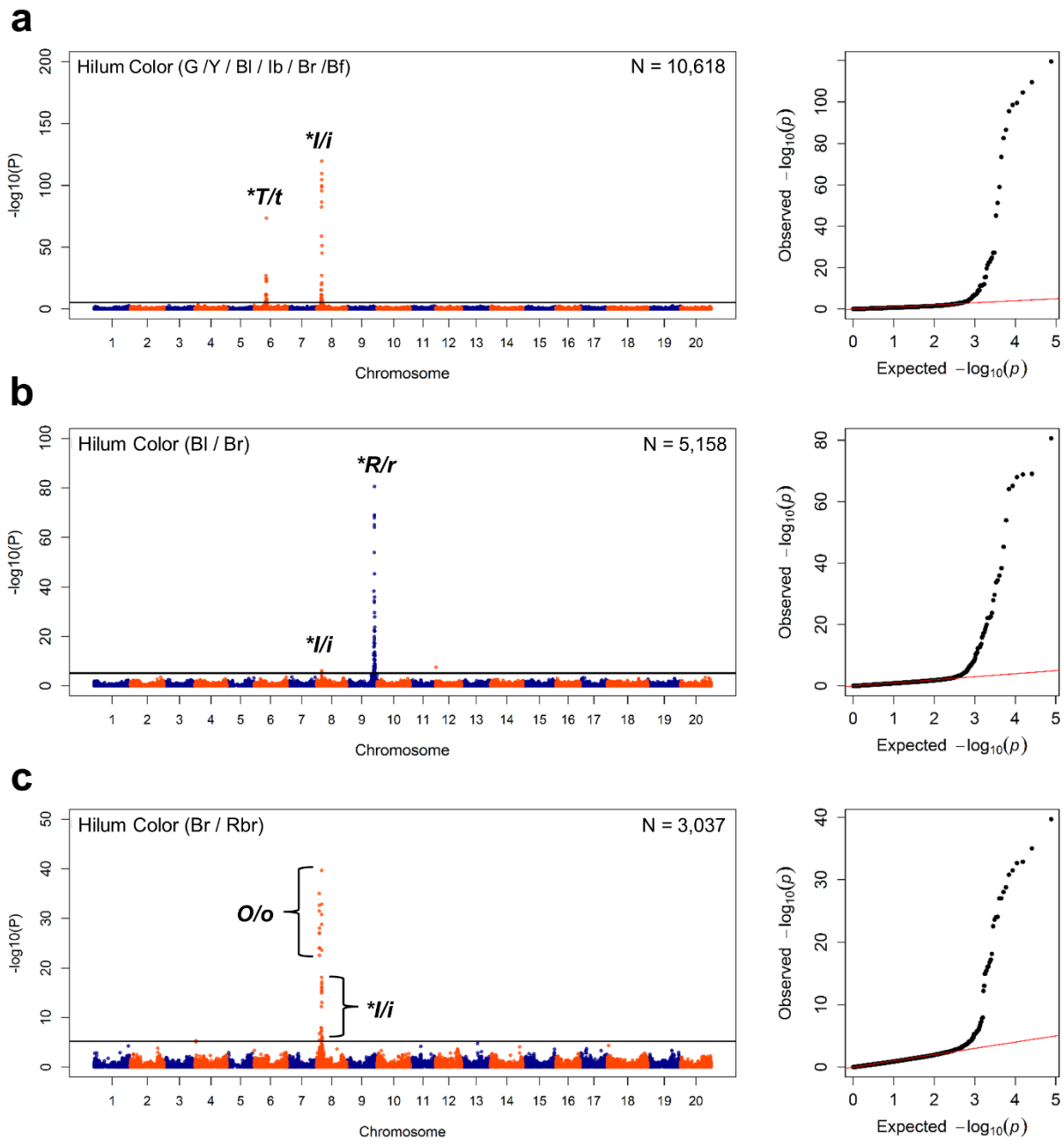
Supplementary Figure S7. Genome-wide association mapping for Descriptor 7 - Pod Color. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$ (P value) (right) is shown for **(a)** all phenotypic categories of black (Bl) / brown (Br) / dark brown (Dbr) / light brown (LBr) / tan (Tn), **(b)** only the categories of Bl / Br / (Tn), and **(c)** only the categories of Bl / Br



Supplementary Figure S8. Genome-wide association mapping for Descriptor 8 - Seed Coat Luster. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$ (P value) (right) is shown for **(a)** all phenotypic categories of bloom (B) / dull (D) / intermediate (I) / shiny (S) / light bloom (Lb), **(b)** only the categories of B / D / S / Lb, and **(c)** only the categories of D / S.



Supplementary Figure S9. Genome-wide association mapping for Descriptor 9 - Seed Coat Color. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$ (P value) (right) is shown for **(a)** all phenotypic categories of green (Gn) / yellow (Y) / black (Bl) / brown (Br) / light green (Lgn) / red brown (Rbr), and **(b)** only the categories of Y / G.



Supplementary Figure S10. Genome-wide association mapping for Descriptor 10 - Hilum Color. A Manhattan plot (left) and quantile-quantile plot of -$\log_{10}$ (P value) (right) is shown for **(a)** all phenotypic categories of gray (G) / yellow (Y) / black (Bl) / brown (Br) / imperfect black (Ib) / brown (Br) / buff (Bf), **(b)** only the categories of Bl / Br, and **(c)** only the categories of Br / Rbr.

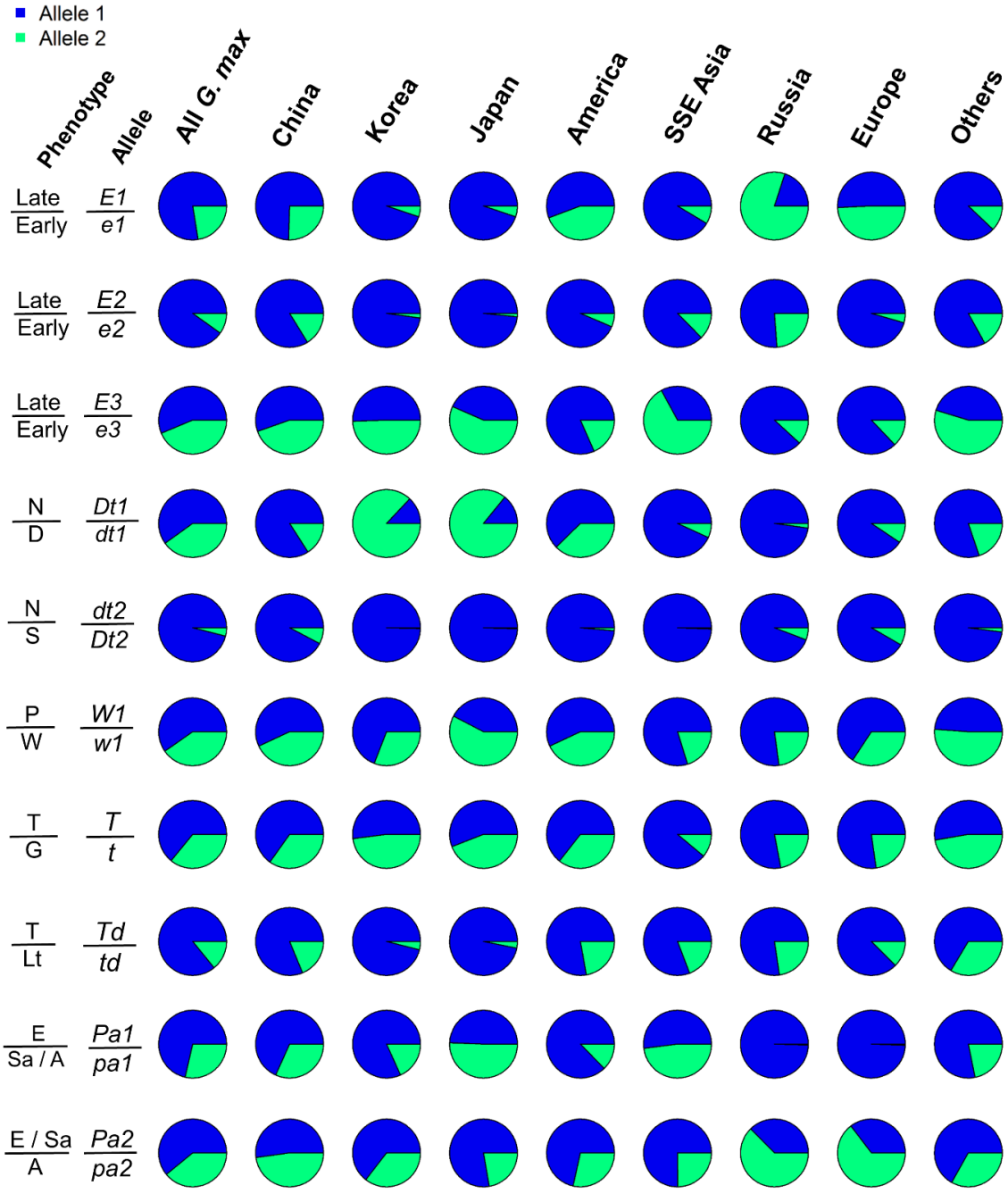


Figure S11. (2-page layout) Global distributions of allelic variation across 13,624 *G. max* accessions, with each subpopulation defined as a specific world region. The first two columns list the locus allelic types and corresponding phenotypes. The remaining columns display pie charts showing the allele frequencies of 20 known gene genes in *G. max* accessions overall, and with respect to eight defined world regions. The pie chart frequencies of two main allelic types of each gene locus are generically colored blue (Allele 1) and light green (Allele 2) in pie chart and correspond on the left to the specific allele symbols of the 20 loci.

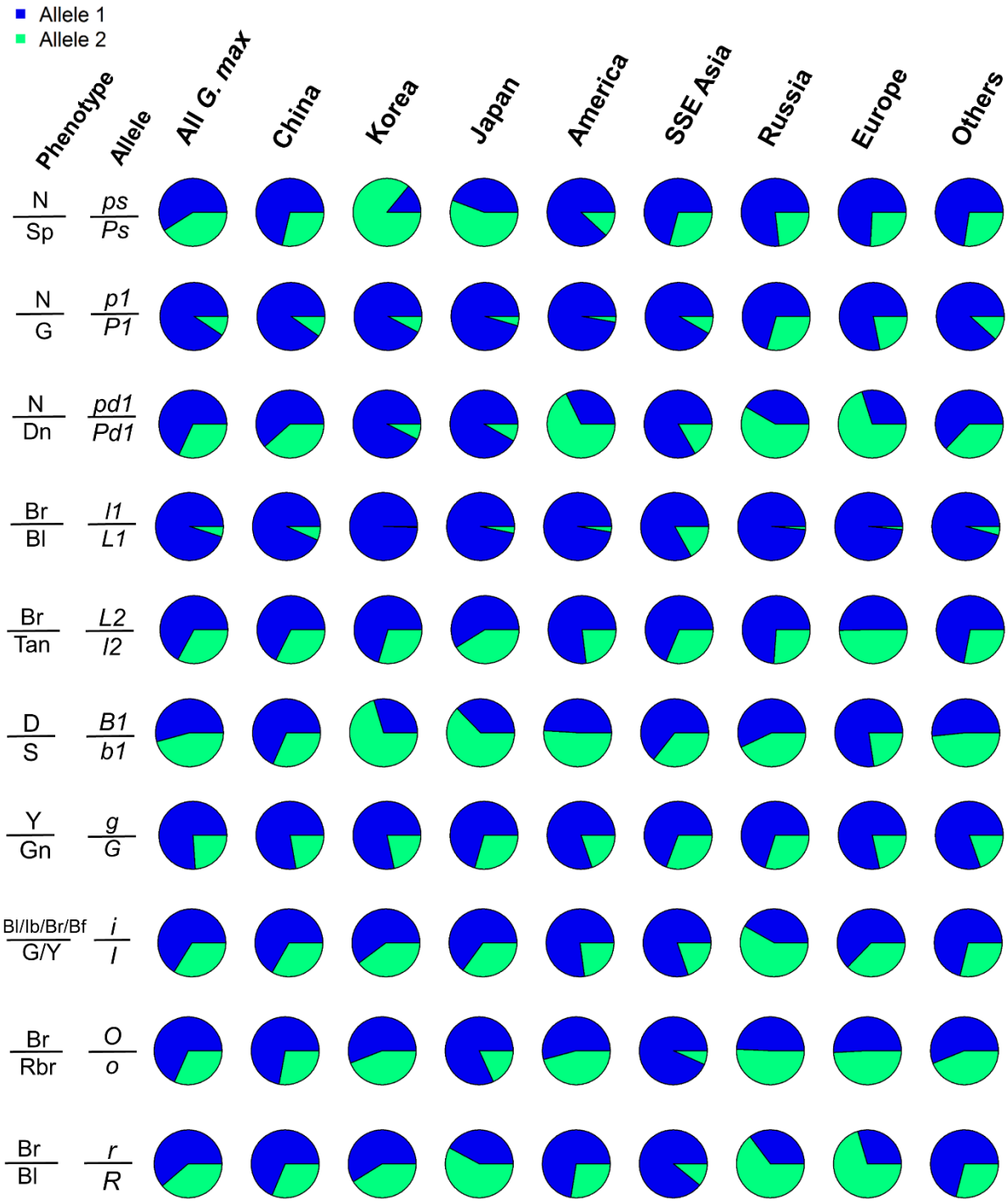


Figure S11. (*Continued*).

CHAPTER 4: DISSECTING THE GENETIC BASIS OF LOCAL ADAPTATION IN SOYBEAN

4. 1 ABBREVIATIONS: GRIN, Germplasm Resource Information Network; EAA, Environmental Association Analysis; GWAS, genome-wide association study; LD, linkage disequilibrium; LG, linkage group; MG, maturity group; MAF, minor allele frequency; PC, principal component; PCA, principal component analysis; QTL, quantitative trait loci; SNP, single nucleotide polymorphism; SP, subpopulation; SPA, Spatial Ancestry Analysis; USDA, United State Department of Agriculture.

4.2 ABSTRACT

Local adaptation is critical for crop species in the face of environmental change but its underlying genetic causes is not clearly elucidated. In this study, we provide first insights on the genetic basis of local adaptation in soybean (*Glycine max*). We exploit natural variation in a large soybean landrace population (N=3,012) within *G. max* maintained in the U.S. Department of Agriculture Soybean Germplasm Collection. We leverage environmental data available for 3,012 landrace accessions to perform environmental association analysis (EAA). Population genomic analysis clearly tracks subpopulation ancestry based on geographic origin. Partitioning of genomic variation revealed that latitude and temperature accounted for most of the explainable genomic variation. Using EAA, we identified numerous small-effect loci contributing to local adaptation. Favorable adaptive alleles are distributed more frequently in the landrace population than in the elite population. We also found a few loci at which allelic effects are determined by geographic origin which may contain important candidate genes for long juvenility in soybean. Selection mapping identified the within-country signatures of selective sweeps which co-localized to loci showing steep gradients in allele frequency that are important for wide-range geographical adaptation. Our approach using a combination of EAA and selection mapping identified important candidate genes related to drought and heat stress, and revealed important signatures of directional selection that are likely involved on geographic divergence of soybean. The geographic extent of environmental and climatic associated SNPs provides the first insight on the genetic architecture of local adaptation and how climate change shapes pattern of genetic variation in *G. max*.

4.3 INTRODUCTION

Global climate is changing with respect to temperature (Yamori et al., 2014), precipitation (Mourtzinis et al., 2015) and soil (Brevik, 2013). Changing environmental conditions force organisms to migrate, adapt, or be phenotypically plastic to avoid extinction (Rellstab et al., 2015). Unlike animals, plants are sessile organisms that cannot escape adverse climatic conditions. One solution to surviving a challenging environment is local adaptation. In this study, we adopt the concept of 'local adaptation' defined as the response to differential selective pressures among populations in a given habitat, which acts on genetically controlled fitness differences among individuals (Williams, 1966; Kawecki and Ebert, 2004; Savolainen et al., 2013). Local adaptation plays an important role in shaping available genetic variation of a population (Kawecki and Ebert, 2004). Directional selection occurs when a local condition favors the best fitted allele, which we refer here as 'adaptive allele', and the frequency of adaptive allele shifts in one direction, this can be interpreted as local adaptation (Kawecki and Ebert, 2004; Abebe et al., 2015). Identification of adaptive alleles is potentially useful for dissecting the genetic basis of crop adaptation (Bragg et al., 2015; Lasky et al., 2015; Russell et al., 2016).

Landscape genomics is a framework that aims to identify environmental factors (e.g., precipitation, temperature, vegetation indices) that shape adaptive genetic variation and genetic variants that drive local adaptation (Rellstab et al., 2015; Bragg et al., 2015). Some earliest examples of local adaptation come from observed concordances between phenotypic traits and environmental variation such as Turesson (1922), who considered genotype as the relevant unit living in different habitats across the distribution of a species. Later, Huxley (1938) coined the term 'ecocline' to describe the cases where phenotypic

variation is correlated with ecological factors. In recent years, a new approach called environmental association analysis (EAA) has emerged as a core part of landscape genomics (Rellstab et al., 2015). One of the earliest examples of EAA was conducted by Mitton et al. (1977) where they found significant variation and association of gene frequencies over geographic distance in *Pinus ponderosa*.

To date, there is a growing body of researches demonstrating the feasibility of landscape genomics for detecting loci related to adaptation. Early work in *Arabidopsis thaliana* has shown that fitness-associated loci exhibited both geographic and climatic signatures of local adaptation (Forunier-Level et al., 2011; Hancock et al., 2011). In maize, Westengen et al. (2012) detected adaptive loci using EAA that respond to precipitation and maximum temperature of a given habitat. Recent studies on some of the world's most important crops such as rice (Meyer et al., 2016), sorghum (Lasky et al., 2015), barley (Russell et al., 2016) and *Glycine soja* (soybean's wild progenitor) (Anderson et al., 2016) provided insights on the genetic architecture and putative genes underlying local adaptation. Landscape genomics frameworks were also found promising in predicting adaptive traits and detecting genetic differences in relative fitness of accessions (Hancock et al., 2011; Lasky et al., 2016).

Statistical methods for EAA have been developed and recently reviewed (Rellstab et al., 2015; Bragg et al., 2015). In a population that is highly structured, a linear mixed model is preferred for controlling neutral genetic background while simultaneously identifying genetic variants strongly associated with environmental variables (Horton et al., 2012; Yoder et al., 2014; Anderson et al., 2016). For example, Yoder et al. (2014) tested for associations of nearly two million SNPs to three climatic factors using a structured-

population in barrel clover (*Medicago truncatula*) and identified more than 20 genes associated with abiotic factors and pathogens. In practice, EAA is often used in concert with other population genomic tools such as outlier analysis (e.g. Fischer et al. 2013). Outlier tests use genomic information to identify signatures of adaptive genetic variation. F_{ST} analysis has been a popular choice to perform outlier tests (Weir and Cockerham, 1984; Excoffier et al., 2009). However, the major drawback of many outlier-based assessments, such as F_{ST}, is that individual genotypes have to be partitioned into discrete populations (Excoffier et al., 2009). Different groupings of the individuals into populations may yield different results, and, thus important signals of selection may be missed (Yang et al., 2012). F_{ST} is also not sensitive to whether allele frequency variation is spatially organized into a steep allele frequency gradient. Alternatively, Spatial Ancestry Analysis (SPA) can be used to identify loci showing extreme frequency gradients, which does not require grouping individuals into populations. In SPA, a continuous function of allele frequency projected onto geographic space, and loci showing steep gradients in allele frequency are inferred to have been subject to selection (Yang et al., 2012).

Understanding the genetic basis of local adaptation is relevant and timely, particularly to the world's most important crops, due to the impact of climate change (Mourtzinis et al., 2015). The C3 legume crop soybean (*Glycine max*) provides an excellent system to investigate the genetic basis of local adaptation. Soybean has a wide range of latitudinal and climatic adaptation in Asia, North and South America (Carter et al., 2004; Anderson et al., 2016). Second, a multitude of localized *G. max* landraces were developed as a result of domestication of *G. max* from its wild species progenitor [*Glycine soja* (Sieb. and Zucc.)] in China ~5,000 years ago (Carter et al., 2004; Hyten et al., 2006). An estimated

45,000 of these unique Asian landraces have been collected and are maintained in *G. max* germplasm collections around the world (Carter et al., 2004; Hyten et al., 2006). The diversity found in *G. max* landraces is the result of a demographic expansion throughout geographically diverse Asia (Carter et al., 2004), which expanded from China to Korea and Japan about 2000 years ago (Kihara, 1969). The geographically diverse landrace accessions possess adaptive traits that can be utilized for crop improvement and thus are suitable for studying local adaptation (Hyten et al., 2006; Song et al., 2015).

Here, we elucidate the genetic basis of local adaptation in soybean by exploring the natural variation available in 3,012 locally adapted landrace accessions from across the geographical range of *G. max* species. We leverage environmental data (geography, temperature, precipitation, and soil) for 3,012 landrace accessions to perform EAA. We then searched for within-country signatures of selective sweeps and compared to loci showing steep gradients in allele frequency that are important for range-wide geographical adaptation. Our approach using a combination of EAA and selection mapping identified important candidate genes related to drought and heat stress, and revealed important signatures of directional selection that are likely involved in geographic divergence of soybean. The geographic extent of environmental and climatic associated SNPs provides the first insight on the genetic architecture of local adaptation and how climate change shapes pattern of genetic variation in *G. max* species.

4.4 MATERIALS AND METHODS

Landrace Collection for Environmental Association and Spatial Ancestry Analysis. The set of landrace accessions used in this study are from the USDA Soybean Germplasm Collection. Only lines with latitude and longitude coordinates were included. These were a subset of the 5,396 accessions previously labeled as landraces (Song et al., 2015), or *G. max* lines added to the USDA collection prior 1945 sourced from China, Japan, North Korea, or South Korea. This threshold was meant to eliminate elite lines developed through modern breeding practices. We then omitted those accessions determined to be genotypic duplicates and accessions that were potential geographic outliers. Filtering left a total of 3,012 landrace accessions that were collected within the geographic range of 22-50°N and 113-143°E. Landrace accessions were distributed in China (N=625), Japan (N=587), South Korea (N=1,737) and North Korea (N=63).

