UNRAVELING THE GENETIC ARCHITECTURE OF AGRONOMIC TRAITS AND DEVELOPING A GENOME WIDE INDEL PANEL IN COMMON BEAN (*PHASEOLUS VULGARIS*)

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

By

Samira Mafi Moghaddam

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Program:
Genomics and Bioinformatics

February 2015

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Unraveling the Genetic Architecture of Agronomic Traits and Developing a Genome Wide Indel Panel in Common Bean (*Phaseolus Vulgaris*)

By

Samira Mafi Moghaddam

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

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Phillip McClean
Chair

Dr. Richard Horsley

Dr. Juan Osorno

Dr. Justin Faris

Dr. Yarong Yang

Approved:

02/06/2015	Dr. Phillip McClean
Date	Department Chair

ABSTRACT

Common bean (*Phaseolus vulgaris*) is an economically important legume. The agronomic characteristics of this crop such as days to flower, growth habit and seed yield affect breeding strategies. However, little is known about the genomic regions controlling these traits. Therefore, discovering the genetic architecture underlying important agronomic traits can accelerate breeding via marker assisted selection (MAS) in addition to providing genomic and biological information. Genome wide association studies (GWAS) are currently the method of choice to find the genomic regions associated with traits of interest using a population of unrelated individuals. It takes advantage of the historic recombinations that exist in the population to map the traits at a higher resolution. The availability of a reference genome in common bean has paved the way for higher throughput and more accurate genomic research including the discovery of new knowledge and development of new tools. In the first experiment we conducted GWAS using a panel of 280 diverse genotypes from the Middle American gene pool and about 15,000 SNPS with minor allele frequency of 5% and greater to map seven important agronomic traits in common bean. We were able to detect known and new genomic regions with strong candidate genes associated with these traits. In the second experiment we used sequence data from 19 genotypes from different bean market classes to develop a panel of insertion-deletion (InDel) markers that can be used for MAS as well as other genetic and genomic studies. These user-friendly, cost-effective, and co-dominant markers were tested for their efficiency and application. They demonstrated utility in a medium throughput genetic map construction and diversity analysis.

ACKNOWLEDGEMENTS

No words can really completely express my gratitude to my advisor, Dr. Philip McClean. He has been my teacher and supporter in both research and life. He has been and will always be my role model. I will never forget his guidance, patience, and the trust he placed on me. I would like to thank my supportive colleagues and friends: Dr. Sujan Mamidi who taught me the basis of my analyses and helped me to stay on my own feet in my PhD project; Rian Lee for his help and support in all aspects of my five year experience in the lab and Fargo; and Matthew Doucette and Shireen Chikara for their help with the lab experiments.

It was a privilege to be part of the bean coordinated agricultural project (BeanCAP) funded by USDA-NIFA and work with great professors and students in the bean community on this project like, Dr. Juan Osorno, Dr. Mark Brick, Dr. James Kelly, Dr. Carlos Urrea, and Dr. Perry Cregan, and Dr. Qijian Song. Thank you to the dry bean breeding program at NDSU, especially my friend and classmate Dr. Angela M. Linares-Ramírez.

I would like to thank my committee members: Dr. Richard Horsley, Dr. Justin Faris, Dr. Yarong Yang and Dr. Juan Osorno for their valuable time and guidance. I also appreciate all the support I received from Dr. Shahryar Kianian.

I would like to give special thanks to my husband, Ali Soltani, for his endless support and love. Without him I wouldn't be able to reach where I stand now. I appreciate the long distance support of my family and friends that warmed up my heart in the coldest days of Fargo.

Many thanks to my friends, classmates and officemates at NDSU who made these five years more joyful for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

GENERAL INTRODUCTION

Knowing the genetic architecture of important agronomic traits has always been of great importance to researchers because of its biological and applied applications. Moreover, as more user-friendly, and cost-effective tools become available for genomic and biological research, the outcome is more applied knowledge. Although valuable research has been conducted on economically important agronomic traits in common bean, there is little known about their genetic and genomic control. The main objective of this dissertation was to provide new genetic discoveries and tools for current and future research in common bean.

The first goal was to use genome wide association study (GWAS) to discover loci affecting important agronomic traits. GWAS is the method of choice at this point of time since it is capable of assaying more recombination events, using a diverse panel of bean genotypes from all over the United States, than is possible with a bi-parental population. Using this method as a first step we identified candidate regions of the genome that control the traits of interest. Markers that are associated with candidate regions will be the practical outcomes of this step. This research informs efforts to discover the underlying gene(s) and causal polymorphisms controlling these traits.