Elite and Landrace Collections for Selection Mapping. Plant materials for selection mapping were comprised of landrace and elite populations recently described (Song et al., 2015). As our objective was to identify genomic regions that were selected locally, we partitioned elite and landrace collections based on country of origin. China had the highest proportion of landrace accessions (N=2,727), followed by South Korea (1,776), and Japan (N=893) (Song et al., 2015). As no landrace accessions originate from North America, we chose the known ancestors of North American soybean (Gizlice et al., 1994; Li and Nelson, 2001; Ude et al., 2003; Hyten et al., 2006) for selection mapping. A total of 65 *G. max* landrace accessions were extracted for North America, all introduced from Asia (Song et al., 2015). The breeding programs of Japan, China and North America have produced a large number of modern cultivars (Carter et al., 1993; Carter et al., 2004). In

this study, the modern cultivar population was comprised of 565 North American cultivars, 364 cultivars from China, 615 cultivars from Japan and 25 cultivars from Korea. We omitted the Korean population for selection mapping analysis because of the small population size for Korean elite lines which may confound the selection mapping results.

Genotype Data

Genotype data from the SoySNP50K platform were downloaded from SoyBase (Grant *et al.* 2010) for all available *G. max* landrace and elite accessions (Song et al., 2015). Ambiguous and heterozygous SNP calls were treated as missing data due to the low outcrossing rate in *G. max* (Carter et al., 2004). The physical map positions of the SoySNP50K SNPs (Song et al., 2013) were mapped into the second genome assembly 'Glyma.Wm82.a2' (Anderson et al., 2016). Any SNP with minor allele frequency (MAF) < 0.01 was removed from the genotype dataset for subsequent analyses. The SNP genotype data set is publicly available at http://www.soybase.org/dlpages/index.php.

Environmental Data

Climate Data. The latitude and longitude coordinates of 3,012 *G. max* accessions were used to query the WorldClim database (see http://www.worldclim.org/) for 112 environmental variables, including bioclimatic variables based on yearly, quarterly, and monthly temperature and precipitation data as well as altitude data at a resolution of 30 arc-seconds (approximately 1 km grids) (Hijmans et al., 2005). The bioclimatic variables represent annual trends, seasonality and extreme or limiting environmental factors that are often used in ecological niche modeling (Hijmans et al., 2005). The unit used for downloaded temperature data are in °C * 10. This means that a value of 231 represents

23.1 °C. Temperature data was converted into °C by dividing the temperature value by 10. The unit used for the precipitation data is millimeter (mm).

Soil Data. The sampling locations of 3,012 landrace accessions were also used to query the ISRIC database (World Soil Information database, Hengl et al., 2014) for seven biophysical variables (pH x 10 in H_2O, percent sand, percent silt, percent clay, bulk density in kg/m^3, cation exchange capacity in cmolc/kg, and organic carbon content (fine earth fraction) in permilles) at a resolution of 30 arc-seconds (see http://www.isric.org/). Available data for seven biophysical variables were taken at six soil depths: 2.5 cm, 10 cm, 22.5 cm, 45 cm, 80 cm, and 150 cm (Hengl et al., 2014). Because of high correlation and less variability in soil variables across depths, we grouped the six measurements per variable into one class by taking the average value across soil depths.

Principle component analysis on the bioclimatic and biophysical variables (first scaled to a mean of 0 and standard deviation of 1) was conducted using the prcomp function in R (R Development Core Team 2015). Pearson correlation coefficients between bioclimatic and biophysical variables were calculated in R. Boxplots for each scaled bioclimatic and biophysical variable were created based on *G. max* localities to examine the distribution for each variable (**Supplementary Fig. 1**).

Population Structure and Linkage Disequilibrium

Principal component analysis using SNPs present in all landrace accessions was conducted using the *prcomp* function in R (version 3.1.0). The Bayesian clustering program

*fast*STRUCTURE was used to calculate varying levels of K (2-10) and the command chooseK.py was used to identify the model complexity that maximized the marginal likelihood (K=2-6). The population structure was visualized using barplot based-function in R. Genome-wide and intra-chromosomal linkage disequilibrium (LD) was estimated using pairwise r^2 between SNPs, which was calculated using PLINK version 1.07 (Purcell et al., 2007).

We calculated the proportion of genome-wide SNP variation among landrace accessions that could be explained by environmental variables (temperature, precipitation and soil) and geographic data (latitude, longitude). We used variance partitioning of redundancy analysis (RDA) implemented in the R package *vegan* (Oksanen et al., 2010). The RDA is an eigenanalysis ordination to assess the explanatory power of multivariate predictors (environmental and geographical variables) for multivariate responses (e.g., SNP data) (van den Wollenberg et al., Lasky et al., 2002; Peres-Neto et al., 2006). The variance components explained by environmental variables were partitioned by fitting different models. The first model considered all environmental and geographic variables as explanatory variables and the SNP data as response variables. Because geographic effects are correlated with the SNP data, we fit a partial model in which the SNP data were conditioned on the effects of geographic coordinates. For both models, significance testing was conducted using Monte Carlo permutations test with 500 runs and α=0.01.

Detection of Selection Footprints

F_{ST} outlier analyses and Spatial Ancestry Analysis (SPA) were used to identify loci that had been differentially selected. To identify loci that had been selected locally, F_{ST}

analyses was conducted between elite and landrace populations within each country ($F_{ST\ within}$). Theta (Θ), the variance-based F_{ST} estimate of Weir and Cockerham (1984), was estimated using the R *hierfstat* package (Goudet, 2005). For visualization, F_{ST} was averaged in sliding windows, with a window size of 5 and a step of 3 SNPs (Anderson et al., 2016). SNPs with F_{ST} values above the 99.9th percentile were identified as outliers. A Mantel test was conducted to explore isolation by distance utilizing great circle distance between geographic locations and pairwise genetic distance using the *vegan* package in R (Oksanen et al., 2007).

SPA was used to detect loci showing steep gradients in allele frequency (Yang et al., 2012). The SPA incorporates geographic and genetic gradients in identifying local clines (Yang et al., 2012). This type of analysis is particularly compelling for species with a continuous distribution and relationship among individuals driven by isolation-by-distance (Yang et al., 2012). The outputs of the SPA model are individual mapping coordinates and coefficients for allele frequency slope functions. Based on the steepness of allele frequency slope, the selection scores from SPA were generated as follows (Yang et al., 2012)

$$SPA_j = \sqrt{\sum_i \left(f_j(x_i) - \frac{\sum_i f_j(x_i)}{N} \right)^2}$$

where $f_j(x_i)$ = 1/1(+exp ($-a_j^T - x_i - b_j$)) for the allele frequency for individual *i* at locus *j*. SNPs with SPA scores above the 99.9th percentile were identified as outliers.

Environmental Association Analysis

To perform environmental association analysis, four mixed linear models were fitted: K, Q+K, P+K, and L+K. The generic Q+K model was fitted using the equation

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{C\gamma} + \mathbf{Zu} + \mathbf{e},$$

where **y** is a vector of environmental variable; **β** is a vector of fixed marker effects; **γ** is a vector of subpopulation effects; **u** is a vector of polygenic effects caused by relatedness , i.e., $\mathbf{u} \sim MVN(0, \mathbf{K}\sigma_u^2)$; **e** is a vector of residuals, i.e., $\mathbf{e} \sim MVN(0, \mathbf{I}\sigma_e^2)$; **X** is a marker matrix; **C** is an incidence matrix containing membership proportions to each of the three genetic clusters identified by the *fast*STRUCTURE analysis; **Z** is the corresponding design matrix for **u;** and **K** is the realized relationship matrix estimated internally in the Factored Spectrally Transformed Linear Mixed Models (FaST-LMM) using the SNP data (Lippert et al., 2011). The K model was the same with the Q+K model except that the term **Cγ** was removed in the model. In the P+K, the incidence matrix C of the Q+K model was replaced with a matrix that contained the first three PCs identified from PCA. In the L+K model, the incidence matrix C of the Q+K model was replaced with a matrix that contained the latitude information where each accession was collected from. All models were implemented using the FaST-LMM algorithm (Lippert et al., 2011).

A comparison wise error rate of $P < 0.0000143$ was used to control the experiment-wise error rate based on the methods of Li and Ji (2005). Briefly, the correlation matrix and eigenvalue decomposition of SNPs with MAF > 0.01 were calculated to determine effective number of independent tests (M_{eff}) (Li and Ji, 2005). The test criteria was then adjusted using the M_{eff}=3,578 tests with the correction (Sidak, 1967) $\boldsymbol{\alpha_p = 1 - (1- \alpha_e)^{1/Meff}}$, where α_p is the comparison-wise error rate and α_e is the experiment-wise error rate. An $\alpha_e = 0.05$ was used in this study. Multiple linear regression was used to estimate the

proportion of phenotypic variance accounted for by significant SNPs after accounting for population structure effects.

Haplotype Analysis

Haplotype analysis was conducted to further investigate the variation in and around putatively selected regions. A haplotype analysis was performed within a range of significant SNPs identified by environmental association and selection mapping. Haplotype blocks were constructed using the four gamete method (4gamete) (Wang et al., 2002) implemented in the software Haploview (Barrett et al., 2005). The 4gamete method creates block boundaries where there is evidence of recombination between adjacent SNPs based on the presence of all four gametic types. A cutoff of 1% was used, meaning that if addition of a SNP to a block resulted in a recombinant allele at a frequency exceeding 1%, then that SNP was not included in the block.

Candidate Gene Annotations and Enrichment Analysis

SNPs identified as outliers through the environmental association mapping, SPA, or F_{ST} approaches were examined for functional annotation using SoyBase (www.soybase.org) (Grant et al., 2010). A sliding window-approach (e.g., 50kb) was used to search for functional genes implemented in bedtools (Quinlan and Hall, 2010). The prediction of candidate genes was based on (a) genes of known function in soybean related to the trait under study, and (b) genes with function-known sequence homologs in *Arabidopsis* related to the trait. For each significant SNP, we collected additional information on genic context, nearby annotated genes, and the inferred Arabidopsis ortholog (TAIR10 best hit provided by Soybase). We performed enrichment analysis to determine if euchromatin, 3′ UTR, 5′ UTR, coding sequence (CDS), and intronic regions

were over or under represented among outliers. Significance of enrichment was assessed by creating a 99% confidence interval around the proportion of SNPs that were found in each category as calculated by bootstrap sampling the number of SNPs in each category 1000 times.

4.5 RESULTS

4.5.1 Genetic Structure of the Soybean Landrace Population

We observed a clear subpopulation (SP) structure in the 3,012 landrace accessions with a clustering pattern based on geographic origin (**Fig. 1**). Based on model complexity and component analysis as computed by *fast*STRUCTURE (Raj et al., 2014), the optimal number of SPs was predicted to be K=2 up to K=6. At K=3, accessions grouped clearly by geographic origin (**Fig. 1a)**; SP1 (green cluster) represents the accessions mostly collected from Korea, SP2 (red cluster) composed predominantly of accessions collected from China, and SP3 (blue cluster) primarily of accessions from Japan. Using principal component (PC) analysis of SNP data, the top two PCs, which mainly accounted for geographic origin differences, explained ~13 % of total genetic variation (**Fig. 1b**). About 71% of the landraces were assigned to groups consistent with a *priori* population definitions based on > 80% ancestry cut-off (**Fig. 1a**).

We also found a substantial number of admixed accessions (ancestry < 80%) highlighting the complex history of landrace population when soybean was domesticated from *G. soja* (Hyten et al., 2006). Twenty-nine percent of the landrace accessions had shared ancestry between mainland (China) and island populations (Japan and Korea). The biggest admixed portion was identified in China (56%) while accession in Korea (25%) and Japan (14%) had homogenous subpopulations with fewer admixed individuals. The geographic distribution of accessions with above 80% ancestry for each of three SPs was plotted **(Fig. 1c).** The clustering corresponds primarily to migration barriers and fits well

with previous studies of population structure of soybean landrace population (Li et al., 2012; Song et al., 2015).

4.5.2 Population Differentiation and Linkage Disequilibrium

On the basis of SNP data, the global genetic differentiation among country ($F_{ST=}$0.14) was modest, the same as rice (F_{ST} =0.14, Huang et al., 2014) and slightly higher than human (F_{ST} =0.12, The International HapMap Consortium, 2005). The pairwise genome-wide F_{ST} between countries ranged from 0.10 to 0.16, with accessions from Japan and China (F_{ST}=0.16) being more-differentiated than accessions from Japan and South Korea (F_{ST}=0.10). A Mantel test showed that isolation by latitudinal distance drives genetic differentiation in landrace population (r = 0.579, p < 0.0001). These results suggests that geographic isolation has impact on shaping genetic differentiation in landrace accessions.

Within this landrace collection, the genome-wide LD in the euchromatic region decays from r^2 = 0.48 to r^2 = 0.15 within 300 Kb, then decays more slowly to r^2 = 0.10 from 500 to 1500 Kb (**Supplementary Fig. 2)**. The genome-wide LD in the heterochromatic region drops from r^2 = 0.58 to r^2 = 0.27 within 500 Kb, reaching half of its initial value at around ~400 Kb, then decays more slowly and extends up 2,000 Kb with r^2 = 0.18 (**Supplementary Fig. 2**). The intra-chromosomal LD was more extensive in the heterochromatic than in the euchromatic regions (**Supplementary Fig. 3**).

4.5.3 Environmental Variability

Climatic and environmental variables we examined can be classified broadly into four categories: geographic-related (latitude, longitude, and altitude), temperature-related

(monthly and seasonal), precipitation-related (monthly and seasonal) and soil-related (**Supplementary Fig. 1**). Variation was observed in 3,012 accessions for 112 climatic and environmental variables (**Supplementary Fig. 1**). Pearson correlation analysis demonstrated that phenotypes within each category are often correlated (**Supplementary Fig. 4**). Across the geographic range, China (mean annual temperature=48.6) is colder and drier than Japan (mean annual temperature=115.9) and Korea (mean annual temperature=118.9) (**Supplementary Fig. 5)**. The soils data indicate that Japan soils have lower bulk density, higher soil organic carbon, higher sand content, and more variable pH than the soils in China and Korea (**Supplementary Fig. 5**). Soil in China has higher bulk density, higher cation exchange capacity and higher average pH (**Supplementary Fig. 5**) than Japan and Korea.

Korea has the highest amount of precipitation and warmest temperature during the months of July and August (**Fig. 2**), while China has consistently low precipitation and temperature except for the month of June (**Fig. 2a, 2b**). Japan has the highest precipitation from January to March and then from September to December (**Fig. 2a**). Overall, higher precipitation occurs during the hottest days of July and August, which is consistent to the peak of the Summer Season in those countries (**Fig. 2**). The growing season for soybean in those countries varies. In Japan and Korea, planting season starts as early as March and harvesting extends as late as October to November. In China, most soybeans are planted from late April to late June and then harvested by September through early October.

A PCA of environmental variables that included 67 bioclimatic and 42 biophysical variables recapitulates the geography (**Supplementary Fig. 6).** The first four principle components explained ~ 86% of the total variation (**Supplementary Fig. 6**). The first PC

was associated with temperature and latitude, the second PC with monthly and extreme precipitation, the third PC soil/precipitation, and the fourth PC with soil. The correlation of PCs to specific environmental variables (i.e., PC1 was correlated with temperature) was almost similar to the analysis reported in *G. soja* (Anderson et al., 2016). Plotting the PC values grouped the landrace accessions by geography, suggesting that the observed variability among landraces is due to geographic isolation.

4.5.4 Partitioning of Genomic Variation Explained By Environmental Variables

A partial redundancy analysis (van den Wollenberg et al., Lasky et al., 2002; Peres-Neto et al., 2006) was used to partition the amount of genomic variation explained by environmental variables. This analysis was performed to condition the effects of geographic effects and separate the genomic variance explained by environmental variables. Partitioning the effects of environmental variables indicated that temperature (5.2%) explained a higher portion of genomic variation than soil (3.4%) and precipitation (3.3%) (**Fig. 3**). Geographic variables explained 6.9% of total genomic variation (**Fig. 3**). All environmental variables (temperature, precipitation and soil) cumulatively explained 7.3% of the total genomic variance after accounting for the geographic effects. Taken together, the combined environmental and geographic effects explained 14.3% of the total genomic variation. The results indicated the importance of geography and temperature in shaping the existing genetic variation in a soybean landrace population.

4.5.5 Environmental Association Analysis

Four mixed linear models (K, PCA + K, Q + K and Latitude + K) were fitted to correct for confounding effects of subpopulation structure and relatedness among landrace

accessions. Among these four methods, the L+K model had a reduced number of associated SNPs (322) mainly due to high correlation of latitude with temperature variables (**Fig. 3)**. Fitting the L+K model resulted in an overcorrection (causing Type 2 error) and a corresponding reduction in SNP significance (**Supplementary Fig. 7**). The Q+K and P+K models had minimal inflation of *p*-values with lowest Λ while the K model resulted in a slight under-correction (**Supplementary Fig. 7**).

Using GWA mapping results from the Q+K model, a total of 73 distinct genomic regions were identified for 112 environmental variables (**Fig. 4a, Supplementary Table 9**). The highest number of associated regions were identified for precipitation (27) followed by temperature (20), geography (19) and soil (17) (**Fig. 4a and 4b)**. The number of associated loci was higher for monthly-specific variables than seasonal and annual-based temperature and precipitation (**Fig. 4c**). About 41% of the associated loci were identified in the genic region,(**Supplementary Fig. 8**). We found that associated loci are significantly enriched in the genic region ($p<0.01$). About 70% of associated loci were identified in euchromatic regions that occur in chromosome ends while the remaining 30% occur near or in heterochromatic regions (**Supplementary Fig. 9a and 9b)**. In the soybean genome, about 78% of the predicted genes occur in chromosome ends (Schmutz et al., 2010).

There was a substantial overlap in associated loci within and between trait(s) category. The SNP sharing among associated loci ranged from 0 to 87.1% between two traits **(Fig. 4b**) and among variables measured on monthly vs annual-basis (**Fig. 4c**). The non-shared SNPs between traits were highest for precipitation (32.6%), followed by soil (22.7%), temperature (19.7%) and geography (12.9%) (**Fig. 4b)**. The results suggest that

pleiotropy is common among adaptive alleles. However, some of these results may be due to correlations among variables, rather than pleiotropy *per se* (**Supplementary Fig. 4**).

4.5.6 F_{ST}-Based Selection Mapping within Each Country

Genome-wide F_{ST} between elite and landrace populations was highest for China (F_{ST}=0.09), followed by America (F_{ST}=0.05) and Japan (F_{ST}=0.01). A total of 24 genomic regions were selected (China =11; Japan=8; North America=5) (**Fig. 5**; **Supplementary Fig. 9a and 9b)**. The regions on chromosomes 4, 16 and 19 had the longest range of selective sweep features, indicating that these chromosomes might be the one most affected during soybean adaptation and improvement. Of 24 selected regions, only two overlapped between countries. The first region on chromosome 4 at 4384695 Bp were preferentially selected in China and North America but was not selected in Japan (**Fig. 5)**. The chromosome 4 region showed the divergence between China (F_{ST} = 0.62) and America (F_{ST}= 0.49) to Japan (F_{ST}=0.01) is a QTL hotspot associated for many traits such as resistance to soybean cyst nematode race 2 (Vhuong et al., 2011), seed size (Orf et al., 1999), pods per nod (Zhang et al., 2004) and seed protein concentration (Specht et al., 2001). The second overlapped region was identified on chromosome 8 between 8451046-8602715 Bp that co-localized with the *I* locus, which controls the distribution of anthocyanin and proanthocyanidin pigments (Todd and Vodkin 1996; Tuteja et al. 2009). The *I* locus region was selected in China and Japan but it was not identified in America (**Fig. 5).** Overall, a substantial number of selected regions between countries were distinctive, indicating that different genes are being selected within each country (**Supplementary Fig. 10**).

4.5.7 Spatial Ancestry Analysis Identified Loci under Selection

Six strongly selected regions were identified by SPA (**Fig. 6**). Like F_{ST} analysis, SPA also identified two selected regions that had a long range of clustered of SNPs; one region on chromosome 16 and another region on chromosome 19 that co-localized with *Pdh1* and *Dt1* loci, respectively. These are well-known genes associated with soybean genetic improvement (Song et al., 2015; Hyten et al., 2006) and widely noted as a target of strong selection in soybean (Song et al., 2015; Wen et al., 2015). We then looked on the common selection signals identified between countries and SPA. Out of the six selected regions, only the *pdh1* locus on chromosome 16 overlapped with the F_{ST} analysis and it was solely identified in America (**Supplementary Fig. 10)**.

4.5.8 Overlapping Signals between Environmental Association and Selection Mapping

We examined the genome for any overlap between the QTLs identified by environmental association and the strongly selected regions identified by F_{ST} and SPA (**Supplementary Figs. 9a and 9b)**. Five associated regions were found to be within the range of selected regions, two of which were near the known genes. Haplotype analyses on selected regions revealed that SNPs indicates a fairly low gene diversity within the range of selected region. Selected region on chromosomes 16 (29517407 – 31181902 Bp) and 19 (44533069-45292930 Bp) co-localized with *pdh1* **(Fig. 7a, 7b)** and *Dt1* (stem termination gene) (**Fig. 8a, 8b)**. Allelic effects of SNPs near *pdh1* and *Dt1* suggests a genic relationship between them. Accessions carrying T *(pdh1)* and G *(Dt1)* alleles predominates under high precipitation and high temperature while accessions carrying the alternative alleles

predominates under low precipitation and low temperature (**Fig. 7c; Fig. 8c)**. The T *(pdh1)* and G *(Dt1)* alleles were nearly fixed in SP1 and SP3 but it was rare in SP2 (**Fig. 7g; Fig. 8g)**.

A signal of selection on chromosome 15 at 9964637 Bp was in complete LD with region between 9840775-10142301 Bp associated with Soil Silt Content. The most significant SNP (for EAA?) at this region is located at 10142301 Bp which is 3.63 Kb away from Glyma.15G127700 that encodes for Root hair defective 3 GTP-binding protein (*RHD3*) (**Fig. 9a, 9b)**. Several studies have shown *RHD3* to affect root epidermis development and is required for appropriate root and root hair cells enlargement in *Arabidopsis* (Zhong et al., 2003; Yuen et al., 2005). Allelic effect estimates indicate that accessions carrying the T allele tend to thrive in soil with higher silt content (**Fig. 9c**). The T allele is frequent in SP2 (q=0.72) (**Fig. 9d, 9e)** while it was rare in SP1 (q=0) and SP3 (q=0) (**Fig. 9d, 9e)**.