Genetic diversity is the main requirement of any breeding program. One of the main concerns in bean breeding programs is that within market class (pinto, navy, black, great northern, and kidney) crosses have limited sequence variability within each market class. Providing market class specific markers will provide geneticists and breeders a tool to better understand the genetic diversity, co-ancestry and relatedness among varieties of beans within and between market classes. Cost-effective and user-friendly market class specific markers will result in a higher utilization of genetic diversity, while also providing a better platform to employ

genomic knowledge in breeding programs through marker assisted selection (MAS). Thus, the second objective of this dissertation was to develop medium throughput market class specific markers that can be used in any laboratory. We developed insertion-deletion (InDel) -based markers as a user-friendly tool. InDels are one additional source of variation along with single nucleotide polymorphisms (SNP) that can be mined as a research tool. These co-dominant markers are distributed throughout the genome and offer breeders and geneticists the flexibility to choose from markers for any region of interest.

CHAPTER 1. LITERATURE REVIEW

Importance of Common Bean

Legumes are among the oldest cultivated plants. They provide high amounts of nutrients per calorie which makes them nutrient-rich crops. Among the legumes, common bean is rich in protein, carbohydrates, vitamins and minerals but low in fat, sodium and contains no cholesterol. Dry bean seed contains lignin which might have an impact in preventing osteoporosis, heart disease, and certain cancers. Indeed, high consumption of dry bean has reduced the risk of advanced colorectal adenoma recurrence among participants in a polyp prevention trial (Lanza et al., 2006). The glycemic index of dry bean is low and varies from 27-42% compared to glucose and 40-59% compared to white bread, respectively. The carbohydrate properties and fiber content of dry bean make it an appropriate food choice to manage abnormalities associated with insulin resistance, diabetes and hyperlipidemia. Dry bean contains both soluble and insoluble fiber. The soluble fiber helps to reduce LDL cholesterol in the blood. Common bean is noted for its high protein content among vegetables and contains up to 25% protein by weight. According to the FAO (FAOSTAT, 2013), common bean provides up to 15% of the total daily calories and more than 30% of the total daily protein requirement in some countries.

Dry edible beans is an economically important legume that was planted on 0.53 million hectares (USDA, 2014) in the United States in 2013 with total production of 47.4 million quintal (USDA-AMS, 2013). The top dry bean producing states in the U.S. are as follows: 1.North Dakota (38%), 2.Michigan (14%), 3.Nebraska (11%), 4.Minnesota (10%), 5.Idaho (7%), 6.California (4%), 7.Whasington (4%), 8.Colorado (3%) (USDA-ERS, 2012). The United States not only produces dry beans for its own consumption but also exports them to other countries

with an average annual value of $361 million from 2008/09 to 2012/13 (Zahniser and Farah Wells, 2014).

Domestication, Diversity, and Commercialization of Common Bean

During the evolution of common bean, some morphological, physiological, and genetic alterations have occurred such as changes in seed, pod, stem, nodes, branches, leaves, growth habit (from indeterminate to determinate), suppression of seed dispersion mechanism, loss of seed dormancy, and other physiological changes such as loss of photoperiod sensitivity (Smartt, 1990; Gepts and Debouck, 1991). Most of the evolutionary changes in beans are due to mutations in a few genes. The study of Koinange et al., (1996), indicated that the genetic control of the domestication syndrome in common bean involves genes with large effect that are responsible for the important phenotypic differences.

Observations on morphological, agronomic, biochemical, and molecular variability for wild and cultivated lines, defined two major centers of domestication. One center of domestication is located in Middle America (northern Mexico) and the other is in the southern Andes (southern Peru, Bolivia, or Argentina) (Mamidi et al., 2011b; Bitocchi et al., 2012; Schmutz et al., 2014). These two major gene pools can be further divided into six races. The Andean gene pool includes races Chile, Nueva Granada, and Peru while the Middle American gene pool consists of races Durango, Jalisco and Mesoamerica (Singh et al., 1991). Mamidi et al., (2011b) proposed a single domestication event in each gene pool with reciprocal migration between wild and landrace genotypes using an approximate Bayesian computation (ABC) approach. It is suggested that approximately 111,000 years ago, two wild type gene pools diverged from one ancestral pool and then each gene pool underwent a bottleneck event which continued for up to 40,000 years. The Middle American gene pool is more diverse than the Andean pool, and gene flow between the two

gene pools is asymmetrical with more migrants from Mesoamerican into Andean gene pool. Therefore, the evidence indicates the presence of a complex population structure in common bean which needs to be considered in genetic and genomic studies.

The United States produces nine commercial market classes of dry beans categorized based on differences in seed size, shape, and color. The major market classes from the Middle American gene pool are pinto, great northern, and pink from race Durango; and black and navy from race Mesoamerica. The main market classes from the Andean gene pool are white, light and dark red kidney, and cranberry, all from race Nueva Granada (Mensack et al., 2010).

Molecular Marker Studies in Common Bean

The molecular basis of many biological phenomena can be studied using genetic variation analysis. Molecular markers are tools to determine the degree of genetic variation (Agarwal, Shrivastava, & Padh, 2008). They are used in QTL analysis, marker assisted selection (MAS), GWAS, and high resolution mapping for gene cloning purposes. An ideal molecular marker should have the following characteristics according to Agarwal et al., (2008) : "1- be polymorphic and evenly distributed throughout the genome, 2- provide adequate resolution of genetic differences, 3-generate multiple, independent and reliable markers, 4- simple, quick and inexpensive, 5-need small amounts of tissue and DNA samples, 6- have linkage to distinct phenotypes". However, there is no ideal molecular marker for all situations.

Since the advent of molecular genetics, researchers in common bean have developed and utilized a wide variety of molecular markers such as RFLP (Khairallah et al., 1990, 1992; Velasquez and Gepts, 1994; De Meaux et al., 2002), RAPD (Haley et al., 1994; Skroch and Nienhuis, 1995; Freyre et al., 1996; Beebe et al., 2000; Franco et al., 2001; Galván et al., 2001, 2006; Maciel et al., 2001), AFLP (Tohme et al., 1996; Papa and Gepts, 2003; Rosales-Serna et

al., 2005), inter simple sequence repeat (Galván et al., 2003; González et al., 2005), SSR (Blair et al., 2003, 2006; Buso, Amaral, Brondani, & Ferreira, 2006; Córdoba, Chavarro, Schlueter, Jackson, & Blair, 2010; Gaitán-Solís, Duque, Edwards, & Tohme, 2002; Galeano, Fernández, Gómez, & Blair, 2009; Gómez, Blair, Frankow-Lindberg, & Gullberg, 2004; Jianchun, Xinwen, S, & E, 2000; Masi, Zeuli, & Donini, 2003; Métais, Hamon, Jalouzot, & Peltier, 2002; K. Yu, Park, & Poysa, 1999), gene-based markers (Kami, Velasquez, Debouck, & Gepts, 1995; Mamidi et al., 2013; Mamidi et al., 2011; McClean, Lee, & Miklas, 2004; McConnell et al., 2010), and insertion/deletions (InDels) (Moghaddam et al., 2014).

Among the different types of markers, InDels and SNPs can be applied at medium and high throughput levels, respectively. Studies has shown that SNPs and InDels are the most abundant variations distributed throughout the genome of many species (Garg, 1999; Drenkard, 2000; Nasu, 2002; Batley et al., 2003). Although InDels are less abundant compared to SNPs, they can be genotyped using easy and simple procedures based on fragment size differences. Moreover, they are co-dominant, PCR-based, and there is little chance of mutation in the InDels of the exact same size and genomics position. This means the shared InDels represent identity-by-descent (Väli et al., 2008). Such characteristics make InDels a suitable choice for medium-throughput application in the laboratory. On the other hand, high throughput automated assays can be employed for SNP detection since their detection is independent of methods that measure DNA fragment sizes (Gaitán-Solís et al., 2008). This results in a reduction of genotyping costs which in turn increases the availability of high-throughput methods to more researchers (Hyten et al., 2010). In addition, SNPs in coding sequences might indicate a functional change in the encoded amino acid which results in an alternative phenotype. The benefits mentioned above make SNPs an appropriate molecular marker system for phenotype-genotype association studies.