On chromosome nine, a selected region was identified between 1054596–1261468 Bp that spanned 206 Kb. Haplotype analysis collapsed the 206 Kb region into four haplotype blocks. Blocks two and three contained highly significant SNPs that spanned only 40Kb (**Supplementary Fig. 11**). Only four candidate genes were narrowed within the 40Kb-region, two (Glyma.09G014700 and Glyma.09G014800) of which are functionally related to environmental stresses based on *Arabidopsis* annotations. Glyma.09G014700 is annotated as a Ca^{2+}-dependent lipid-binding (CaLB domain) family protein which operates when intracellular Ca^{2+} rises due to environmental stresses. CaLB binds Ca^{2+} to undergo a conformational change to activate stress-responsive genes. Silva et al. (2011) identified a novel transcriptional regulator, a Ca^{2+}-dependent lipid-binding protein (AtCLB) containing

a C2 domain that binds specifically to the promoter of the *Arabidopsis thaliana* synthase gene, whose expression is induced by gravity and light. The loss of the AtCLB gene function confers an enhanced drought and salt tolerance and a modified gravitropic response in T-DNA insertion knockout mutant lines (Silva et al., 2011). On the other hand, Glyma.09G014800 is annotated as an oxidoreductase, 2OG-Fe (II) oxygenase family protein. In *Arabidopsis*, the mutant downy mildew resistant 6 encodes a putative 2OG-Fe (II) oxygenase that is defense-associated but required for susceptibility to downy mildew (Damme et al., 2008). In soybean, 2OG-Fe (II) oxygenase was identified as an important candidate for early iron deficiency chlorosis signaling in soybean roots and leaves (Lauter et al., 2014).

A selected region on chromosome 17 spanning 608 Kb between 3857335–4466291 Bp was associated with longitude, soil and temperature variables and was identified only in North? America (**Supplementary Fig. 12)**. Haplotype analysis collapsed the 608 Kb region into nine haplotype blocks (**Supplementary Fig. 12b)**. Haplotype block three had the most significant SNP, spanning only 30Kb and contained two plausible candidate genes including a calmodulin-binding factor and a heat-shock transcription factor (**Supplementary Fig. 12c)**. Pandey et al. (2013) showed that a calmodulin binding transcription activator, *CAMTA1*, regulates drought responses in *A. thaliana*. This calmodulin-binding factor is known to be functionally related to the Ca^{2+}-dependent lipid-binding (CaLB domain) identified on chr nine (**Supplementary Fig. 11**). Interestingly, Dobney et al. (2009) revealed that the Calmodulin-related Calcium Sensor (CML42) plays a role in trichome branching. In adverse environments, trichomes are implicated in local adaptation by protecting plants from abiotic stresses including UV damage (Liakopoulos

et al., 2006; Yan et al., 2012) and heat load reduction (Espigares and Peco, 1995). Further, trichome density is negatively correlated with the rate of transpiration (Benz and Martin, 2006) and carbon dioxide diffusion (Galmes et al., 2007). Finally, heat-shock proteins (HSP) mainly respond to high light intensity and heat stress (Xu et al., 2011; Gurley et al., 2000; Kotak et al., 2010).

Another interesting region was identified on chromosome 20 between 45864382–47884469 Bp because of a high cluster of associated SNPs (17) with precipitation and geographic variables. Though the evidence of selection is only modest, this region co-localized with genes related to drought and cold stress (**Supplementary Fig. 13)**. A haplotype analysis covering the chromosomal interval between 45864382–47884469 Bp separated the set of plausible candidate genes, with at least three genes represented per haplotype block (**Supplementary Fig. 13b**). We focused on the most significant haplotype block which contained Glyma.20G225400, with an *Arabidopsis* ortholog that encodes Dehydration-Responsive Element Binding Protein2a (DREB2A) (**Supplementary Fig. 13c)**. The DREB2A is a well-known gene for drought response and cold tolerance (Sakuma et al., 2006; Qin et al., 2008; Lata and Prasad, 2011) which encodes a transcription factor that controls the expression of water deficit-inducible genes (Qin et al., 2008). The two significant SNPs within this largest haplotype block was between 45864382–46882334 Bp which were identified for Annual Precipitation and Longitude. Allelic effects of the closest SNP tagging DREB2A indicates that the T allele has a wide a range of annual precipitation. The T allele was predominant in SP2 (q=0.47) but had lower frequency in SP1 (q=0.16) and SP3 (q=0.19) (**Supplementary Fig. 13d, 13e)**. The DREB2A gene is of interest to get a better understanding of drought tolerance in soybean.

The T allele was predominant in SP2 (q=0.47) but had lower frequency in SP1 (q=0.16) and SP3 (q=0.19) (**Supplementary Fig. 13d, 13e)**. The DREB2A gene is of interest to get a better understanding of drought tolerance in soybean.

4.6 DISCUSSION

4.6.1 Environmental Association and Selection Mapping Provide Insight on Genetic Basis of Local Adaptation

Detecting adaptive genetic variation in response to environmental variation helps to better understand the local adaptation that has occurred in soybean (Hoffmann and Willi, 2008). In this study, we did not identify any large-effect variants underlying local adaptation. Instead, we found many small-effect variants that cumulative explained up to 10% of total phenotypic variation, with the largest effect locus explaining ~5%. This result is not surprising given the cumulative effects of all environmental variables included in this study only explained 14% of the total genomic variation. This might be the upper limit of how much genetic variation can be attributed to local adaptation using the soybean landrace population. The genomic variance explained by environmental variables in this study (14%) is lower compared to recent studies of local adaptation in sorghum (31%, Lasky et al., 2016) and barley (40%, Abebe et al., 2015). A possible reason is that genetic diversity in soybean (Hyten et al., 2006) is lower compared to sorghum and barley (Zhu et al., 2003; Hamblin et al., 2004). Other possible reasons can be attributed to differences in population size, the SNP coverage of the entire genome and the decay rate of linkage disequilibrium across the genomes, which are all specific inherent properties of a population.

Exploring the distribution of adaptive alleles based on allele frequency spectrum of associated loci, we found that adaptive alleles in this study were mostly frequent in the landrace population but less frequent in the elite population. When compared our results to previous study (Song et al., 2013) and we found out that virtually the same pattern of allelic

distribution was observed on associated loci. About 35% of associated loci have allele frequency less than 5% in elite population with 20% being completely absent in the elite population. The corresponding alleles, however, have allelic frequency that ranged up to 49 % in the landrace population. Our result is in agreement with the analysis of Anderson et al. (2016) using a population comprised of *G. soja* accessions. Anderson et al. (2016) found that adaptive allele were less frequent in elite population than landrace and *G. soja* populations. These results indicate that genetic variation associated with environmental variables is controlled by a multitude of rare or low-frequency alleles in elite population. A possible explanation for this result is that elite cultivars contain most of the common variation of the Asian landrace collection, most of which are related to genetic improvement unrelated to our environmental assessment of local adaptation. This could be due to the fact that a substantial portion of alleles (mostly adaptive) were lost due to introduction bottlenecks when soybean was introduced to different countries (Hyten et al., 2006). This agrees with commonly held belief that genetic bottlenecks result in the loss of many valuable alleles that may play an important role for local adaptation. More importantly, our results highlight the need to increase the frequency of favorable adaptive alleles in modern breeding lines.

Local adaptation is expected to lead to elevated levels of differentiation among populations at selected loci (Bragg et al., 2105). This is because an allele that confers an advantage under particular environmental conditions is likely to occur at elevated frequency in populations where that condition is prevalent (Abebe et al., 2015; Bose and Bartholomew, 2013). This scenario holds true in this study where the five strongly selected regions tend to follow a directional pattern of selection. We found the best fitted alleles

were prevalent under specific environmental condition (mostly extreme conditions) and then had fixed or near to fixed allelic frequency due to selection (either conscious or unconscious) for a particular subpopulation. One major impact of directional selection is that it removes the alternative allele at particular locus and reduces the diversity in the nearby SNPs of selected sites (Bragg et al., 2015). This is indeed the case of selected regions of SPA and F_{ST} where there were unusually long regions with SNPs that are all in LD. Across the genome we found that most adaptive loci identified by EAA had low to modest F_{ST}. This results indicates that a large fraction of the soybean genome had not been selected, for local adaptation purposes, by farmers nor by breeders or our methods weren't able to detect this selection. In a modern soybean breeding program, the main objective of breeders is to increase the frequency of favorable alleles and purge deleterious mutations that affect yield, seed protein and oil content and disease resistance. Selecting genomic regions for target traits may not always results to positive selection along with the adaptive alleles. Possible reasons are: 1) landraces that possess adaptive allele are always inferior agronomically compared to commercial cultivars which hinder the use of landraces as parents for breeding, 2) adaptive alleles are possibly not linked to breeder's target traits, and 3) the limited available genetic diversity within a locality hinders the development of adaptive cultivars, and 4) isolation by distance may limit gene flow of favorable alleles among populations.

We also found a few loci at which allelic effects are determined by geographic variables. About 50% of associated loci, most of which were identified for temperature variables, were correlated with geographic variables (e.g., latitudinal variation). This result is expected given that soybean is adapted photoperiodic response associated with a

latitudinal gradient, and it lacks the genetic flexibility to be farmed successfully at latitudes that diverged more than about 2◦ (equivalent to a few hundred kilometers) (Carter et al., 2004). For example, moving an adapted soybean genotype out of its zone 2◦ north could delay its maturity sufficiently to risk frost damage, while moving the same genotype south 2◦ would reduce plant height to a degree unacceptable for farming. This might be the phenomenon of initial local adaptation in soybean that consequently resulted into a clear subpopulation structure that what we observed in this study which could have been aided further by the self-pollinating nature of soybean (>99%).

We also identified loci that co-localized with *E* genes that affects flowering and maturity in soybean. Interestingly, we also identified a region that did not co-localize with bp positions of reported *E* gene, which are of interest, for evaluation of long juvenility in soybean (Harada et al., 2015). The long juvenility genes are of interest for delaying the onset of flowering of genotypes adapted to non-equatorial environments so that these genotypes will then have greater yield potential when grown in near-equatorial short-day latitudes of soybean production (Cober, 2010). For example, an interesting region on chromosome 5 (41415712 Bp) associated with 30 temperature variables, co-localized with Glyma.05G238300 that encodes for Rubisco methyltransferase family protein (**Supplementary Fig. 14)**. This putative gene encodes protein methylase that is highly expressed in leaves during daylight and thought to be involved in the regulation of Rubisco during photosynthesis (Raunser et al., 2009).

4.6.2 Implications on Developing Climate-Ready Soybean Cultivars

Soybean is one of the most important crop plants for seed protein and oil content, and for its capacity to fix atmospheric nitrogen through symbioses with soil-borne microorganisms (Wilson, 2008). The United States produces 38% of the world's soybean and production is concentrated in the upper Midwest region (Specht et al., 2014). Climate change has been shown to adversely affect soybean production in the USA (Specht et al., 2014; Mourtzinis et al., 2015). Particularly, the combined year-to-year changes in precipitation and temperature suppressed the U.S. average yield gain by around 30 percent, leading to a loss of $11 billion (Mourtzinis et al., 2015). In-season temperature had a greater impact on soybean yields than in-season precipitation. Averaging across the United States soybean yields fell by around 2.4% for every 1 °C rise in growing season temperature (Mourtzinis et al., 2015). In the Midwestern region of the US, higher summer temperature has become the norm which has resulted in a remarkable geographic shift in the location of soybean production (Specht et al., 2014). During the 33-yr time frame between 1979 and 2011, soybean production shifted northward to become more concentrated in the north-central United States (Specht et al., 2014). The most noteworthy aspect of this shift was the movement of soybean production into the northern Great Plains states of South and North Dakota (Specht et al., 2014). This highlights the needs to develop a climate-ready soybean cultivars (i.e., soybean cultivar with greater temperature tolerance).

The results of this study is one step towards understanding the soybean's capability for local adaptation. Overall, most of the candidates underlying associated loci co-localized with abiotic-stress responsive genes (e.g., DREB, ERD, HSP) which may play a central role in stress tolerance and can be an important mechanism of soybean for local adaptation.

These narrowed candidate genes are essential to get a better understanding of drought and heat tolerance in soybean. We also found a candidate that is functionally related to photosynthesis (e.g., Rubisco methylase) which is of interest for increasing soybean yield in environments with varying temperature. Further, this in turn could assist breeders to identify individual landrace accession that have adaptive alleles that could be used as donor parent for breeding climate-ready soybean cultivars.

4.7 REFERENCES

Abebe, T.D., A. A. Naz, and J. Léon. 2015. Landscape genomics reveal signatures of local adaptation in barley (*Hordeum vulgare* L.). Front Plant Sci. 6:813.

Allendorf, F.W., P.A. Hohenlohe, and G. Luikart. 2010. Genomics and the future of conservation genetics. Nat Rev Genet 11:697-709.

Anderson, J. E., T. J. Kono, M. B. Kantar, and P. L. Morrell. 2015. Environmental association analyses identify candidates for abiotic stress tolerance in *Glycine soja*, the wild progenitor of cultivated soybeans. G3 6:835-843.

Bandillo, N., D. Jarquin, Q. Song, R. Nelson, P. Cregan, J. Specht, and A. Lorenz. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Gen 8:1-13.

Benz, B. W., and C. E. Martin. 2006. Foliar trichomes, boundary layers, and gas exchange in 12 species of epiphytic *Tillandsia* (Bromeliaceae). J. Plant Physiol. 163 648-656.

Bernard, R.L. 1972. Two genes affecting stem termination in soybeans. Crop Sci. 12:335-239.

Bose, R., and A. J. Bartholomew. 2013. Macroevolution in deeptime. Briefs Evol. Biol. 3:1-59.

Brady, K. U., A. R. Kruckeberg, and H. D. Bradshaw Jr. 2005. Evolutionary ecology of plant adaptation to serpentine soils. Annu. Rev. Ecol. Evol. Syst. 36: 243-266.

Bragg, J. G., M. A. Supple1, R. L. Andrew, and J. O. Borevitz. 2015. Genomic variation across landscapes: insights and applications. New Phytol. 207: 953-967.

Brevik, E. 2013. The potential impact of climate change on soil properties and processes and corresponding influence on food security. Agriculture 3:398-417.

Carter, T. E., R. L. Nelson, C. H. Sneller, and Z. Cui. 2004. Genetic Diversity in Soybean. Soybeans: Improvement, Production, and uses (American Society of Agronomy Monograph Series): 303-416.

de Silva, K., B Laska, C. Brown , H. W. Sederoff, and M. Khodakovskaya. *Arabidopsis thaliana* calcium-dependent lipid-binding protein (AtCLB): a novel repressor of abiotic stress response. J Exp Bot 62:2679-89.

Eckert, A. J., J. van Heerwaarden, J. L. Wegrzyn, C. D. Nelson, J. Ross-Ibarra, S. C. Gonzalez-Martınez, and D. B. Neale. 2010. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics 185: 969-982.

Espigares, T. and B. Peco. 1995. Mediterranean annual pasture dynamics: impact of autumn drought. J Ecol 83:135-142.

Excoffier, L., T. Hofer and M. Foll. 2009. Detecting loci under selection in a hierarchically structured population. Heredity 103:285-298.

Feng, X., A. Porporato, and I. Rodriguez-Iturbe. 2013. Changes in rainfall seasonality in the tropics. Nat Clim Chang 3: 811-815.

Fischer, M. C., C. Rellstab, and A. Tedder. 2013. Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. Mol Ecol 22:5594-5607.

Fournier-Level, A., A. Korte, M. D. Cooper, M. Nordborg, J. Schmitt, and A. M. Wilczek. 2011. A map of local adaptation in *Arabidopsis thaliana*. Science 334: 86-89.

Funatsuki H., M Suzuki., A. Hirose, A Inaba, T. Yamada, M. Hajika, K. Komatsu, T. Katayama, T. Sayama, M. Ishimoto, and K. Fujino. 2014. Molecular basis of a shattering resistance boosting global dissemination of soybean. PNAS 111:17797-17802.

Galmes J., H. Medrano, and J. Flexas. 2007. Photosynthesis and photoinhibition in response to drought in a pubescent (var. minor) and a glabrous (var. palaui) variety of *Digitalis minor*. Environ. Exp. Bot. 60:105-111.

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

Goudet, J. 2005. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. Mol. Ecol. Notes 5:184-186

Grant, D., R. T. Nelson, S. B. Cannon, and R. C. Shoemaker. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38:D843–D846.

Gurley, W. B. 2000. HSP101: a key component for the acquisition of thermotolerance in plants. Plant Cell 12:457-460.

Hancock, A.M., B. Brachi, N. Faure, M. W. Horton, L. B. Jarymowycz, F. G. Sperone, C. Toomajian, F. Roux, and J. Bergelson. 2011. Adaptation to climate across the *Arabidopsis thaliana* genome. Science 334: 83-86.

Harada, A., L.S.A. Gonçalves, R.A.S. Kiihl, and D. Destro. 2015. Flowering under short days: Juvenile period and inductive phase estimates in soybean genotypes. Agronomy Science and Biotechnology 1:10-16.

Harvell, C. D., C. E. Mitchell, J. R. Ward, S. Altizer, A. P. Dobson, R. S. Ostfeld, and M. D. Samuel. 2002. Climate warming and disease risks for terrestrial and marine biota. Science 296: 2158-2162.

Hengl, T., J. M. de Jesus, R. A. MacMillan, N. H. Batjes, G. B. M. Heuvelink, E. Ribeiro, A. Samuel-Rosa, B. Kempen, J. G. B. Leenaars, M. G. Walsh, and M. R. Gonzalez . 2014. SoilGrids1km--global soil information based on automated mapping. PLoS One 9: e105992.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25: 1965-1978.

Holsinger, K.E., and B.S. Weir. 2009. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nat. Rev. Genet. 10:639-650.

Hu, Y., R. Zhong., W. H. Morrison, and Z.H.Ye. 2003. The *Arabidopsis RHD3* gene is required for cell wall biosynthesis and actin organization. Planta 217:912-921.

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, et al. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961-967.

Huxley, J. 1938.Clines: an auxiliary taxonomic principle. Nature 142: 219-220.

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

Kawecki, T.J., and D. Ebert. 2004. Conceptual issues in local adaptation. Ecol Lett. 7:1225-1241.

Kihara, H. 1969. History of biology and other sciences in Japan in *Glycine max* retrospect. Proc. XII Intern. Cong. Genetics. 3:49-70

Kotak, S. J. Larkindale, U. Lee, P. von Koskull-Döring, E. Vierling, and K.D. Scharf. 2007. Complexity of the heat stress response in plants. Curr. Opin. Plant Biol. 10:310-316

Lasky J. R., D. L. Des Marais, J. K. McKay, J. H. Richards, T. E. Juenger, and T. H. Keitt. 2012. Characterizing genomic variation of *Arabidopsis thaliana*: The roles of geography and climate. Mol. Ecol. 21:5512-5529.

Lasky J.R., H. Upadhyaya, P. Ramu, S. Deshpande, T. Hash, J. Bonnette, T. Juenger, K. Hyma, C. Acharya, S. Mitchell, E. Buckler, Z. Brenton, S. Kresovich, and G. Morris. 2016. Genome-environment associations in sorghum landraces predict adaptive traits. Sci Adv 1. doi:10.1126/sciadv.1400218

Lasky, J. R., D.L. Des Marais, D. B. Lowry et al.2014. Natural variation in abiotic stress responsive gene expression and local adaptation to climate in *Arabidopsis thaliana*. Mol Biol and Evol 31:2283-2296.

Lata, C., and M. Prasad. 2011. Role of DREBs in regulation of abiotic stress responses in plants. J. Exp. Bot. doi:10.1093/jxb/err210.

Lauter, A. N., G. A. Peiffer, T. Yin, S. A. Whitham, D. Cook, R. C. Shoemaker, and M. A. Graham. 2014. Identification of candidate genes involved in early iron deficiency chlorosis signaling in soybean (*Glycine max)* roots and leaves. BMC Genomics 15:702.

Lewontin, R. C., and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175-195.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R Durbin. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

Li, J., and L. Ji. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb.) 95:221-227.

Li, Z.L., and R.L. Nelson. 2001. Genetic diversity among soybean accessions from three countries measured by RAPDs. Crop Sci. 41:1337-1347.

Liakopoulos, G., S. Stavrianakou, and G. Karabourniotis. 2006. Trichome layers versus dehaired lamina of *Olea europaea* leaves: differences in flavonoid distribution, UV-absorbing capacity, and wax yield. Environ. Exp. Bot. 55:294-304.

Lippert, C., J. Listgarten, Y. Liu, C.M. Kadie, R.I. Davidson, and D. Heckerman. 2011. FaST linear mixed models for genome-wide association studies. Nature Methods 8:833-U94.

Meyer, R.S. J. Y. Choi, M. Sanches, A. Plessis, J. M. Flowers, J. Amas, K. Dorph, A. Barretto, B. Gross, D. Q. Fuller, I. K. Bimpong, M. Ndjiondjop, K. M. Hazzouri, G. B. Gregorio, and M. D. Purugganan. 2016. Domestication history and geographical adaptation inferred from a SNP map of African rice. Nat. Genet. doi:10.1038/ng.3633.

Mitton, J.B., Y.B. Linhart, J. L. Hamrick, and J.S. Beckman. 1977. Observations on genetic structure and mating system of ponderosa pine in Colorado front range. Theor Appl Genet 51: 5-13.

Mourtzinis, S., J. Specht, L. Lindsey, W. Wiebold, J. Ross, E D. Nafziger, H. Kandel, N. Mueller, P. Devillez, F. Arriaga, and S. Conley. 2015. Climate-induced reduction in US-wide soybean yields underpinned by region- and in-season specific responses. Nat. Plants doi: 10.1038/nplants.2014.26.