Different genetic linkage maps have been developed in common bean using molecular markers for important traits such as disease resistance (Haley et al., 1994; Adam-Blondon et al., 1994; Johnson et al., 1995; Jung et al., 1996; Young and Kelly, 1997; Bai et al., 1997; Park et al., 1999; Ariyarathne et al., 1999; Miklas et al., 2001; Schneider et al., 2001), morphological traits (Jung et al., 1996), seed size (Park et al., 2000), canning quality (Walters et al., 1997), drought stress tolerance, (Schneider et al., 1997), and traits affected by domestication processes (Koinange et al., 1996). The core linkage map was first developed by Freyre et al., (1998) through merging shared RFLP markers among different maps. It is 1226 cM in length and includes 563 markers, e Mbracing 120 RFLP and 430 RAPD markers, and a few isozyme and phenotypic marker loci. The population used to construct the core linkage map was derived from a cross between BAT93 and Jalo EEP558. The reason for selecting these parents is their divergent evolutionary origins and often contrasting disease responses. Many studies later increased the density of this and other maps (Blair et al., 2003; Grisi et al., 2007; Hanai et al., 2010; McClean, 2002; McConnell et al., 2010). Schlueter et al., (2008), published a draft physical map of common bean by asse Mbling BAC-end sequences. The genomic sequences generated by this method covered 9.54% of the genome. Analysis of these data and 1,404 shotgun sequences derived from cultivar BAT7 led to the conclusion that common bean genome contains 49.2% repetitive sequence and 29.3% genic sequences. The amount of repetitive DNA is higher in *Phaseolus* compared to other legumes. With the advent of high-throughput genotyping and sequencing in plants, Hyten et al., (2010) developed the first high-throughput SNP chip including 827 working SNPs. Recently two 6K Illumina iSelect SNP arrays became available to common bean (Cregan- unpublished data). Schmutz et al., (2014) sequenced the G19833 line with a genome assembly of 521.1 Mb. The assembly was based on a genetic linkage

map constructed using ~ 7000 SNPs in a Stampede × Redhawk population (Schmutz et al., 2014). Thus, the latest QTL and GWAS studies in common bean can employ the high throughput genotyping systems (Linares-Ramirez, 2013; Agarwal, 2014).

Genome Wide Association Studies

Molecular markers are essential for linkage mapping and genetic mapping of phenotypic traits. Historically these studies used bi-parental mapping populations. However, there are some limitations to bi-parental populations such as impossibility of appropriate designed cross (not possible for all species like trees), small population size, and sampling only two alleles at a locus (Gupta et al., 2005). GWAS or association mapping has the advantage of capturing many recombination events from unrelated individuals. GWAS is based on linkage disequilibrium (LD) which is defined as "non-random association of alleles at different loci" (Falconer & Mackay, 1996). This approach can identify causal polymorphisms (Palaisa, 2003; Palaisa et al., 2004) and/or haplotype blocks that aid QTL mapping. To conduct a GWAS, first the genotypes that make up the population of the study should be selected from a natural population or germplasm collection in a way that includes a wide phenotypic range. Secondly, the population should be genotyped using the preferred molecular marker system and phenotyped in different environments with replicates. The genotypic information from molecular markers will be used to estimate LD decay, population structure, and coefficient of relatedness (kinship). The final step involves statistical analyses to identify the association between a phenotype and a marker locus (Balding, 2006). The first LD-based association study of a QTL using candidate genes in plants was the study of flowering time and *dwarf*8 gene in maize (Thornsberry et al., 2001). GWAS has been performed in Arabidopsis (Aranzana et al., 2005; Atwell et al., 2010), and other plant species such as maize (Thornsberry et al., 2001; Buckler et al., 2009; Tian et al., 2011; Robbins et al., 2011),

barley (Cockram et al., 2010; Wang et al., 2012b; a), rice (Muers, 2011; Li et al., 2012), soybean (Mamidi et al., 2011a, 2014; Hwang et al., 2014) and recently common bean (Agarwal, 2014; Shi, Navabi, & Yu, 2011).

Statistical Methods in Genome Wide Association Studies

The simplest association test is a Pearson 2-df test or a Fisher exact test which test the null hypothesis of no association between the marker and phenotype. This approach has been used in case-control phenotypes but is not very powerful for complex traits when the contribution of the marker to the phenotype is additive. For continuous traits, analysis of variance (ANOVA) is analogous to Pearson 2-df test, and it assumes the same null hypothesis. The other method for continuous traits is linear regression which reduces the degree of freedom from two to one. Both assume a normal residual distribution. Logistic regression is a more appropriate approach when the phenotype is binary and not normally distributed such as case-control studies. In human GWAS, where the disease outcome is categorical, multinomial regression can be applied (Balding, 2006).

Association Analysis and Dealing with Population Structure

Population structure and kinship are important issues in association mapping. Divergence from the ideal state of a panmictic population is called population structure which can result in spurious association in GWAS by causing LD between unlinked loci. Thus, it is crucial to have knowledge about the structure in the species and the population used for a GWAS. *P. vulgaris* is a highly structured species as explained by its domestication history. Many association analysis statistical methods have been developed to control for population structure and kinship because the greatest statistical power is achieved when minimal population structure exists, and the trait is normally distributed (Yu et al., 2006). Two statistical methods that are widely used to account for

population structure are principal component analysis (PCA) (Price et al., 2006) and a model-based approach using unlinked markers as proposed by Pritchard et al., (2000). The two approaches can then be incorporated into a linear regression model that is used for continuous traits GWAS. Yu et al., (2006) proposed a mixed-model method for association studies which accounts for both fixed (SNP effect and population structure) and random effects (kinship) by integrating population structure and familial relatedness into the regression model. This method is flexible because it is applicable for both family-based and population-based samples.

References

Adam-Blondon, A. F., Sévignac, M., Bannerot, H., & Dron, M. (1994). SCAR, RAPD and RFLP markers linked to a dominant gene (Are) conferring resistance to anthracnose in common bean. *Theor. Appl. Genet.* , *88*(6-7), 865–70.

Agarwal, C. (2014). Association mapping of agronomic traits of dry beans using breeding populations. North Dakota State University.

Agarwal, M., Shrivastava, N., & Padh, H. (2008). Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Reports*, *27*(4), 617–31.

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C. , Molitor,J., Shindo, C., Tang, C., Toomajian,C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P., Nordborg, M. (2005). Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genetics*, *1*(5), e60.

Ariyarathne, H. M., Coyne, D. P., Jung, G., Skroch, P. W., Vidaver, A. K., Steadman, J. R., Miklas, P.N., Bassett, M. J. (1999). Molecular Mapping of Disease Resistance Genes for Halo Blight, Common Bacterial Blight, and Bean Common Mosaic Virus in a Segregating Population of Common Bean. *J. Amer. Soc. Hort. Sci.*, *124*(6), 654–662.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T. Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Caroline Dean, J. C., Debieu, M., de Meaux, J., Ecker, J. R., Faure, N., Kniskern, J. M., Jones, J. D. G., Michael, T., Nemri, A., Roux, F., Salt, D. E., Tang, C., Todesco, M., Traw, M. B., Weigel, D., Marjoram, P., Borevitz, J. O., Bergelson, J., Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, *465*(7298), 627–31.

Bai, Y., Michaels, T. E., & Pauls, K. P. (1997). Identification of RAPD markers linked to common bacterial blight resistance genes in *Phaseolus vulgaris* L. *Genome*, *40*(4), 544–551.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews. Genetics*, *7*(10), 781–91.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K. J., & Edwards, D. (2003). Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.* , *132*(1), 84–91.

Beebe, S., Skroch, P. W., Tohme, J., Duque, M. C., Pedraza, F., & Nienhuis, J. (2000). Structure of genetic diversity among common bean landraces of middle american origin based on correspondence analysis of rapd. *Crop Sci*, *40*(1), 264. doi:10.2135/cropsci2000.401264x

Bitocchi, E., Nanni, L., Bellucci, E., Rossi, M., Giardini, A., Zeuli, P. S., Logozzo G., Stougaard J., McClean P., Attene G., Papa, R. (2012). Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. *Proc. Natl. Acad. Sci. U.S.A.*, *109*(14), E788–96.

Blair, M. W., Giraldo, M. C., Buendía, H. F., Tovar, E., Duque, M. C., & Beebe, S. E. (2006). Microsatellite marker diversity in common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* , *113*(1), 100–9.

Blair, M. W., Pedraza, F., Buendia, H. F., Gaitán-Solís, E., Beebe, S. E., Gepts, P., & Tohme, J. (2003). Development of a genome-wide anchored microsatellite map for common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* , *107*(8), 1362–74.

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia1,A., Glaubitz, J. C., Goodman, M. M., Harjes, C., Guill1,K., Kroon, D. E., Larsson,S., Lepak, N. K., Li,H., Mitchel, S. E., Pressoir, G., Peiffer, J. A., Rosas, M. O., Rocheford, T.R., Romay, M.C., Romero, S., Salvo, S., Villeda, H.S., da Silva, H. S., Sun, Q., Tian, F., Upadyayula, N., Ware1,D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., McMullen, M. D. (2009). The genetic architecture of maize flowering time. *Science (New York, N.Y.)*, *325*(5941), 714–8.

Buso, G. S. C., Amaral, Z. P. S., Brondani, R. P. V., & Ferreira, M. E. (2006). Microsatellite markers for the common bean *Phaseolus vulgaris*. *Mol. Ecol. Notes*, *6*(1), 252–254.