Oksanen, J., R. Kindt, P. Legendre, B. O'Hara, M.H.H.Stevens, and M. J. Oksanen. 2007. The vegan package. Community ecology package. 10:631-637

Pandey, N., A. Ranjan,, P. Pant, R. K. Tripathi, F. Ateek, H. P. Pandey, U.D. Patre, and S. V. Sawant. 2013. CAMTA 1 regulates drought responses in Arabidopsis thaliana. BMC Genomics 14:216.

Peres-Neto P. R., P. Legendre, S. Dray, and D. Borcard. 2006. Variation partitioning of species data matrices: Estimation and comparison of fractions. Ecology 87:2614-2625.

Purcell S., Neale B., K. Todd-Brown, L. Thomas, M. A.Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P.C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559-575.

Qin, F., Y. Sakuma, L.S. Tran, K. Maruyama, S. Kidokoro, Y. Fujita, M. Fujita, T. Umezawa, Y. Sawano, K. Miyazono, M. Tanokura, K. Shinozaki, K. Yamaguchi-Shinozaki. 2008. Arabidopsis DREB2A interacting proteins function as RING E3 ligases and negatively regulate plant drought stress-responsive gene expression. Plant Cell 20:1693-1707.

Quinlan A. R. and I.M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

R Development Core Team. 2015. R: a language and environment for statistical computing. http://www.R-project.org.

Raj A., M. Stephens, and J. K. Pritchard. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197: 573-589.

Raunser, S., R. Magnani, Z. Huang, R. L. Houtz, R. C. Trievel, P. A. Penczek, and T. Walz. 2009. Rubisco in complex with Rubisco large subunit methyltransferase. PNAS 106: 3160-3165.

Rellstab, C., F. Gugerli, A. Eckert, A. Hancock, and R. Holderegger. 2015. A practical guide to environmental association analysis in landscape genomics. Mol. Ecol. Res 24:4348-4370.

Sakuma, Y., K. Maruyama, Y. Osakabe, F. Qin, M. Seki, K. Shinozaki, and K. Yamaguchi-Shinozaki. Functional analysis of an *Arabidopsis* transcription factor, DREB2A, involved in drought-responsive gene expression. Plant Cell 18:1292-309.

Savolainen, O., M. Lascoux, and J. Merilca. 2013. Ecological genomics of local adaptation. Nature Rev Genet. 14:807-820.

Schmutz, J., S. B. Cannon, J. Schlueter, J. Ma, T. Mitros et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463: 178-183.

Sidak, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc. 62:626-633

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, and P.B. Cregan. 2013. Development and evaluation of SoySNP50K, a highdensity genotyping array for soybean. PLoS ONE 8:e54985.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Ficus, R.L. Nelson, and P.B. Cregan. 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3 5:1999-2006.

Specht, J. E., B. W. Diers, R. L. Nelson, J. Francisco, F. de Toledo, J. A. Torrion, and Patricio Grassini. 2014. Yield Gains in Major US Field Crops CSSA. Special Publication 33 (eds Smith, S., Diers, B., Specht, J. & Carver, B.) Ch. 12, 199–243 (American Society of Agronomy, Crop Science Society of America, Soil Science Society of America).

Strasburg, J. L., N. A. Sherman, K. M. Wright, L. C. Moyle, J. H. Willis, L. H. Rieseberg. 2012. What can patterns of differentiation across plant genomes tell us about adaptation and speciation? Philosophical Transactions of the Royal Society B-Biological Sciences 367:364-373.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299-1320.

Tiffin, P., and J. Ross-Ibarra. 2014. Advances and limits of using population genetics to understand local adaptation. Trends Ecol. Evol. 29: 673-680.

Todd, J.J. and L.O. Vodkin. 1996. Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. Plant Cell 8:687-699

Turesson, G. 1922. The genotypical response of the plant species to the habitat. Hereditas 3: 211-350.

Turner, S. D. 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv DOI: 10.1101/005165.

Tuteja, J.H., G. Zabata, K. Varala, M. Hudson, and L.O. Vodkin. 2009. Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in *Glycine max* seed coats. Plant Cell 21:3063-3071.

Ude, G.N., W.J. Kenworthy, J.M. Costa, P.B. Cregan, and J. Alvernaz. 2003. Genetic diversityof soybean cultivars from China, Japan, North America, and North American ancestral lines determined by amplified fragment length polymorphism. Crop Sci. 43:1858-1867.

van Damme M., R. P. Huibers, J. Elberse, and G. Van den Ackerveken. 2008. Arabidopsis DMR6 encodes a putative 2OG-Fe(II) oxygenase that is defense-associated but required for susceptibility to downy mildew. Plant J 54:785-793

van den Wollenberg, A. L. 1977. Redundancy analysis an alternative for canonical correlation analysis. Psychometrika 42, 207-219.

Wang, N., J.M. Akey, K. Zhang, R. Chakraborty, and L. Jin. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. Am. J. Hum. Genet. 71:1227-1234.

Weir, B. S., and C. C. Cockerham. 1984. Estimating *F*-Statistics for the analysis of population structure. Evolution 38:1358-1370.

Wen, Z., J.F. Boyse, Q. Song, P.B. Cregan, and D. Wang. 2015. Genomic consequences of selection and genome-wide association mapping in soybean. BMC Genomics 16:671.

Westengen, O.T., P. R. Berg, M.P., Kent, and A.K. Brysting. 2012. Spatial structure and climatic adaptation in African maize revealed by surveying SNP diversity in relation to global breeding and landrace panels. PLoS ONE 7:e47832.

Williams, G. C.1966. Adaptation and Natural Selection. Princeton University Press, Princeton, New Jersey.

Wilson, R.F. 2008. Soybean: Market driven research needs. In: G. Stacey, editor, Genetics and genomics of soybean. Springer Science+Business Media, New York. p. 3–15

Yamori W., K. Hikosaka, and D.Way. 2014. Temperature response of photosynthesis in C3, C4, and CAM plants: temperature acclimation and temperature adaptation. Photosynth Res 119:101-117.

Yan A., J. Pan, L. An, Y. Gan, and H. Feng. 2012. The responses of trichome mutants to enhanced ultraviolet-B radiation in *Arabidopsis thaliana*. J. Photochem. Photobiol. B Biol. 113:29-35.

Yang, W.Y., J. Novembre, E. Eskin, and E. Halperin. 2012. A model-based approach for analysis of spatial structure in genetic data. Nature Genetics 44: 725–731

Yoder, J. B., J. Stanton-Geddes, P. Zhou, R. Briskine, N. D. Young, and P. Tiffin. 2014. Genomic signature of adaptation to climate in *Medicago truncatula*. Genetics 196:1263-1275.

Yuen C.Y., J. C., Sedbrook, R. M. Perrin, K. L. Carroll, P.H. Masson. 2005. Loss-of-function mutations of ROOT HAIR DEFECTIVE3 suppress root waving, skewing, and epidermal cell file rotation in Arabidopsis. Plant Physiol. 138:701-714.

4.8 FIGURES

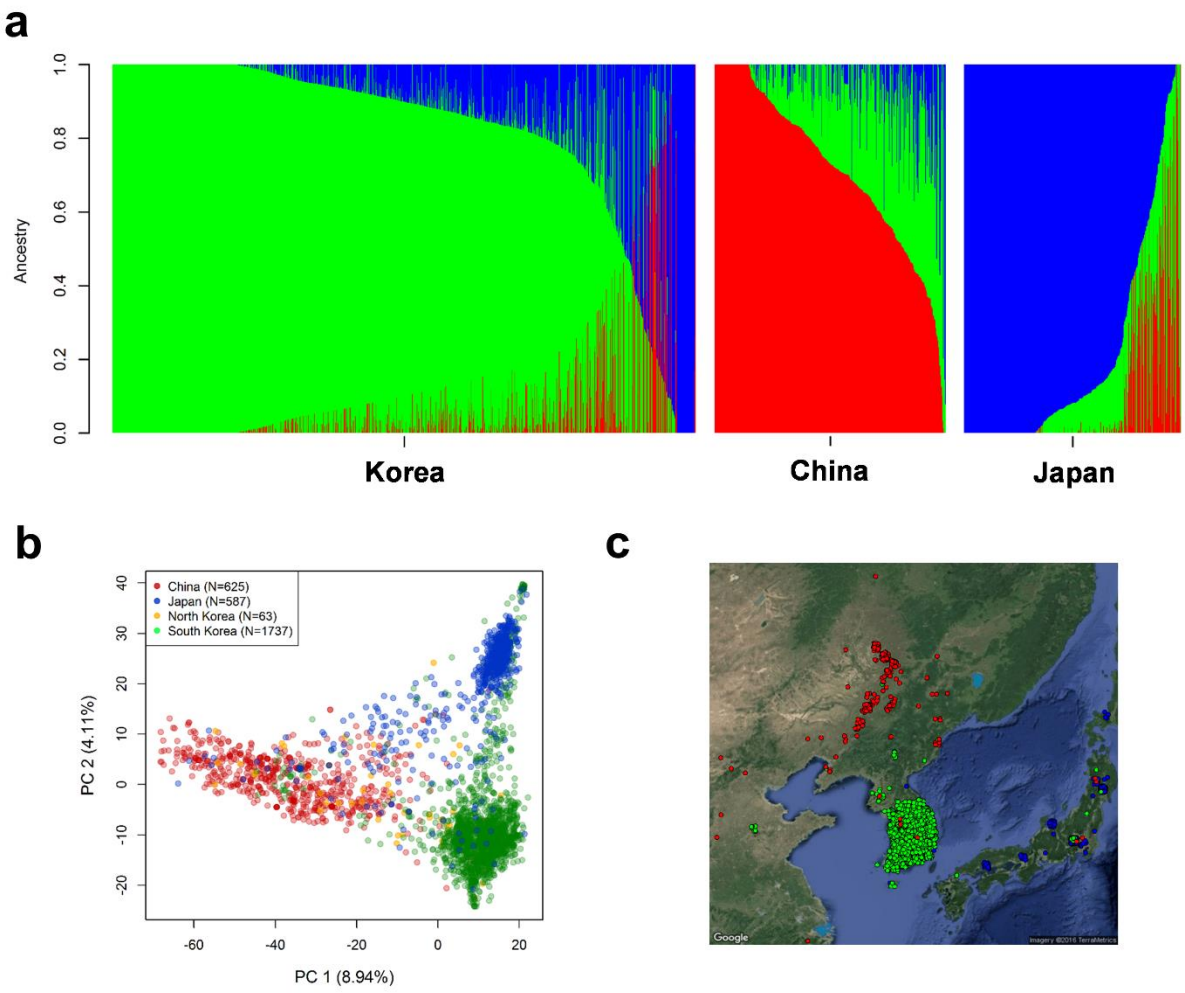


Figure 1. Population structure in the soybean landrace collection inferred by *fastSTRUCTURE* and principal component analysis. a) The number of clusters (K) present in the entire population of 3,012 accessions was judged to be K=3. Each colored vertical line in the barplot represents an individual accession that was assigned proportionally to the one of the three clusters. b) Principal component (PC) analysis of 3,012 landrace accessions. The top two PCs accounted for geographic origin differences which explained ~13 % of total genetic variation. c) The geographical location in which each landrace accession (with subpopulation ancestry > 80%) was collected. The spot colors correspond to the *fastSTRUCTURE* assignment of each accession. The assignment of samples into three genetic clusters generally accords with geography. The spots have been jittered to show overlapping samples.

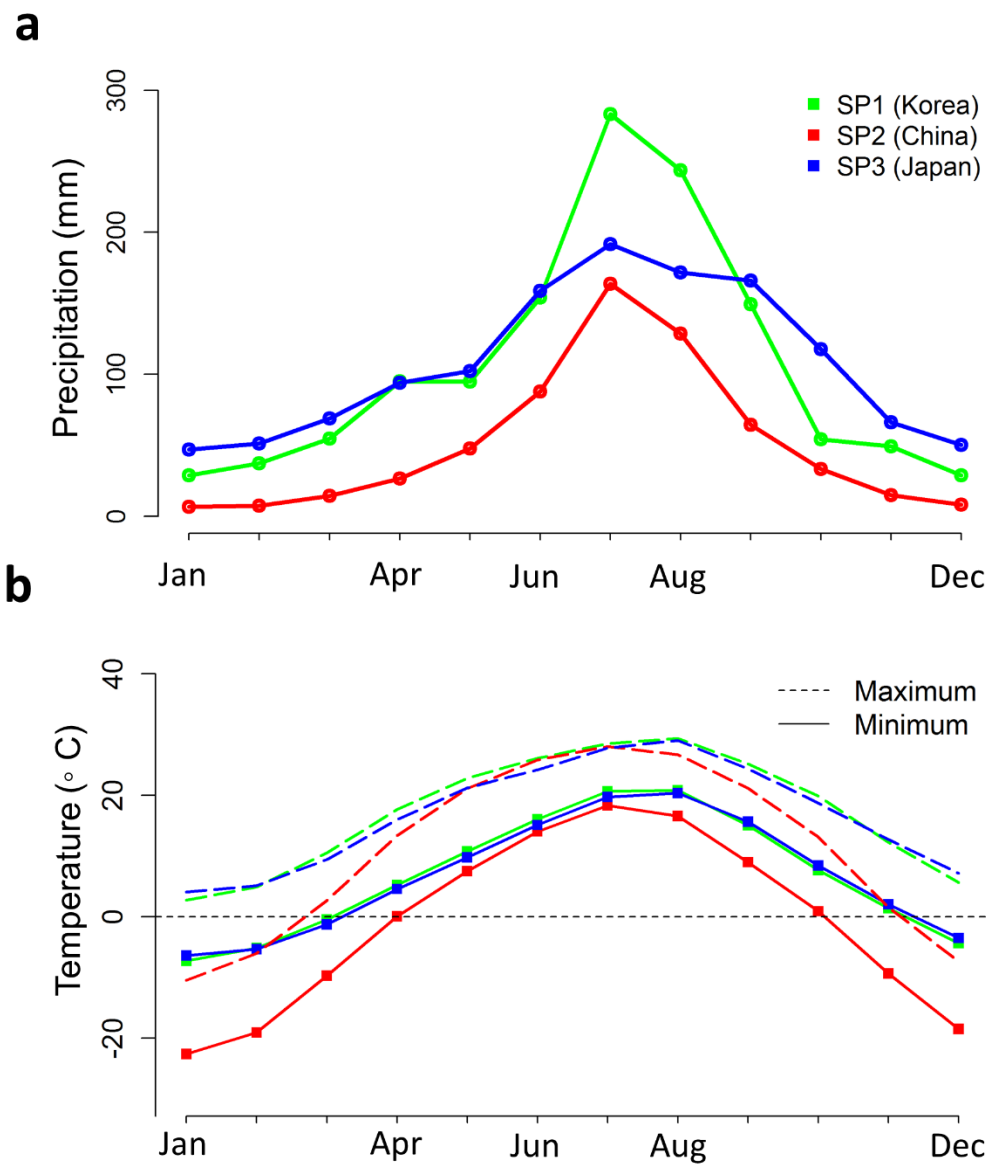


Figure 2. Monthly series analysis for mean precipitation and minimum and maximum temperatures for each of the three subpopulations inferred from *fastSTRUCTURE.* SP1 represents accessions collected from Korea; SP2 represents accessions collected from China; SP3 represents accessions collected from Japan.

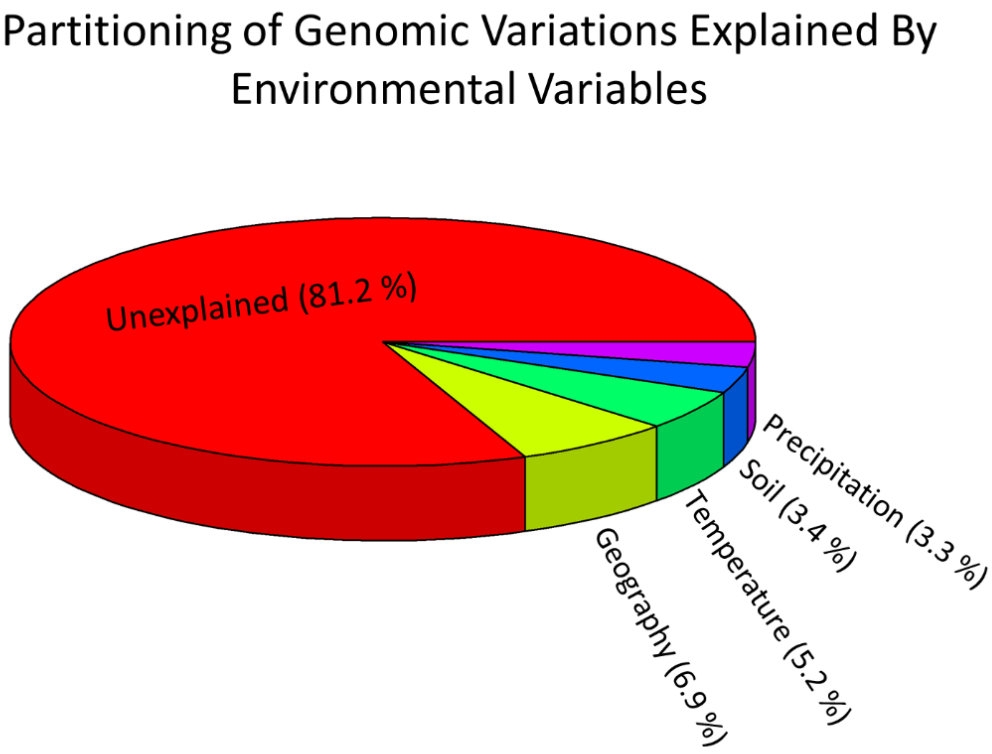


Figure 3. Partitioning of genomic variation due to environmental variation and geographic variables using a partial redundancy analysis. Genomic variation was partitioned based on four categories of grouped environmental variables (geography, temperature, precipitation, soil).

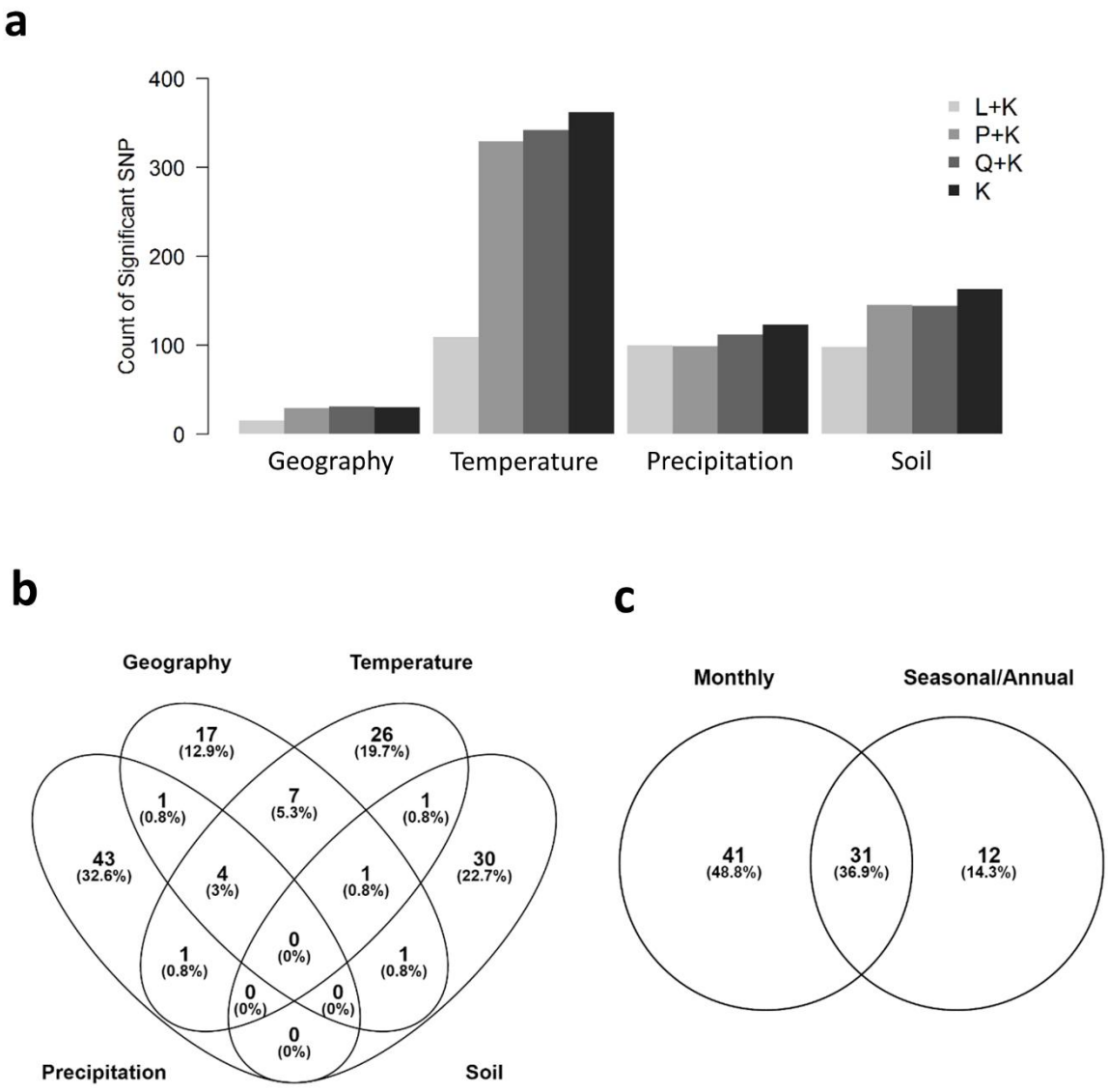


Figure 4. Genome-wide association mapping of 112 environmental variables using 3,012 landrace accessions. a) Summary of genome-wide significant associations identified using four linear mixed models. b) Summary of unique and overlapped significant associations among four categories (geography, temperature, precipitation, soil) of environmental variables. c) Summary of unique and overlapped significant associations between monthly and seasonal/annual variables.