Cockram, J., White, J., Zuluaga, D. L., Smith, D., Comadran, J., Macaulay, M., Luoc, Z., Kearseyc, M. J., Wernerd, P., Harrapd, D., Tapselld, C., Liub, H., Hedleyb, P. E., Steine, N., Schultee, D., Steuernagele, B., Marshallb, D. F., Thomasb, W. T. B., Ramsayb, L., Mackaya, I., Baldingf, D. J., The AGOUEB Consortium, Waugh, R., O'Sullivan, D. M. (2010). Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proc. Natl. Acad. Sci. U.S.A.*, *107*(50), 21611–6.

Córdoba, J. M., Chavarro, C., Schlueter, J. A., Jackson, S. A., & Blair, M. W. (2010). Integration of physical and genetic maps of common bean through BAC-derived microsatellite markers. *BMC Genomics*, *11*(1), 436.

De Meaux, J., Cattan-Toupance, I., Lavigne, C., Langin, T., & Neema, C. (2002). Polymorphism of a complex resistance gene candidate family in wild populations of common bean (*Phaseolus vulgaris*) in Argentina: comparison with phenotypic resistance polymorphism. *Mol. Ecol.*, *12*(1), 263–273.

Drenkard, E. (2000). A Simple Procedure for the Analysis of Single Nucleotide Polymorphisms Facilitates Map-Based Cloning in Arabidopsis. *Plant Physiol.* , *124*(4), 1483–1492.

Falconer, D., & Mackay, T. (1996). Introductionto Quantitative Genetics (p. 464). Essex, UK: Longman Group Ltd.

FAOSTAT. (2013). FAO statistical databases and data sets. *Food and agriculture organization of the United Nations*. Retrieved from http://faostat.fao.org/

Franco, M. C., Cassini, S. T. A., Oliveira, V. R., & Tsai, S. M. (2001). Caracterização da diversidade genética em feijão por meio de marcadores RAPD. *Pesquisa Agropecuária Brasileira*, *36*(2), 381–385.

Freyre, R., Ríos, R., Guzmán, L., Debouck, D. G., & Gepts, P. (1996). Ecogeographic distribution ofPhaseolus spp. (*Fabaceae*) in Bolivia.*Econ. Bot.*, *50*(2), 195–215.

Freyre, R., Skroch, P. W., Geffroy, V., Adam-Blondon, A.-F., Shirmohamadali, A., Johnson, W. C., Llaca, V., Nodari, R. O., Pereira, P. A., Tsai, S.-M., Tohme, J., Dron, M., Nienhuis, J.,Vallejos, C. E., Gepts, P. (1998). Towards an integrated linkage map of common bean. 4. Development of a core linkage map and alignment of RFLP maps. *Theor. Appl. Genet.*, *97*(5-6), 847–856.

Gaitán-Solís, E., Choi, I.-Y., Quigley, C., Cregan, P., & Tohme, J. (2008). Single Nucleotide Polymorphisms in Common Bean: Their Discovery and Genotyping Using a Multiplex Detection System. *The Plant Genome Journal*, *1*(2), 125.

Gaitán-Solís, E., Duque, M. C., Edwards, K. J., & Tohme, J. (2002). Microsatellite repeats in common bean . *Crop Sci*, *42*(6), 2128.

Galeano, C. H., Fernández, A. C., Gómez, M., & Blair, M. W. (2009). Single strand conformation polymorphism based SNP and Indel markers for genetic mapping and synteny analysis of common bean (*Phaseolus vulgaris* L.). *BMC Genomics*, *10*(1), 629.

Galván, M. Z., Aulicino, M. B., Medina, S. G., & Balatti, P. A. (2001). Genetic diversity among Northwestern Argentinian cultivars of common bean (*Phaseolus vulgaris* L.) as revealed by RAPD markers. *Genetic Resources and Crop Evolution*, *48*(3), 251–260.

Galván, M. Z., Bornet, B., Balatti, P. A., & Branchard, M. (2003). Inter simple sequence repeat (ISSR) markers as a tool for the assessment of both genetic diversity and gene pool origin in common bean (*Phaseolus vulgaris* L.). *Euphytica*, *132*(3), 297–301.

Galván, M. Z., Menéndez-Sevillano, M. C., De Ron, A. M., Santalla, M., & Balatti, P. A. (2006). Genetic diversity among wild common beans from northwestern argentina based on morpho-agronomic and RAPD data. *Genetic Resources and Crop Evolution*, *53*(5), 891–900.

Garg, K. (1999). Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using Asse Mbled Expressed Sequence Tags. *Genome Research*, *9*(11), 1087–1092.

Gepts, P., & Debouck, D. G. (1991). Origin, domestication, and evolution of the common bean (*Phaseolus vulgaris* L.). In A. van Schoonhoven & O. Voysest (Eds.), *Common beans: Research for crop improvement* (pp. 7–53). Wallingford, UK and CIAT, Cali, Colombia: CAB Intl.

Gómez, O. J., Blair, M. W., Frankow-Lindberg, B. E., & Gullberg, U. (2004). Molecular and phenotypic diversity of common bean landraces from Nicaragua. *Crop Sci*, *44*(4), 1412.

González, A., Wong, A., Delgado-Salinas, A., Papa, R., & Gepts, P. (2005). Assessment of inter simple sequence repeat markers to differentiate sympatric wild and domesticated populations of common bean. *Crop Sci*, *45*(2), 606.

Grisi, M. C. M., Blair, M. W., Gepts, P., Brondani, C., Pereira, P. A. A., & Brondani, R. P. V. (2007). Genetic mapping of a new set of microsatellite markers in a reference common bean (*Phaseolus vulgaris*) population BAT93 x Jalo EEP558. *Genetics and Molecular Research*, *6*(3), 691–706.

Gupta, P. K., Rustgi, S., & Kulwal, P. L. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Molecular Biology*. Springer-Verlag.

Haley, S. D., Miklas, P. N., Afanador, L., & Kelly, J. D. (1994). Random Amplified Polymorphic DNA (RAPD) marker variability between and within gene pools of common bean. *J. Amer. Soc. Hort. Sci.*, *119*(1), 122–125.

Hanai, L. R., Santini, L., Camargo, L. E. A., Fungaro, M. H. P., Gepts, P., Tsai, S. M., & Vieira, M. L. C. (2010). Extension of the core map of common bean with EST-SSR, RGA, AFLP, and putative functional markers. *Mol. Breed. : New Strategies in Plant Improvement*, *25*(1), 25–45.

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

Hyten, D. L., Song, Q., Fickus, E. W., Quigley, C. V, Lim, J.-S., Choi, I.-Y., Hwang, E-Y., Pastor-Corrales, M., Cregan, P. B. (2010). High-throughput SNP discovery and assay development in common bean. *BMC Genomics*, *11*(1), 475.

Jianchun, G., Xinwen, H., S, Y., & E, Y. (2000). Isolation and characterization of microsatellites in snap bean. *Acta Botanica Sinica*, *42*(11), 1179–1183.

Johnson, E., Miklas, P. N., Stavely, J. R., & Martinez-Cruzado, J. C. (1995). Coupling- and repulsion-phase RAPDs for marker-assisted selection of PI 181996 rust resistance in common bean. *Theor. Appl. Genet.* , *90*(5), 659–64.

Jung, G., Coyne, D. P., Skroch, P. W., Nienhuis, J., Arnaud-Santana, E., Bokosi, J., Ariyarathne, H.M., Steadman, S. R., Beaver, J. S., Kaeppler, S. M. (1996). Molecular markers associated with plant architecture and resistance to common blight, web blight, and rust in common beans. *J. Amer. Soc. Hort. Sci.*, *121*(5), 794–803.

Kami, J., Velasquez, V. B., Debouck, D. G., & Gepts, P. (1995). Identification of presumed ancestral DNA sequences of phaseolin in Phaseolus vulgaris. *Proc. Natl. Acad. Sci. U.S.A.*, *92*(4), 1101–1104.

Khairallah, M. M., Adams, M. W., & Sears, B. B. (1990). Mitochondrial DNA polymorphisms of Malawian bean lines: further evidence for two major gene pools. *Theor. Appl. Genet.* , *80*(6), 753–61.

Khairallah, M. M., Sears, B. B., & Adams, M. W. (1992). Mitochondrial restriction fragment length polymorphisms in wild Phaseolus vulgaris L.: insights on the domestication of the common bean. *Theor. Appl. Genet.* , *84*(7-8), 915–22.

Koinange, E. M. K., Singh, S. P., & Gepts, P. (1996). Genetic Control of the Domestication Syndrome in Common Bean. *Crop Sci*, *36*(4), 1037.

Lanza, E., Hartman, T. J., Albert, P. S., Shields, R., Slattery, M., Caan, B., Paskett E., Iber F., Kikendall J. W., Lance P., Daston C., Schatzkin, A. (2006). High dry bean intake and reduced risk of advanced colorectal adenoma recurrence among participants in the polyp prevention trial. *J. Nutr.*, *136*(7), 1896–1903.

Li, X., Yan, W., Agrama, H., Jia, L., Jackson, A., Moldenhauer, K., Yeater, K., McClung, A., Wu, D. (2012). Unraveling the complex trait of harvest index with association mapping in rice (*Oryza sativa* L.). *PloS One*, *7*(1), e29350.