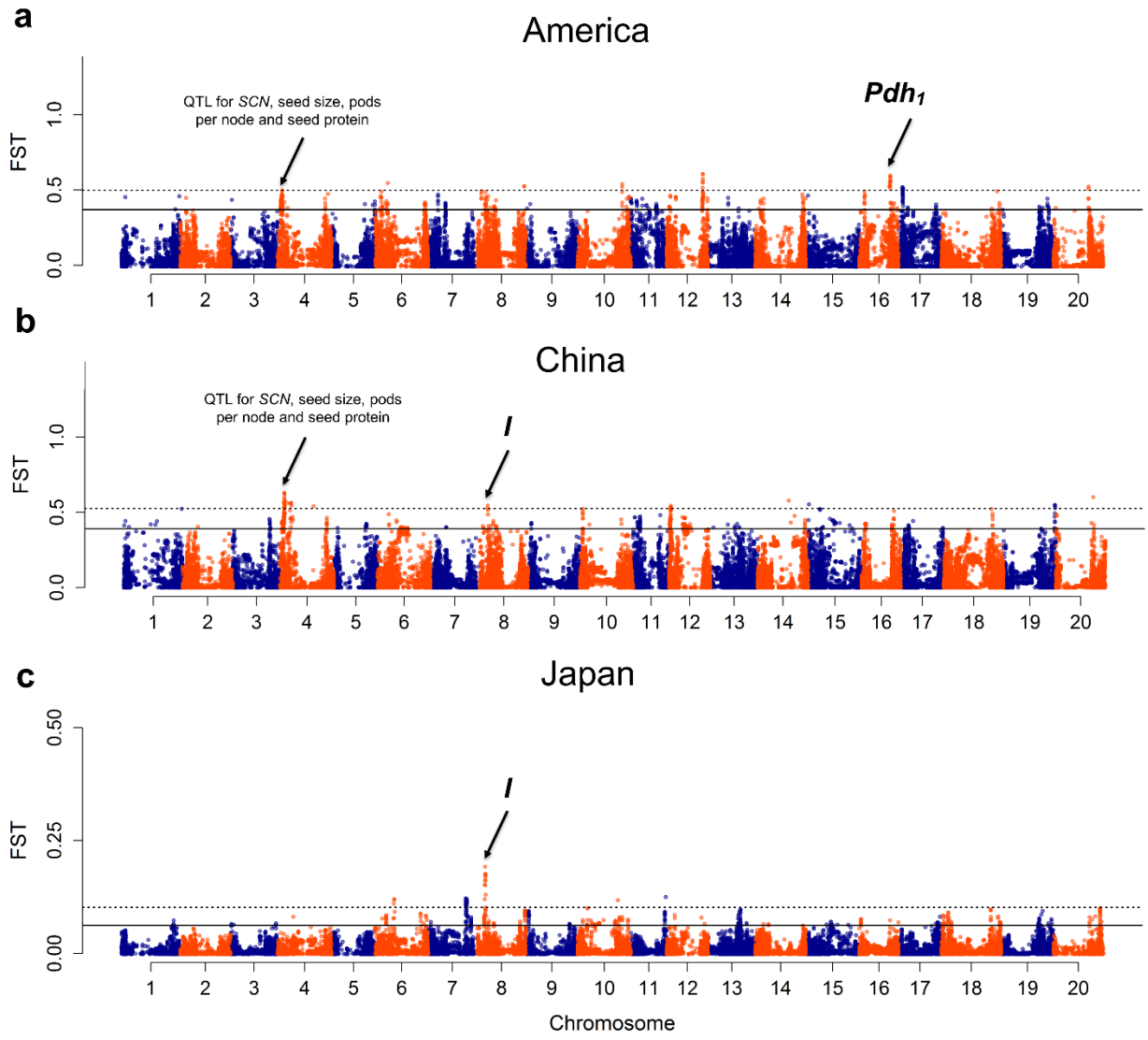


Figure 5. Differential selection between elite and landrace population within a) America, b) China and c) Japan using F_{ST} analysis. The F_{ST} values are plotted against the base pair position on 20 chromosomes of soybean. The dashed horizontal line denotes the calculated F_{ST} value based based on 99.9th percentile for declaring a selected region. The solid horizontal line denotes the calculated F_{ST} value based on 99th percentile for declaring a selected region. Strong selection signals that co-localized with known genes or QTL are indicated by an arrow.

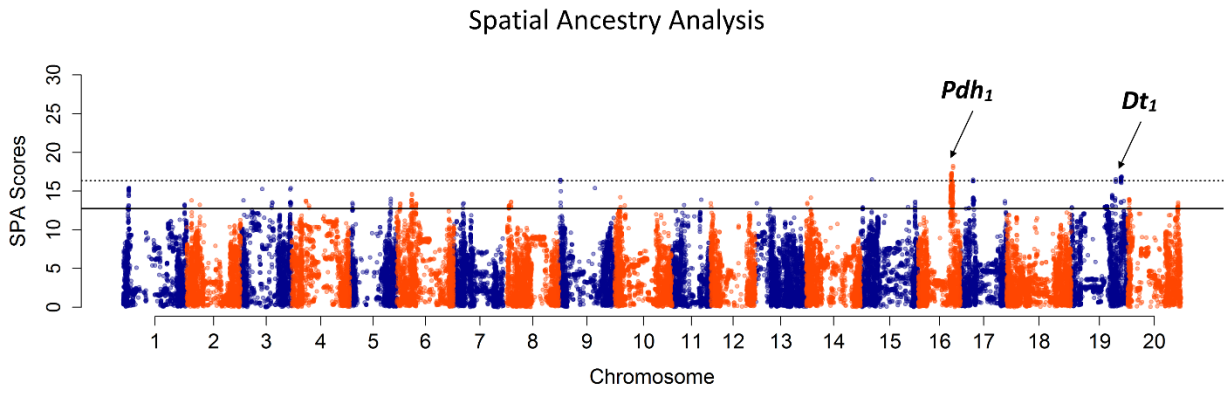


Figure 6. Spatial ancestry analysis (SPA) using 3,012 landraces within *G. max* accessions. The SPA selection scores values are plotted against the base pair position on 20 chromosomes of soybean. Strong selection signals that co-localized with known genes or QTL are indicated by an arrow. The dashed horizontal line denotes the calculated SPA threshold score based on 99.9th percentile for declaring a selected region. The solid horizontal line denotes the calculated SPA threshold score based on 99th percentile for declaring a selected region.

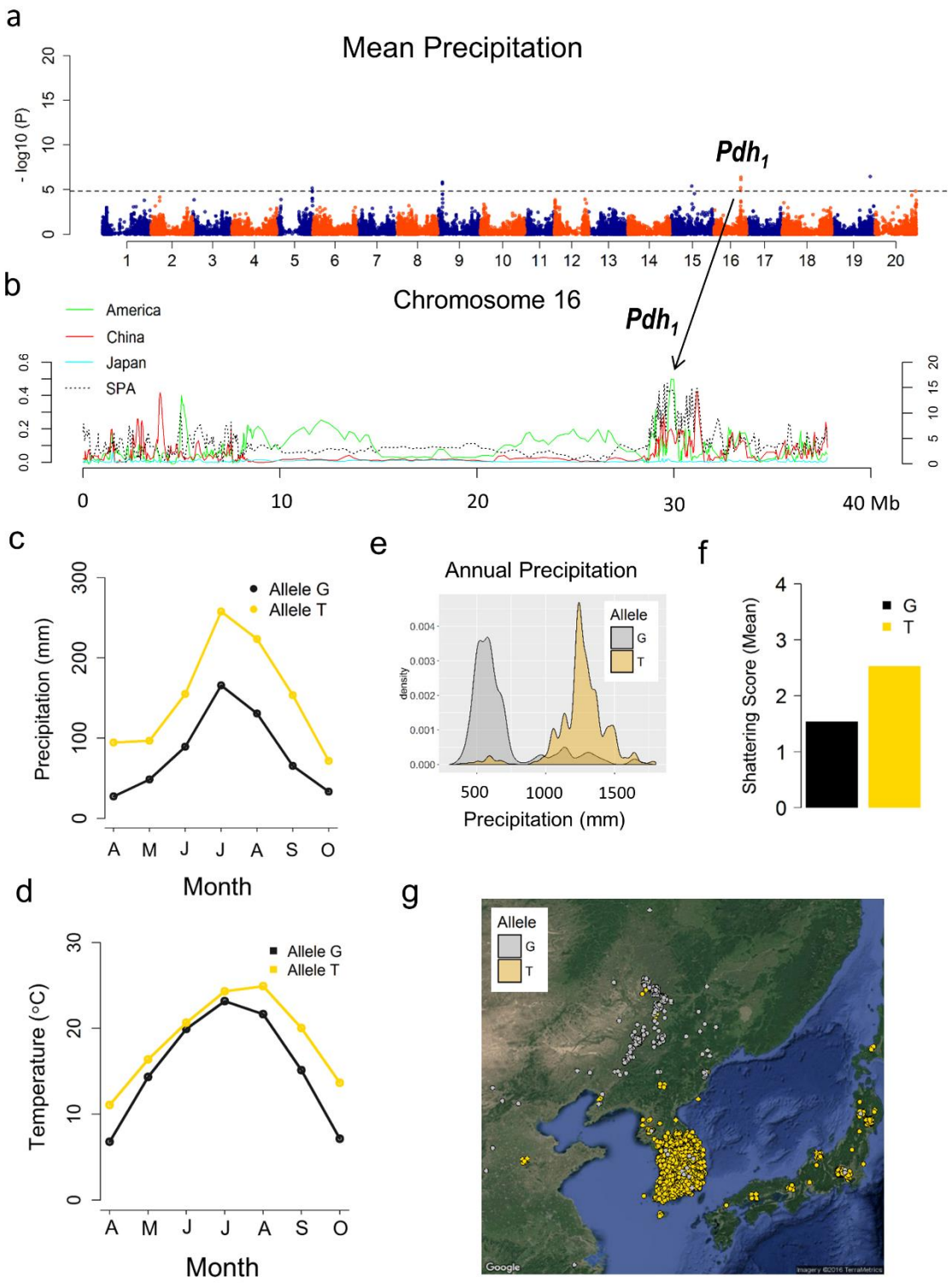


Figure 7. Environmental association and spatial ancestry analysis (SPA) and F_{ST} analyses identified a selected region on chromosome 16 of the *Glycine max* genome. a) Environmental association identified the *Pdh1* locus for mean precipitation, b) SPA and

F_{ST} values were plotted based on a sliding-window approach. The gray solid vertical line denotes the position of significant associations. Notably, highest SPA and F_{ST} values overlapped with significant associations between 29517407 - 31181902 Bp that co-localized with *pdh1*, a major QTL responsible for the reduction of pod shattering in soybean (Fonatsuki et al., 2014). c, d, e) Significant allelic effects of nearest SNP tagging *pdh1* was more significant for precipitation than temperature, f) Allelic effects indicate that G allele is favorable for shattering resistance, g) Geographic location of individuals with the "G" allele (gray) or "T" allele (gold) with jitter added to show overlapping samples. Only individuals with ancestry >80% based on *fastSTRUCTURE* results was plotted. Allelic frequency distribution in three subpopulations defined by *fastSRUCTURE*. The G allele was directionally selected in SP2 (China) while it was selected against in SP1 (Korea) and SP3 (Japan).

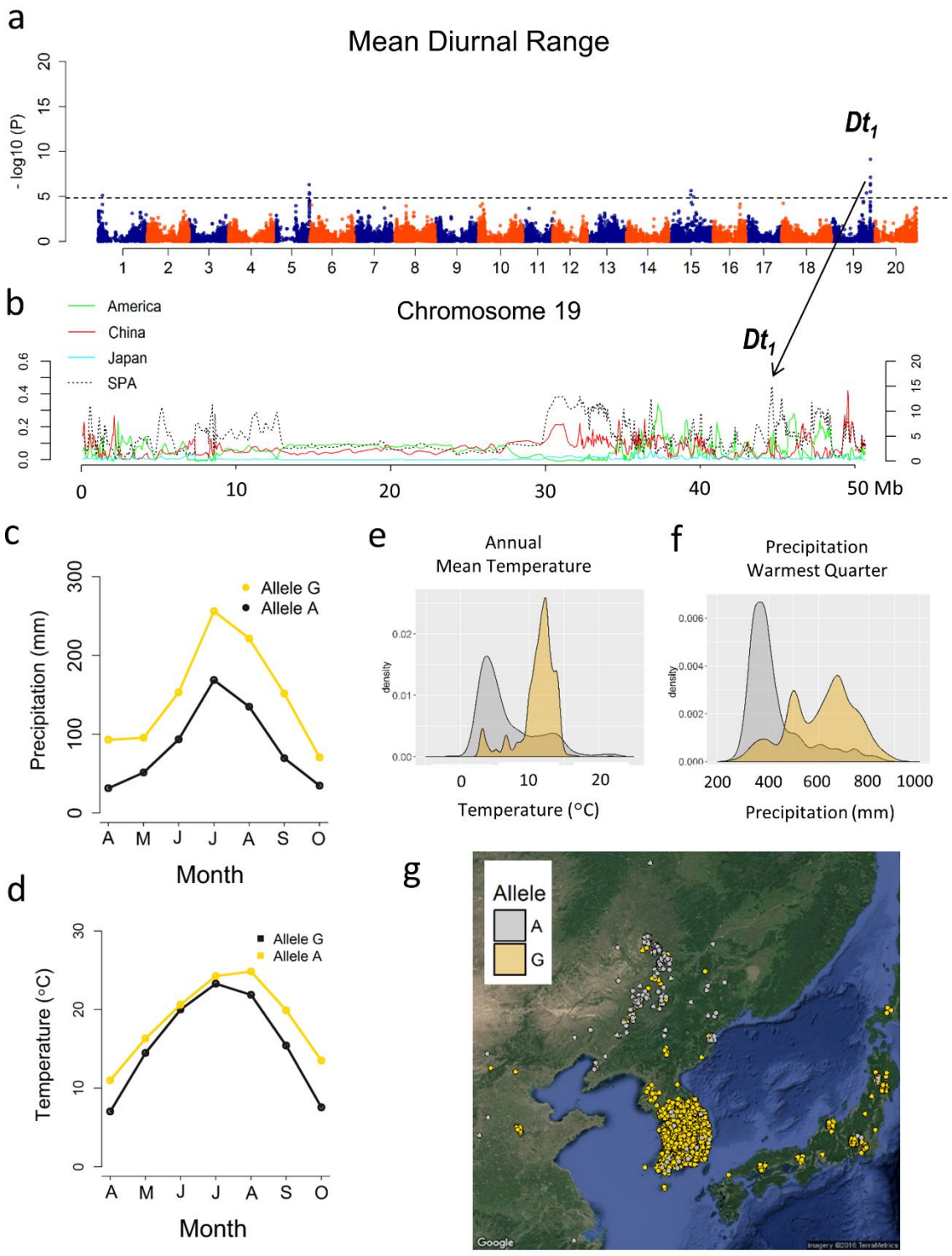


Figure 8. Environmental association and spatial ancestry analysis (SPA) and F_{ST} analyses identified a selected region on chromosome 19 of the *Glycine max* genome. a) Environmental association identified the *Dt1* locus for mean diurnal range. b) SPA and

F_{ST} values were plotted based on a sliding-window approach. The gray solid vertical line denotes the position of significant associations. Notably, highest SPA and F_{ST} values overlapped with significant associations between 29517407 - 31181902 Bp that co-localized with *Dt1*, a major QTL responsible for the reduction of pod shattering in soybean (Fonatsuki et al., 2014). c, d, e, f) Significant allelic effects of nearest SNP tagging *Dt1* was more significant for precipitation than temperature. g) Geographic location of individuals with the "G" allele (gray) or "A" allele (gold) with jitter added to show overlapping samples. Only individuals with ancestry >80% based on *fastSTRUCTURE* results was plotted. Allelic frequency distribution in three subpopulations defined by *fastSRUCTURE*. The G allele was directionally selected in SP2 (China) while it was selected against in SP1 (Korea) and SP3 (Japan).

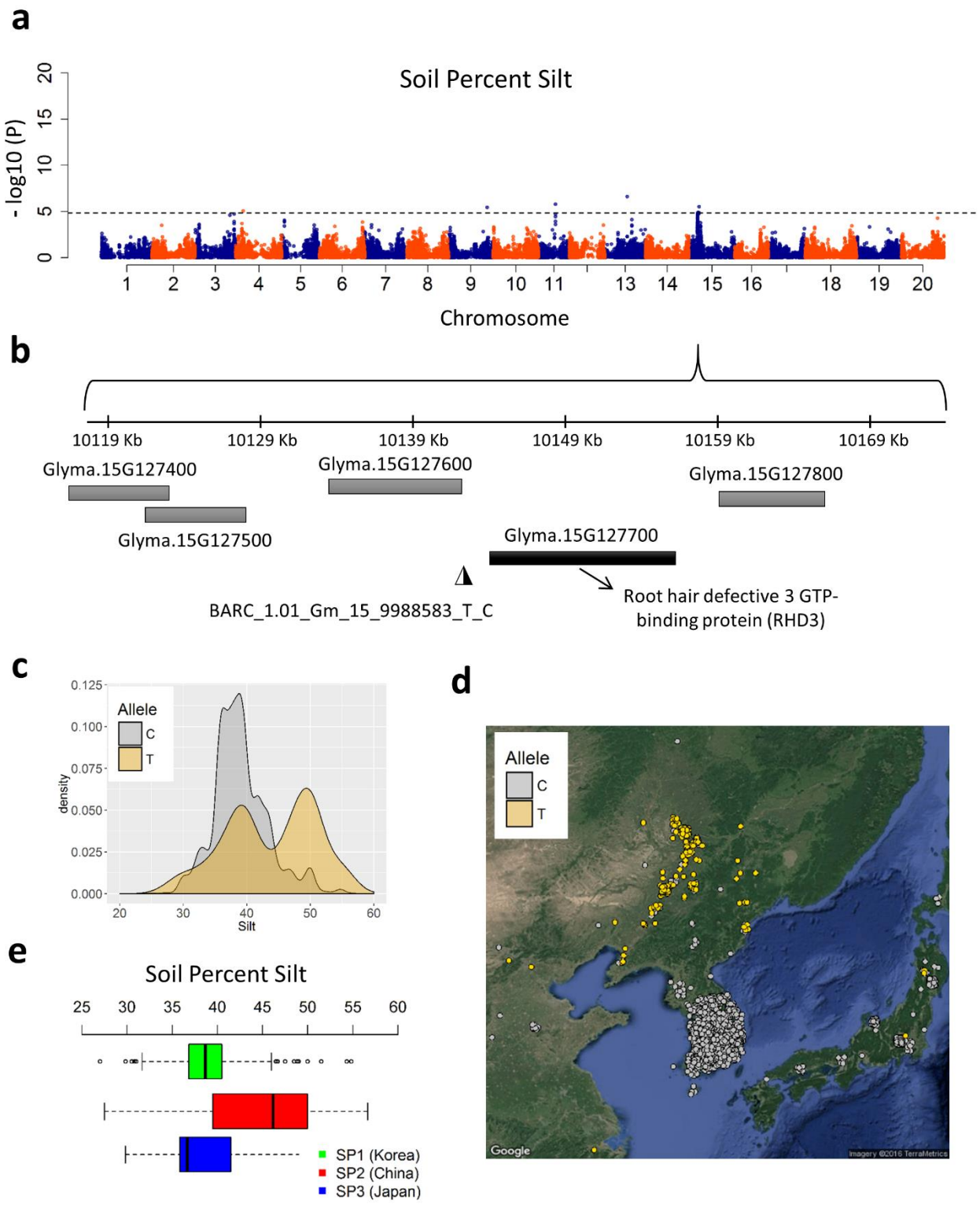
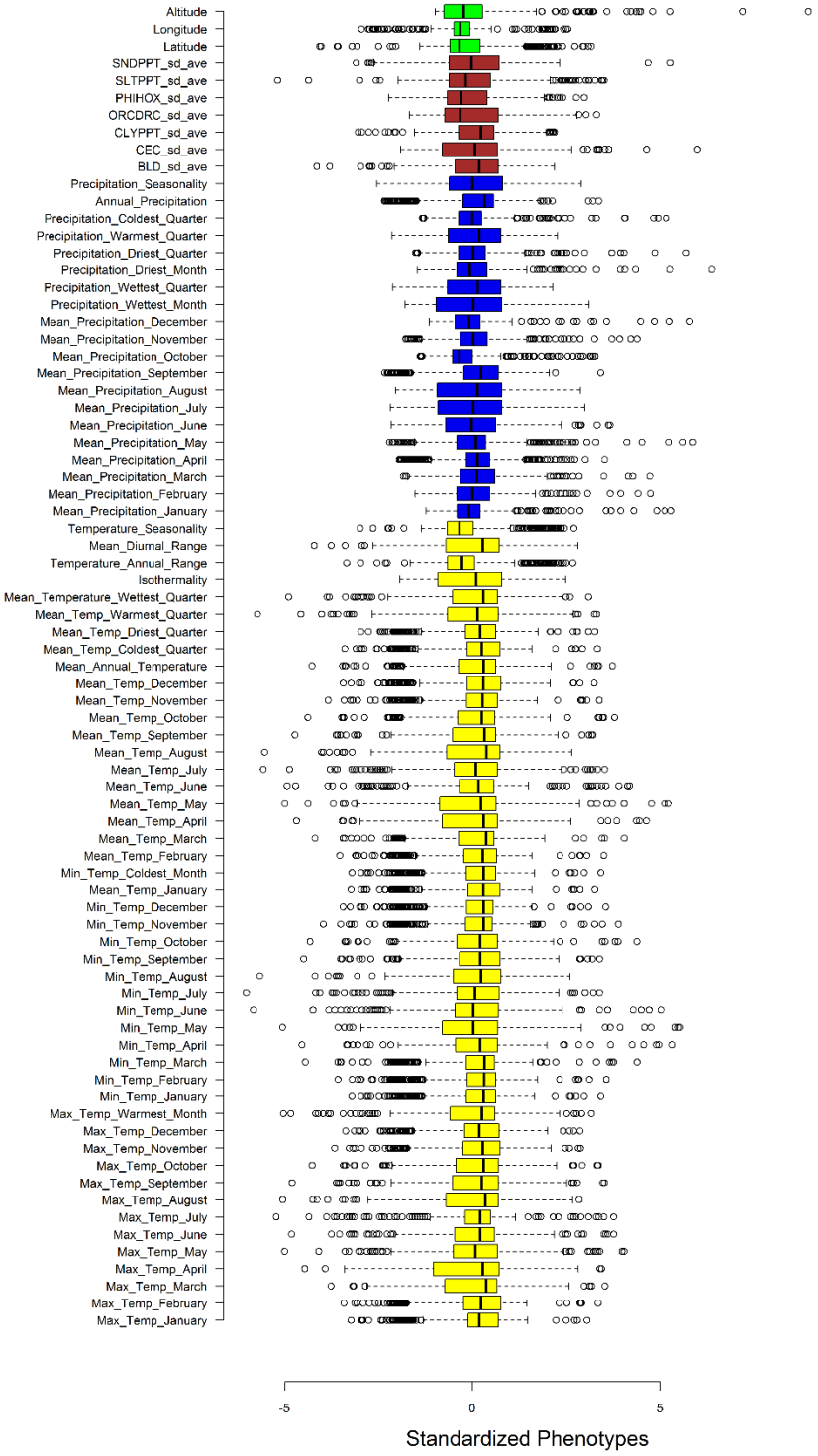
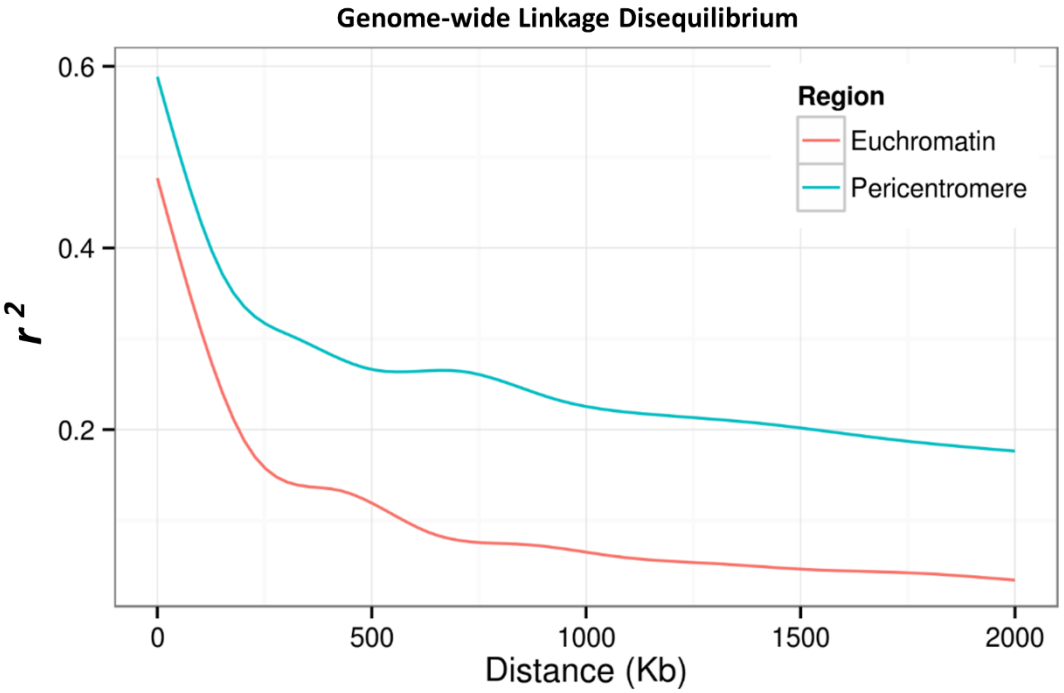

a
Soil Percent Silt
- log10 (P)
0
5
10
15
20
1
2
3
4
5
6
7
8
9
10
11
13
14
15
16
18
19
20
Chromosome
b
10119 Kb
10129 Kb
10139 Kb
10149 Kb
10159 Kb
10169 Kb
Glyma.15G127400
Glyma.15G127500
Glyma.15G127600
Glyma.15G127700
Glyma.15G127800
BARC_1.01_Gm_15_9988583_T_C
Root hair defective 3 GTP-binding protein (RHD3)
c
Allele
C
T
density
0.125
0.100
0.075
0.050
0.025
0.000
20
30
40
50
60
Silt
d
Allele
C
T
Google
e
Soil Percent Silt
25
30
35
40
45
50
55
60
SP1 (Korea)
SP2 (China)
SP3 (Japan)

Figure 9. Genome-wide association results of soil percent silt. a) Genome wide view of association results for soil percent silt. A cluster of significant associations was identified on chromosome 15 across six soil depths. b) Zoom in on 50 kb region around the significant marker BARC_1.01_Gm15_9988583_T_C, the most significant hit for soil percent silt. The Arabidopsis ortholog for the nearest gene, Glyma.15g127700, is Root hair defective 3 GTP-binding protein (RHD3), a gene that affects root epidermis development and is required for appropriate root and root hair cells enlargement in *Arabidopsis* (Lockwood et al., 1997; Zhong et al., 2003; Yuen et al., 2005). c) Density plot of allele frequency distribution for Percent Silt. The "T" allele at this locus is associated with high silt environments while the C 'allele' is associated with low silt environment. d) Geographic location of individuals with the "C" allele (gray) or "T" allele (gold) with jitter added to show overlapping samples. Only individuals with ancestry >80% based on *fastSTRUCTURE* results was plotted. The "T" allele is absent in SP1 and SP3 while it is directionally selected in SP2, which is predominated by accessions from China. e) Boxplot analysis of variation among subpopulations for soil percent silt.

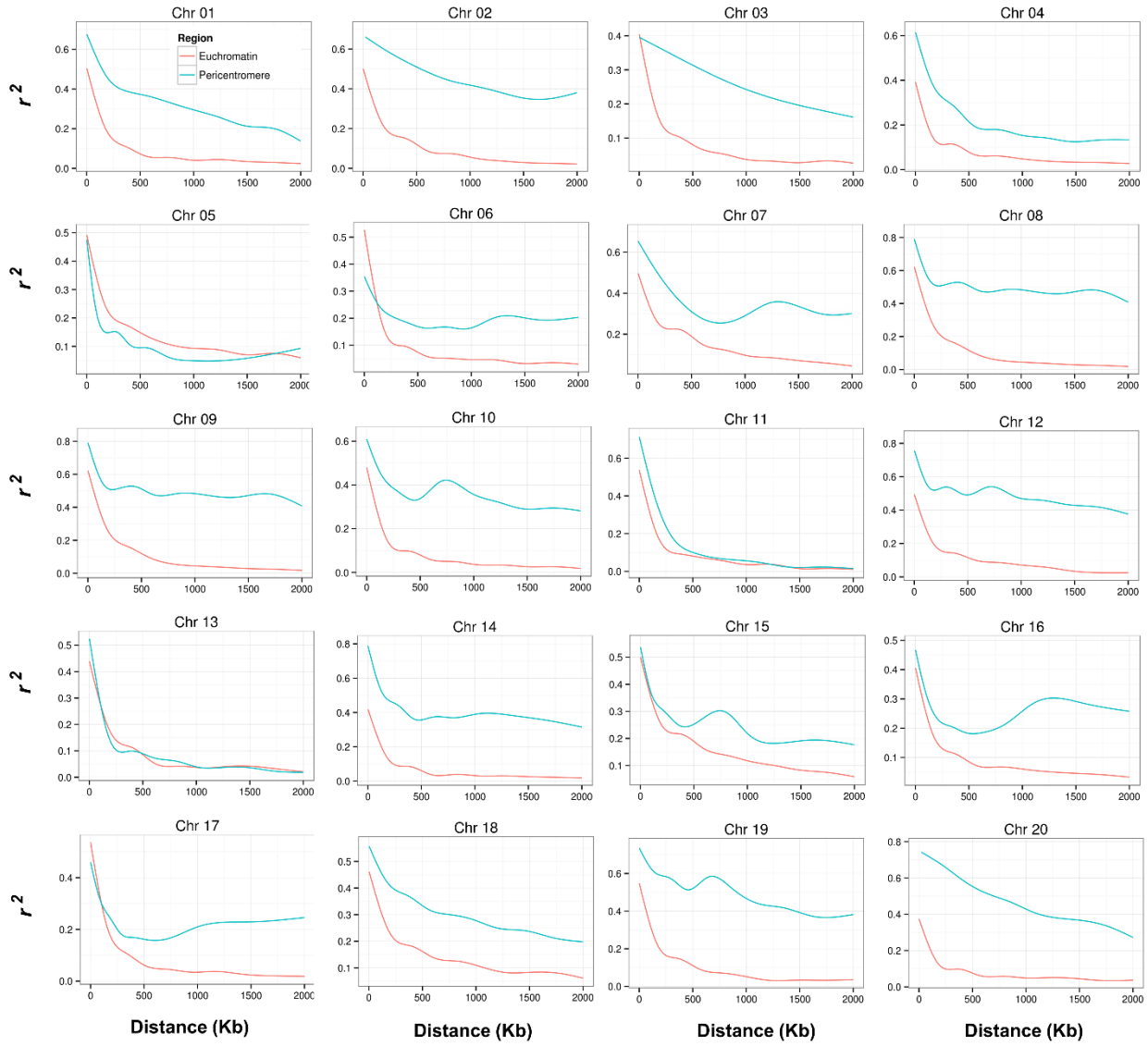
4.9 APPENDIX



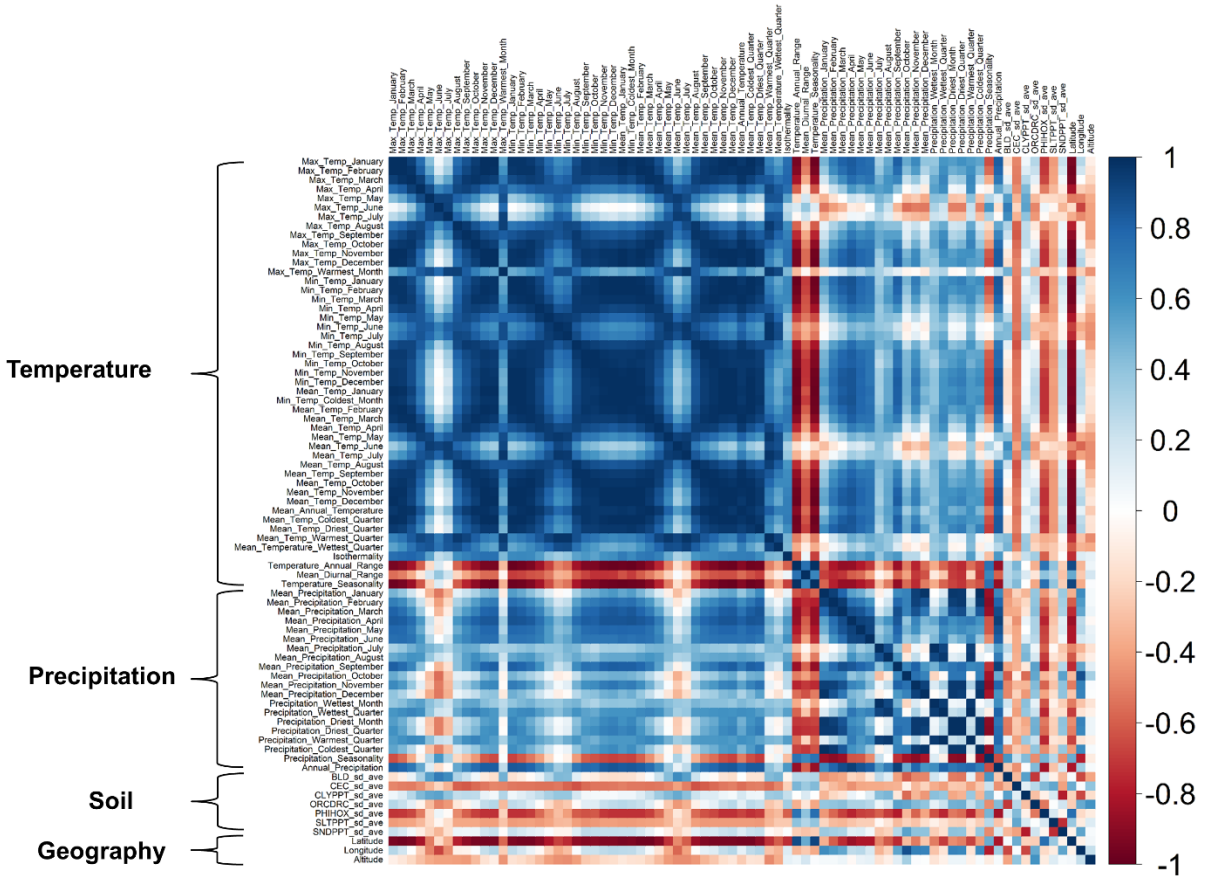
Supplementary Figure 1. Standardized distributions of spatial (green), soil (brown), precipitation (blue) and temperature (yellow) variables.



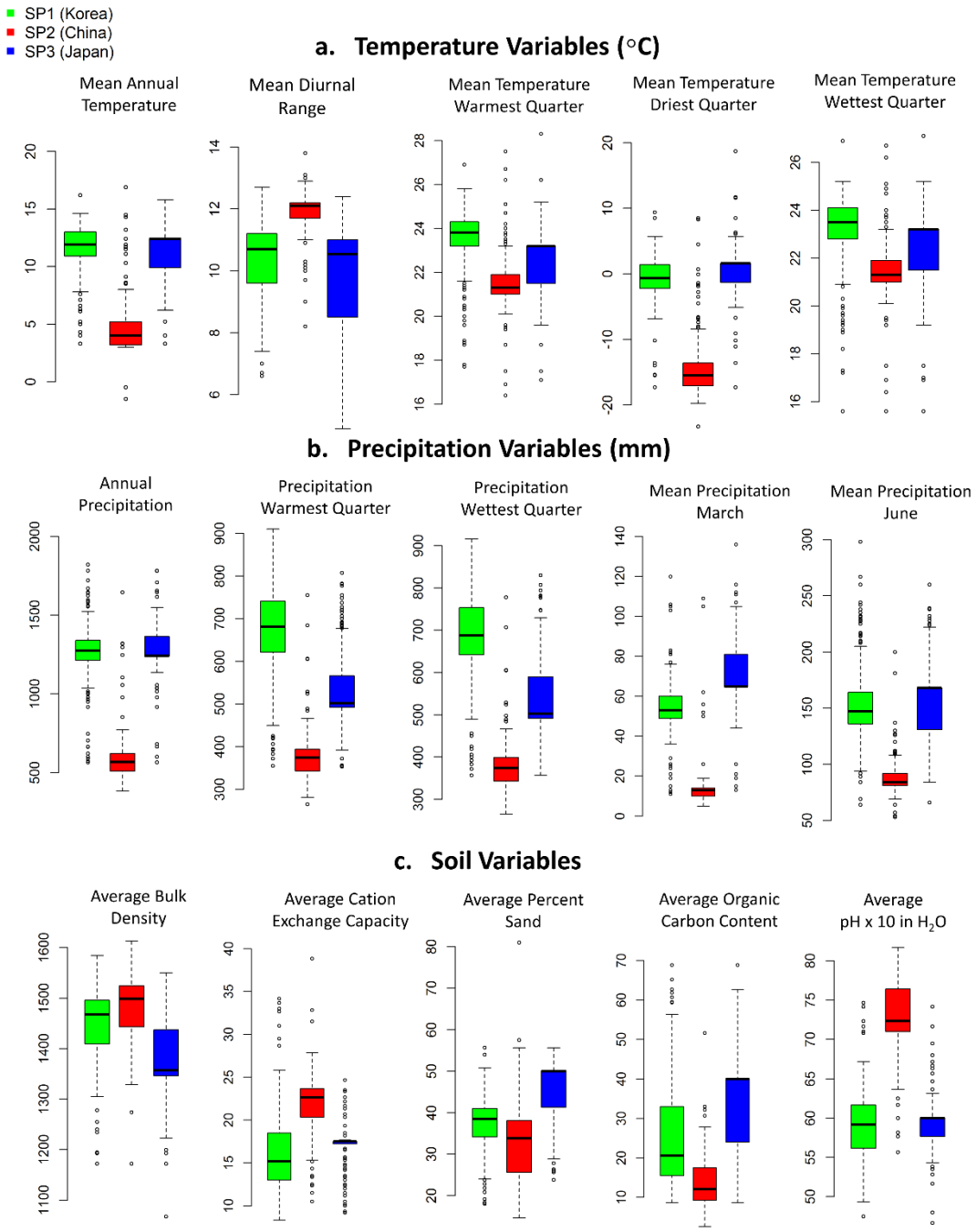
Supplementary Figure 2. Genome-wide linkage disequilibrium decay in 3,012 landrace *G. max* accessions.



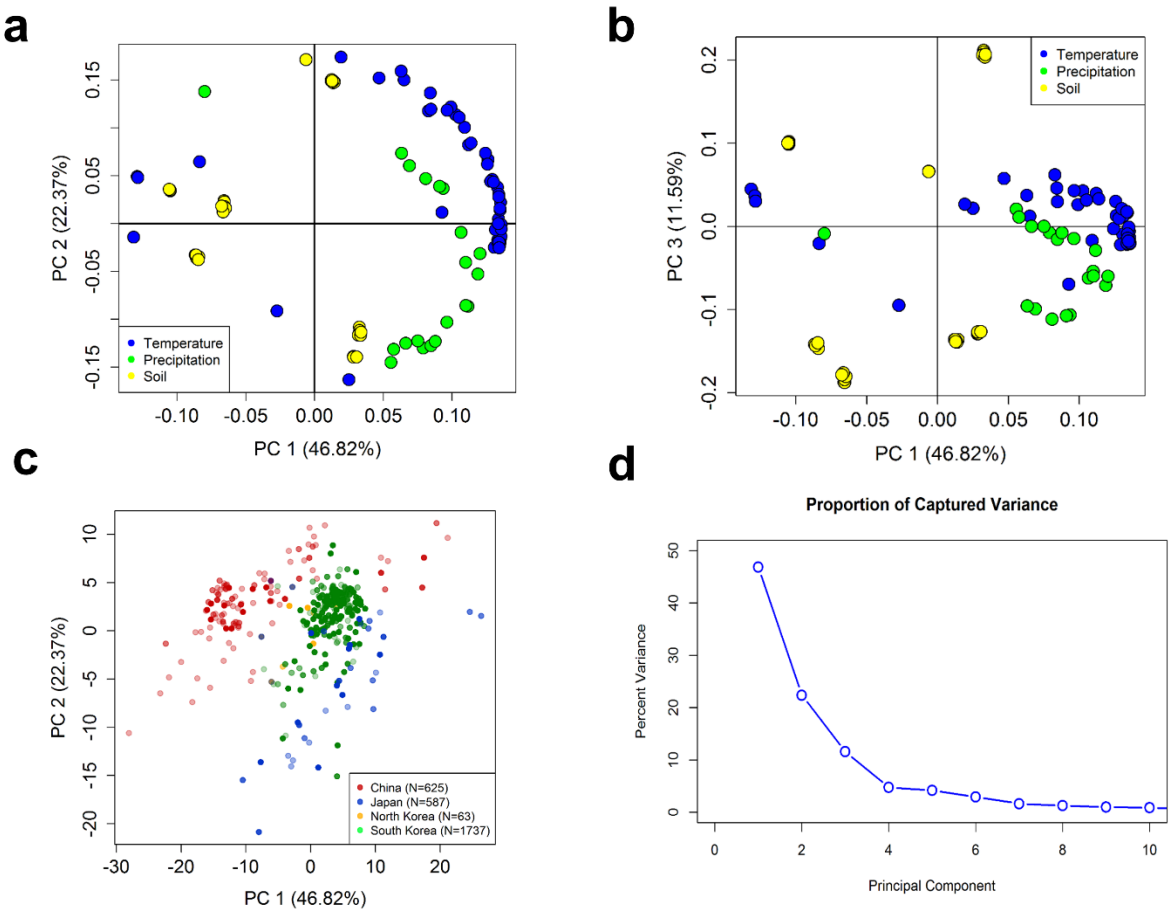
Supplementary Figure 3. Intra-chromosomal pattern of linkage disequilibrium decay in 3,012 landrace *G. max* accessions.



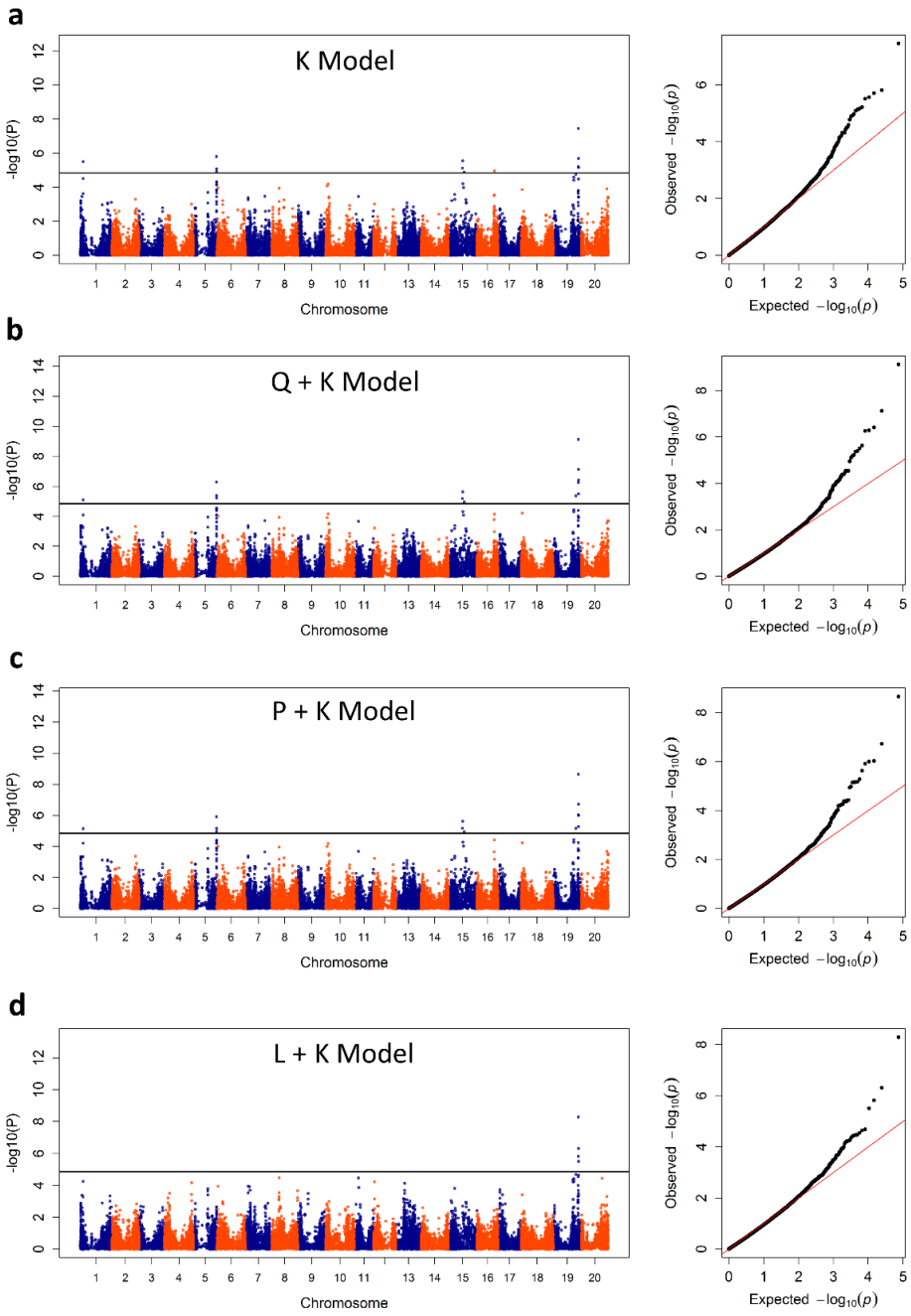
Supplementary Figure 4. Pearson correlation between biophysical and bioclimatic variables. Blue indicates a high positive correlation, white indicates a correlation near zero, and red indicates a high negative correlation.



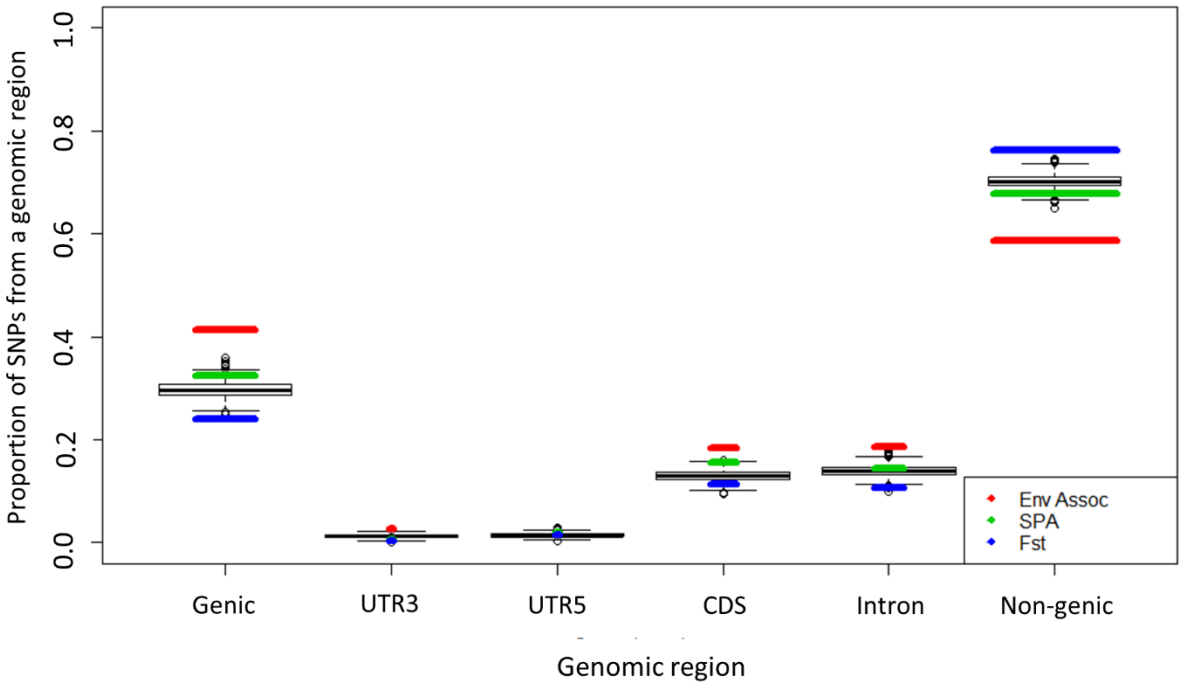
Supplementary Figure 5. Phenotypic variation among subpopulations for a) selected temperature variables, b) selected precipitation variables and c) selected soil variables. Boxplot analysis was used to display phenotypic variation among subpopulations defined by *fast*STRUCTURE.



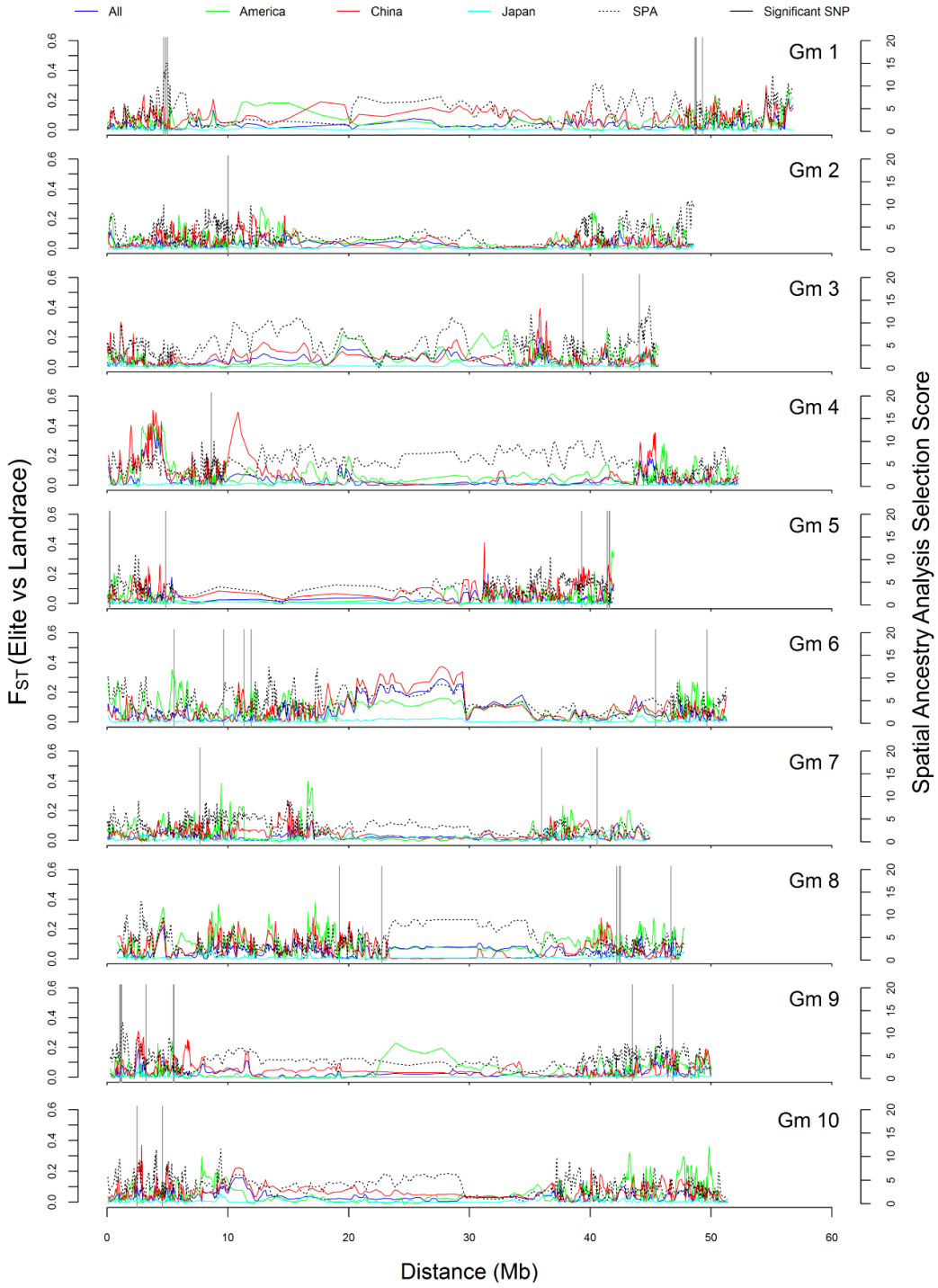
Supplementary Figure 6. Principal component analysis of phenotypic data in 3,012 landrace *G. max* accessions. a, b) The first three PCs were used to infer relationship among variables. c) The first two PCs were used to infer relationship among 3,012 landrace *G. max* accessions. d) The proportion of variance explained by each PC.



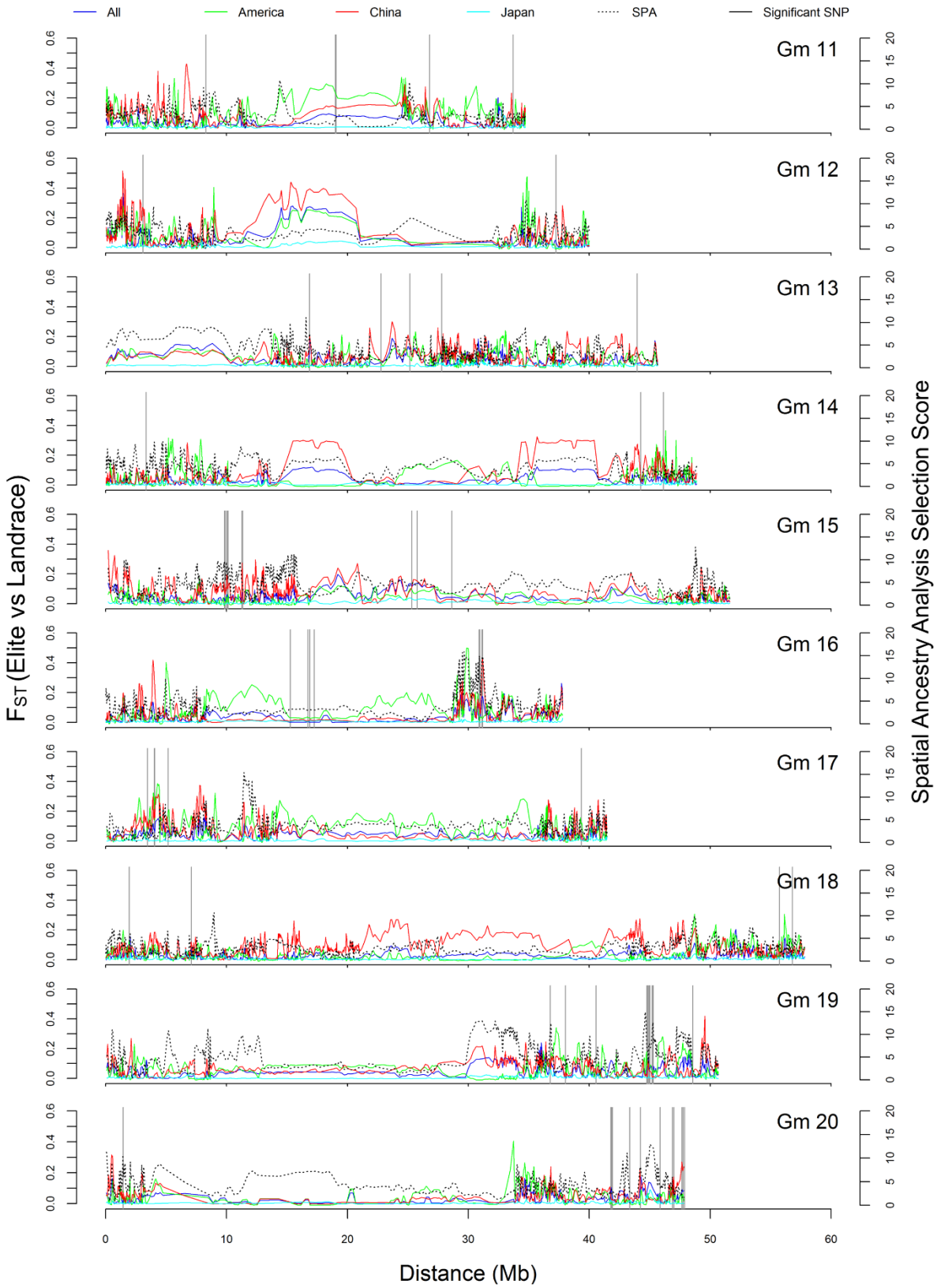
Supplementary Figure 7. Manhattan and quantile-quantile (QQ) plots generated from environmental association analysis using four linear mixed models: a) K model, b) Q+K model c) P+K model and d) L+K model. The QQ plots are displayed to compare the distribution of observed p-values to the expected distribution under the null hypothesis of no association in four different models. The level of significance and the number of associated signals was reduced using the L+K model.



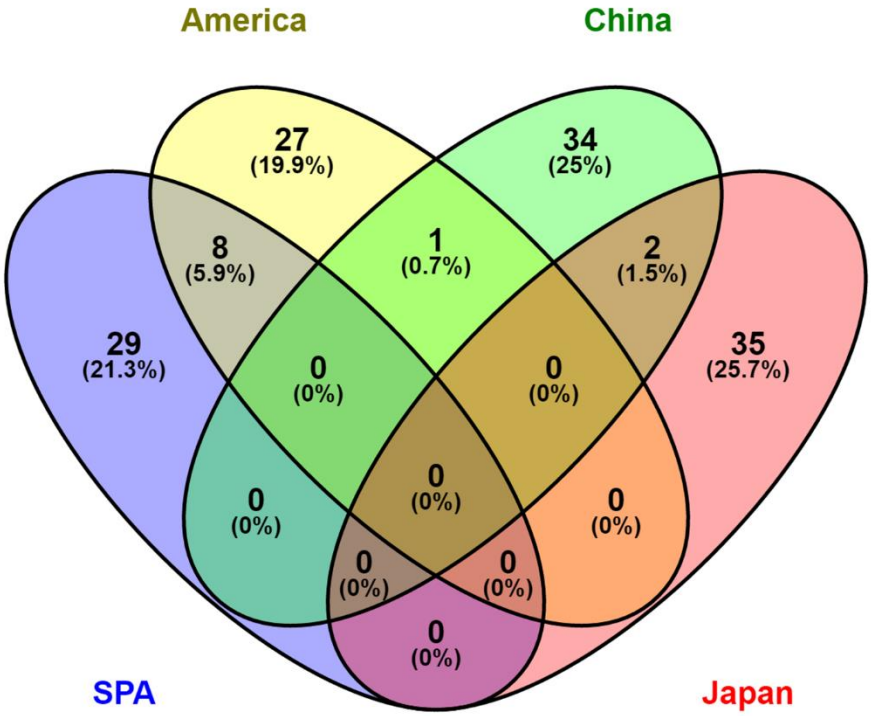
Supplementary Figure 8. Enrichment analysis for genomic region. Enrichment analysis was performed to determine if euchromatin, 3′ UTR, 5′ UTR, coding sequence (CDS), and intronic regions we over or under represented among outliers and significant loci.



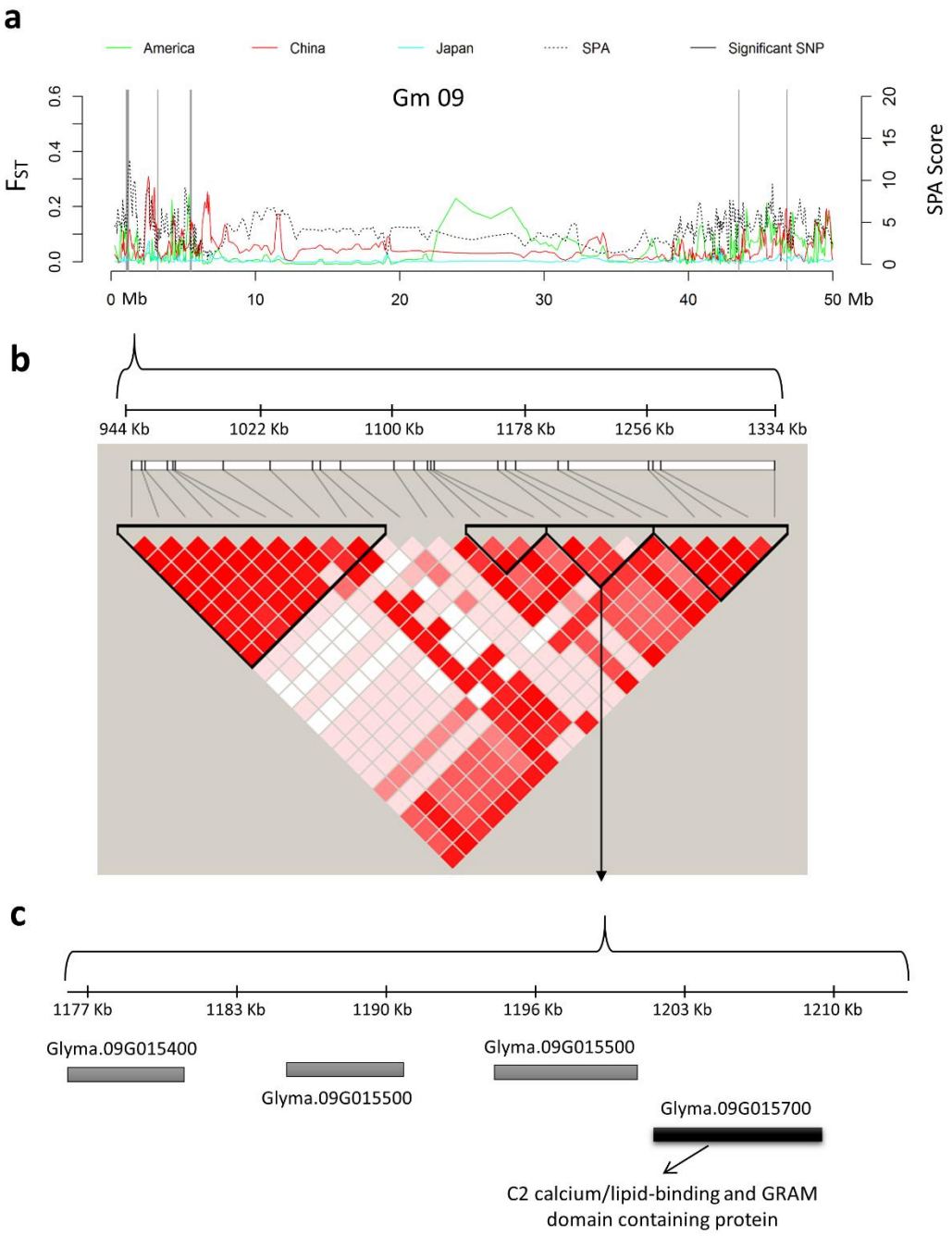
Supplementary Figure 9. Summary selected regions identified by F_{ST} and SPA of and significant associations detected by environmental associations. The gray solid vertical line denotes the position of significant associations. A sliding window approach (see methods) was used to plot F_{ST} and SPA values across the 20 chromosomes. A total of five selected regions overlapped with significant association which is demarked by a green triangle.



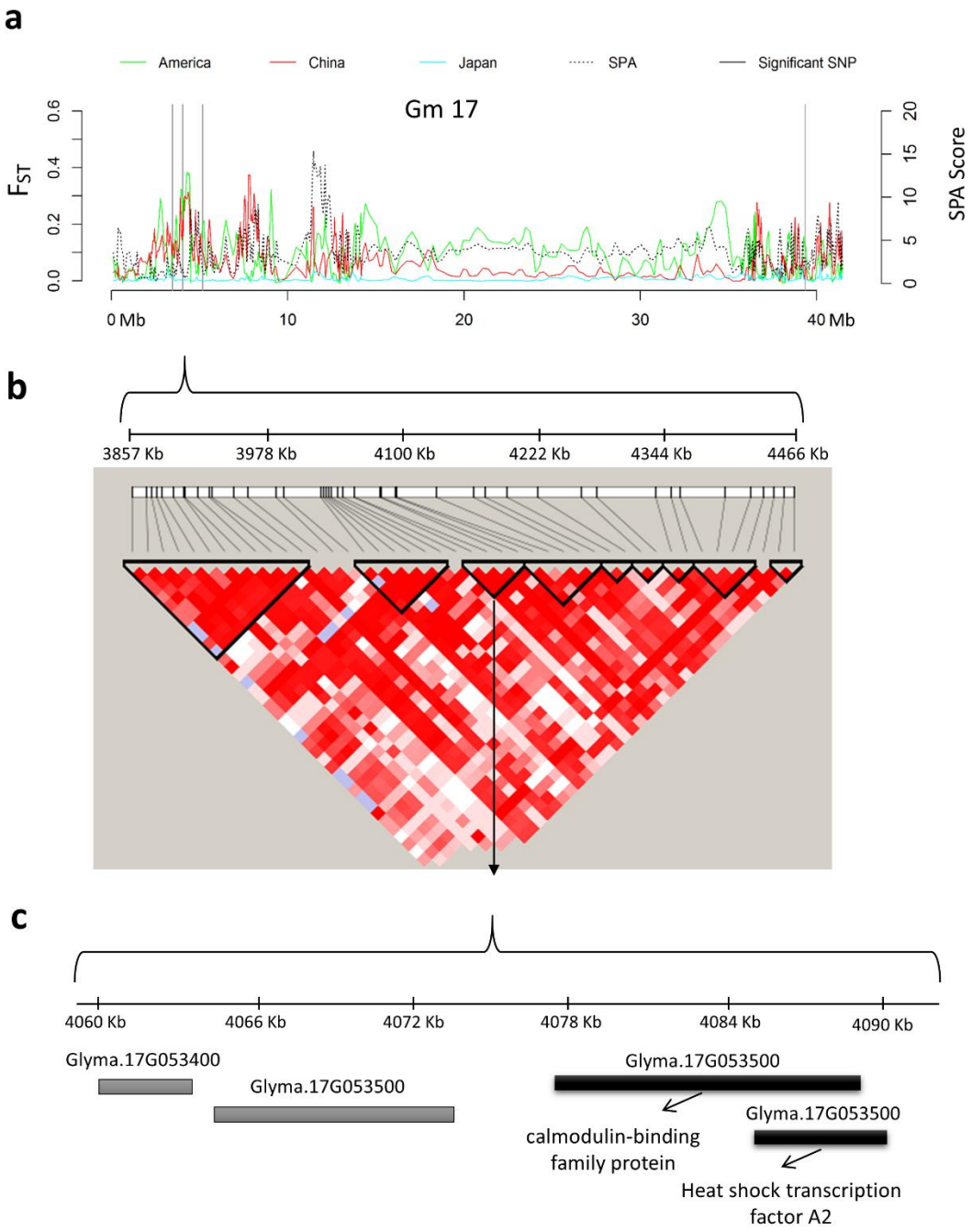
Supplementary Figure 9 (*Continued*).



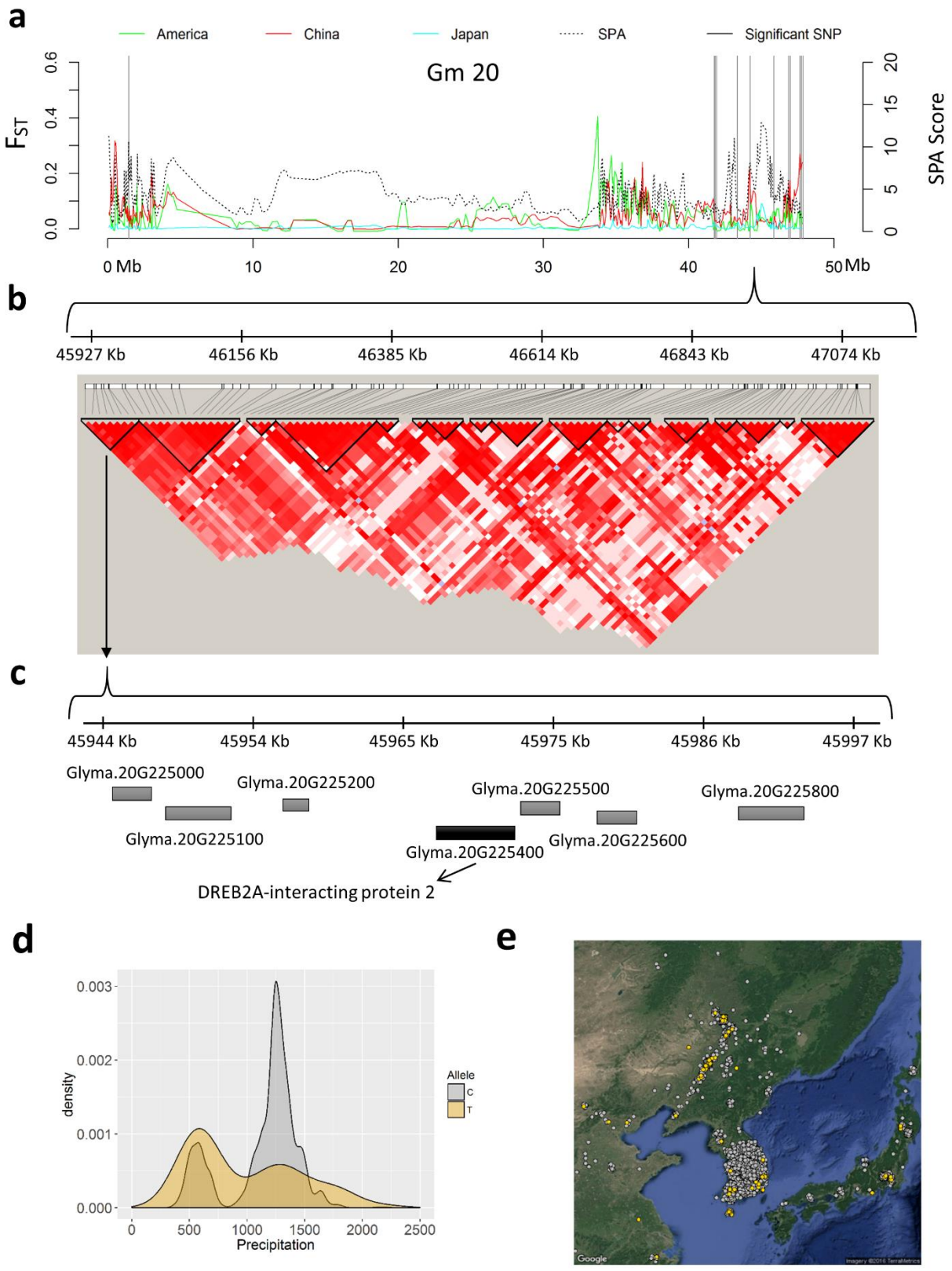
Supplementary Figure 10. Summary of strong selection signals identified using Spatial Ancestry Analysis (SPA) and F_{ST} between elite and landrace population within each country. A Venn diagram was constructed to determine the number of overlaps between and among countries and SPA. The percentage selection signals was put at the bottom of the total detected signals.



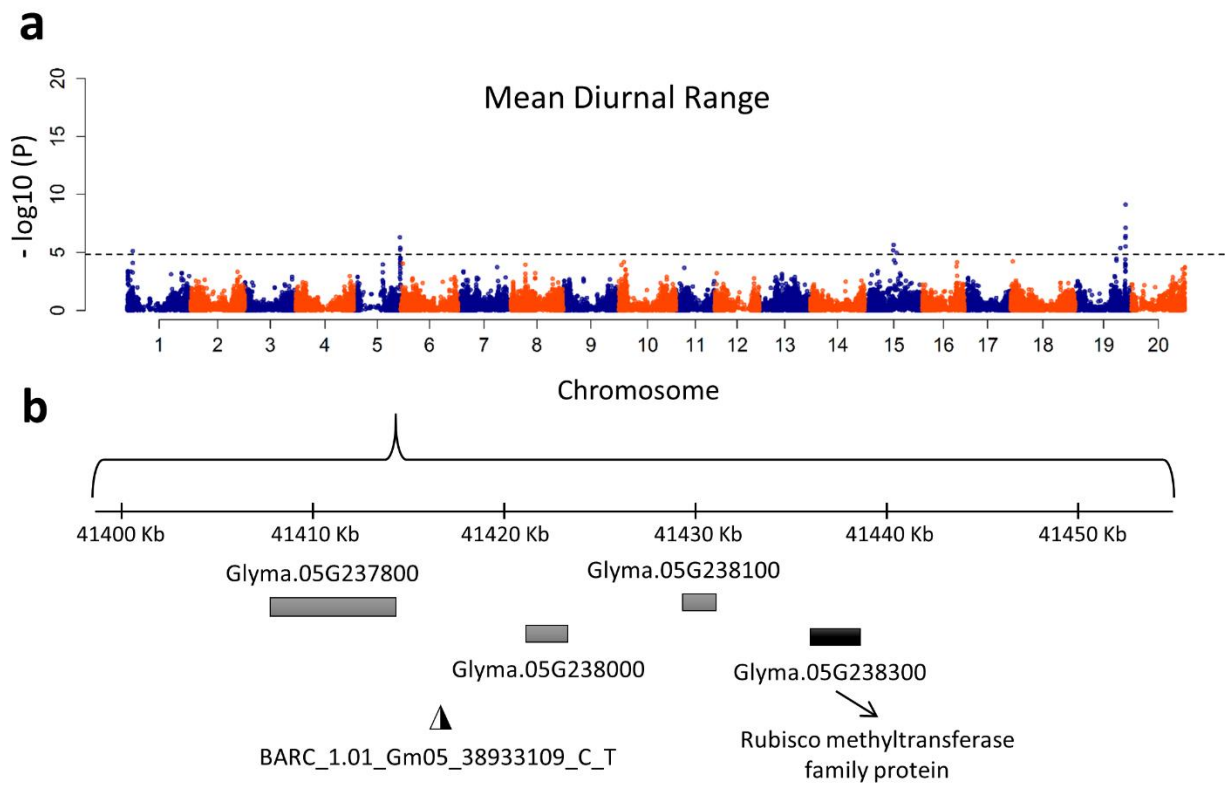
Supplementary Figure 11. Spatial ancestry analysis (SPA), F_{ST}, and significant associations identified on chromosome 9 of the *Glycine max* genome. a) SPA and F_{ST} values were plotted based on a sliding-window approach. The gray solid vertical line denotes the position of significant associations. Notably, highest SPA and F_{ST} values overlapped with significant associations between 1054596 – 1261468 Bp. b) LD and haplotype analysis using the four gamete algorithm of selected region between 1054596 – 1261468 Bp. c) Zoom in on 40 kb region around the significant markers. The Arabidopsis ortholog for the nearest genes, Glyma.09G014700 and Glyma.09G014800, are annotated as Ca^{2+}-dependent lipid-binding (CaLB domain) family protein and oxidoreductase, 2OG-Fe (II) oxygenase family protein, respectively.

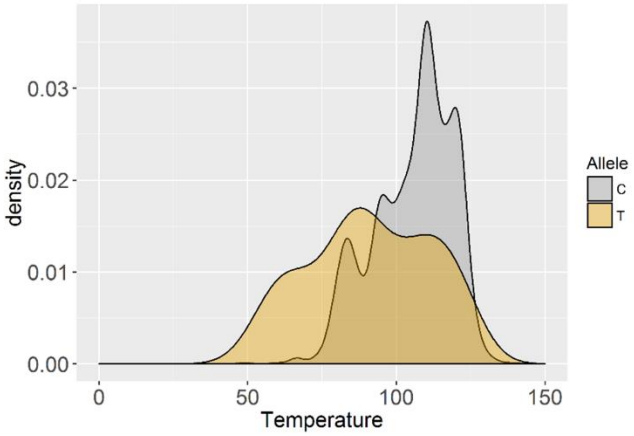


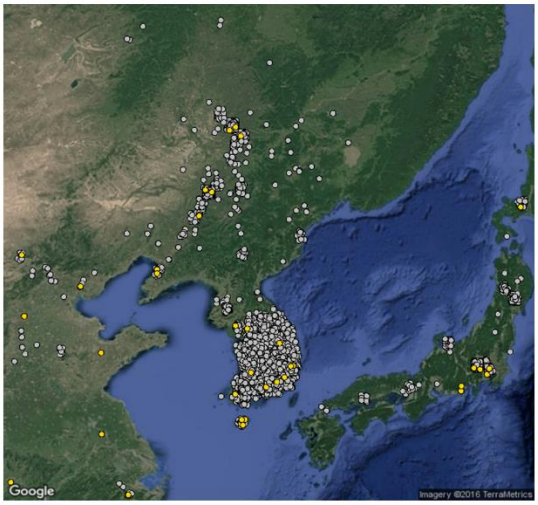
Supplementary Figure 12. Spatial ancestry analysis (SPA), F_{ST}, and significant associations identified on chromosome 17 of the *Glycine max* genome. a) SPA and F_{ST} values were plotted based on a sliding-window approach. The gray solid vertical line denotes the position of significant associations. Notably, highest SPA and F_{ST} values overlapped with significant associations between 3857335 – 4466291 Bp. b) LD and haplotype analysis using the four gamete algorithm of selected region between 3857335 – 4466291 Bp. c) Zoom in on 30 kb region around the significant markers. The Arabidopsis ortholog for the nearest genes are annotated as calmodulin-binding factor and heat-shock transcription factors.


a
America
China
Japan
SPA
Significant SNP
Gm 20
F_{ST}
SPA Score
0 Mb
10
20
30
40
50 Mb
b
45927 Kb
46156 Kb
46385 Kb
46614 Kb
46843 Kb
47074 Kb
c
45944 Kb
45954 Kb
45965 Kb
45975 Kb
45986 Kb
45997 Kb
Glyma.20G225000
Glyma.20G225200
Glyma.20G225500
Glyma.20G225800
Glyma.20G225100
Glyma.20G225400
Glyma.20G225600
DREB2A-interacting protein 2
d
density
Precipitation
Allele
C
T
e
Google

Supplementary Figure 13. Spatial ancestry analysis (SPA), F_{ST}, and significant associations identified on chromosome 20 of the *Glycine max* genome. a) SPA and F_{ST} values were plotted based on a sliding-window approach. The gray solid vertical line denotes the position of significant associations. Notably, highest SPA and F_{ST} values overlapped with significant associations between 45864382 – 47884469 Bp. b) LD and haplotype analysis using the four gamete algorithm within the range of significant SNPs. c) Zoom in on narrowed region around the significant markers. The Arabidopsis ortholog for the nearest gene, Glyma.20G225400, is annotated as Dehydration-Responsive Element Binding Protein2a (DREB2A). d) Density plot of allele frequency distribution for mean diurnal range. The "C" allele at this locus is associated with high temperature environment while the T 'allele' is associated with low temperature environment. e) Geographic location of individuals with the "C" allele (gray) or "T" allele (gold) with jitter added to show overlapping samples.



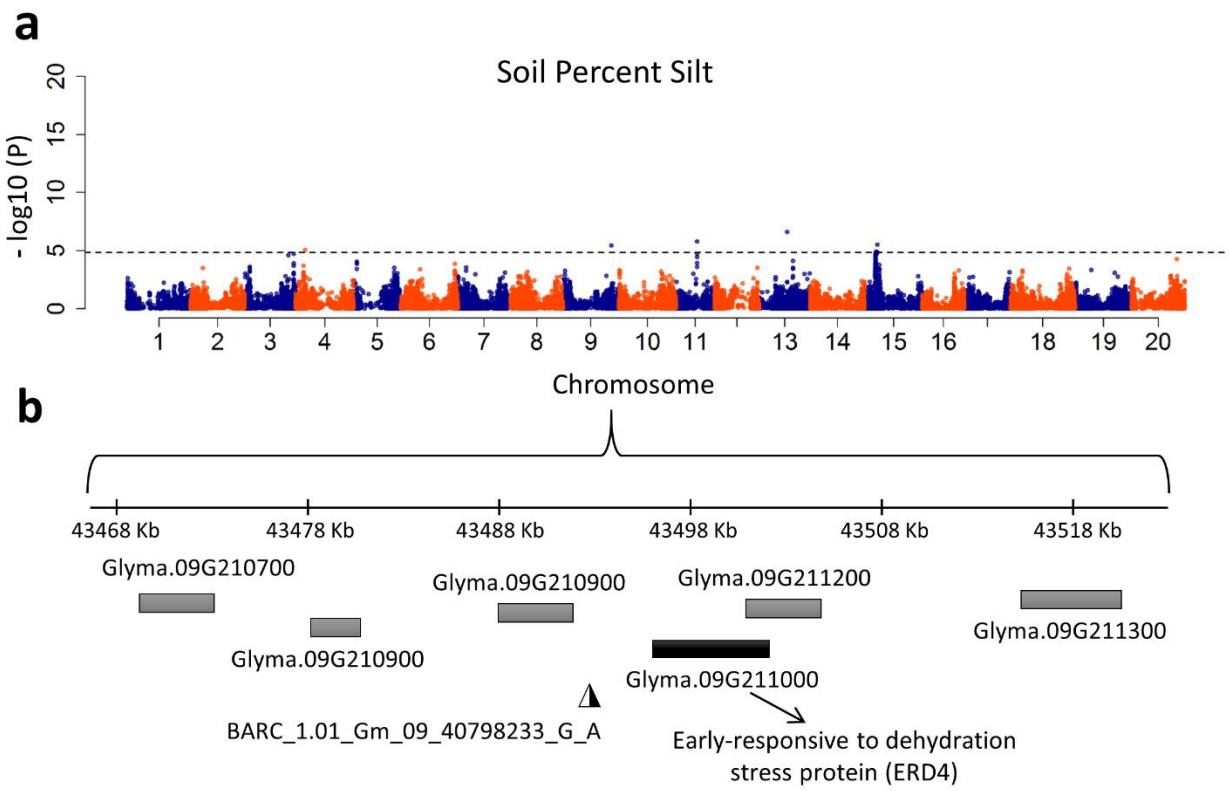


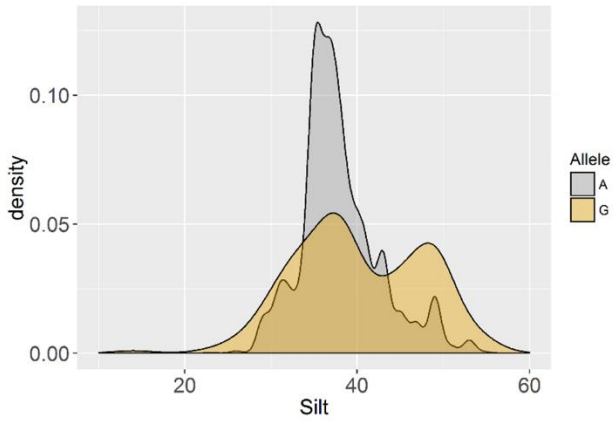


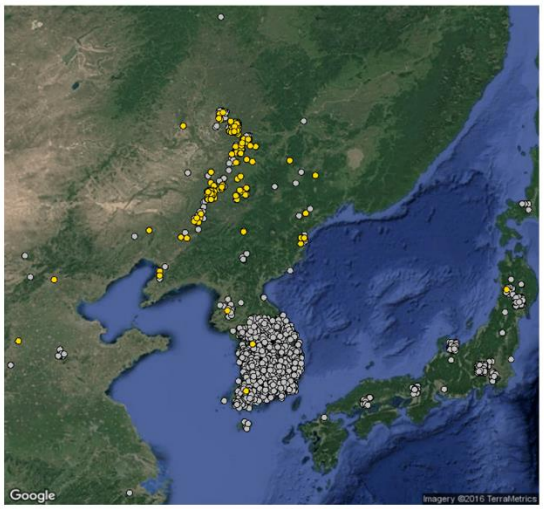
e

Genotypic Class	This study	Song et al (2013)			Elite Population		
	Landrace (N=3,012)	Elite	G.*max*	G. *soja*	America (N=565)	China (N=364)	Japan (N=615)
T Allele	0.017	0.279	0.324	0.431	0.534	0.269	0.167

Supplementary Figure 14. Genome-wide association results for mean diurnal range. a) Genome wide view of association results for mean diurnal range. A cluster of significant associations was identified on chromosomes 1 9, 15 and 20. b) Zoom in on 50 kb region around the most significant marker at 41415712 Bp on chromosome 9. The Arabidopsis ortholog for the nearest gene, Glyma.05G238300 that encodes for Rubisco methyltransferase family protein. c) Density plot of allele frequency distribution for mean diurnal range. The "C" allele at this locus is associated with high temperature environment while the T 'allele' is associated with low temperature environment. d) Geographic location of individuals with the "C" allele (gray) or "T" allele (gold) with jitter added to show overlapping samples. Only individuals with ancestry >80% based on *fastSTRUCTURE* results was plotted. e) Allelic frequency distribution in three subpopulations defined by *fastSRUCTURE* and in Song et al (2013) populations.







e

Genotypic Class	This study	Song et al (2013)			Elite Population		
	Landrace (N=3,012)	Elite	G.*max*	G. *soja*	America (N=565)	China (N=364)	Japan (N=615)
G Allele	0.063	0.372	0.221	0	0.064	0.104	0.074

Supplementary Figure 15. Genome-wide association results for soil percent. a) Genome wide view of association results for soil percent silt. b) Zoom in on 50 kb region around the most significant marker on chromosome 9. The Arabidopsis ortholog for the nearest gene, Glyma.09G211000 that encodes for Early-responsive to dehydration stress protein. c) Density plot of allele frequency distribution for mean diurnal range. The "A" allele at this locus is associated with high low silt environment while the G 'allele' is associated with high silt environment. d) Geographic location of individuals with the "A" allele (gray) or "G" allele (gold) with jitter added to show overlapping samples. Only individuals with ancestry >80% based on *fastSTRUCTURE* results was plotted. e) Allelic frequency distribution in three subpopulations defined by *fastSRUCTURE* and in Song et al (2013) populations.

CHAPTER 5: CONCLUSIONS

The wealth of phenotypic diversity available in the USDA Soybean Germplasm Collection should be mined to help meet the demands of food production in the face of climate change and ever-evolving pathogens. The Collection contains many valuable genes that hold potential to improve the cultivated version of soybean. Finding the genes relevant for breeding and genetics would be benefited by knowing more about the ancestry and diversity of the soybean collection as well as the association between genetic markers and economically important traits. Overall, our findings indicated how samples in the Collection relate to one another and the importance of country of origin and maturity group in determining relatedness. Accessions originating from Japan were relatively homogenous and distinct from the Korean accessions. As a whole, both Japanese and Korean accessions diverged from the Chinese accessions. The ancestry of founders of the American accessions derived mostly from two Chinese subpopulations, which reflects the composition of the American accessions as a whole.

We found several strong associations between genetic markers and phenotypic traits, which help narrowing the search for genes controlling these economically important traits. A 12,000 accession GWAS conducted on seed protein and oil is the largest reported to date in plants and identified SNPs with strong signals on chromosomes 20 and 15. A chromosome 20 region previously reported to be important for protein and oil content was further narrowed and now contains only three plausible candidate genes. The haplotype effects show a strong negative relationship between oil and protein at this locus, indicating negative pleiotropic effects or multiple closely linked loci in repulsion phase linkage. The

vast majority of accessions carry the haplotype allele conferring lower protein and higher oil.

Genome-wide association mapping was also applied to categorical phenotypic data available for ten descriptive traits in a collection of ~13,000 *G. max* accessions. A total of 23 known genes were identified as well as several heretofore unknown genes controlling the phenotypic variants for ten descriptive traits. Because some of those genes had been cloned, we were able to show that the narrow SNP signal regions had chromosomal base pair spans that, with few exceptions, bracketed the base pair region of the cloned gene coding sequences, despite variation in SNP number/distribution of chip SNP set.

By leveraging environmental data, we also elucidate the genetic basis of local adaptation in soybean by exploring the natural variations available in 3,012 locally adapted landrace accessions from across the geographical range of *G. max* species. Our approach of using a combination of EAA and selection mapping identified important candidate genes related to drought and heat stress, and revealed important signatures of directional selection that are likely involved on geographic divergence of soybean.

Overall, the results reported herein will assist soybean researchers in their pursuit of genes that can be used to further improve soybean for agricultural production. Others surely to flow from this valuable resource for providing a fuller understanding of the distribution of genetic variation contained within the collection and its relation to phenotypic variation for economically important traits. Further characterization of the phenotypic diversity and its relationship to the genomic diversity will ultimately facilitate a more efficient and effective introgression of this diversity into elite varieties for continued genetic improvement.