Linares-Ramirez, A. (2013). Selection of dry bean genotypes adapted for drought tolerance in the northern Great Plains. North Dakota State University.

Maciel, F. L., Gerald, L. T. S., & Echeverrigaray, S. (2001). Random amplified polymorphic DNA (RAPD) markers variability among cultivars and landraces of common beans (*Phaseolus vulgaris* L.) of south-Brazil. *Euphytica*, *120*(2), 257–263.

Mamidi, S., Chikara, S., Goos, R. J., Hyten, D. L., Annam, D., Moghaddam, S. M., Lee, R. K., Cregan, P. B., McClean, P. E. (2011a). Genome-wide association analysis identifies candidate genes associated with iron deficiency chlorosis in soybean. *The Plant Genome Journal*, *4*(3), 154.

Mamidi, S., Lee, R. K., Goos, J. R., & McClean, P. E. (2014). Genome-wide association studies identifies seven major regions responsible for iron deficiency chlorosis in soybean (*Glycine max*). *PloS One*, *9*(9), e107469.

Mamidi, S., Rossi, M., Annam, D., Moghaddam, S., Lee, R., Papa, R., & McClean, P. (2011b). Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. *Funct. Plant Biol.* , *38*(12), 953.

Mamidi, S., Rossi, M., Moghaddam, S. M., Annam, D., Lee, R., Papa, R., & McClean, P. E. (2013). Demographic factors shaped diversity in the two gene pools of wild common bean *Phaseolus vulgaris* L. *Heredity*, *110*(3), 267–76.

Masi, P., Zeuli, P. L. S., & Donini, P. (2003). Development and analysis of multiplex microsatellite markers sets in common bean (Phaseolus vulgaris L.). *Mol. Breed.*, *11*(4), 303–313.

McClean, P. E. (2002). Molecular and Phenotypic Mapping of Genes Controlling Seed Coat Pattern and Color in Common Bean (*Phaseolus vulgaris* L.). *Heredity*, *93*(2), 148–152.

McClean, P. E., Lee, R. K., & Miklas, P. N. (2004). Sequence diversity analysis of dihydroflavonol 4-reductase intron 1 in common bean. *Genome / National Research Council Canada = Génome / Conseil National de Recherches Canada*, *47*(2), 266–80.

McConnell, M., Mamidi, S., Lee, R., Chikara, S., Rossi, M., Papa, R., & McClean, P. (2010). Syntenic relationships among legumes revealed using a gene-based genetic linkage map of common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* , *121*(6), 1103–16.

Mensack, M. M., Fitzgerald, V. K., Ryan, E. P., Lewis, M. R., Thompson, H. J., & Brick, M. A. (2010). Evaluation of diversity among common beans (*Phaseolus vulgaris* L.) from two centers of domestication using "omics" technologies. *BMC Genomics*, *11*(1), 686.

Métais, I., Hamon, B., Jalouzot, R., & Peltier, D. (2002). Structure and level of genetic diversity in various bean types evidenced with microsatellite markers isolated from a genomic enriched library. *Theor. Appl. Genet.* , *104*(8), 1346–1352.

Miklas, P. N., Johnson, W. C., Delorme, R., & Gepts, P. (2001). QTL Conditioning Physiological Resistance and Avoidance to White Mold in Dry Bean. *Crop Sci*, *41*(2), 309.

Moghaddam, S. M., Song, Q., Mamidi, S., Schmutz, J., Lee, R., Cregan, P., Osorno, J. M., McClean, P. E. (2014). Developing market class specific InDel markers from next generation sequence data in *Phaseolus vulgaris* L. *Frontiers in Plant Science*, *5*, 185.

Muers, M. (2011). Complex traits: Genome-wide association mapping in rice. *Nature Reviews Genetics*, *12*(11), 741–741.

Nasu, S. (2002). Search for and analysis of Single Nucleotide Polymorphisms (SNPs) in rice (*Oryza sativa*, *Oryza rufipogon*) and establishment of SNP markers. *DNA Research*, *9*(5), 163–171.

Palaisa, K. A. (2003). Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *The Plant Cell Online*, *15*(8), 1795–1806.

Palaisa, K., Morgante, M., Tingey, S., & Rafalski, A. (2004). Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proc. Natl. Acad. Sci. U.S.A.*, *101*(26), 9885–90.

Papa, R., & Gepts, P. (2003). Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. *Theor. Appl. Genet.* , *106*(2), 239–50.

Park, S. O., Coyne, D. P., Bokosi, J. M., & Steadman, J. R. (1999). Molecular markers linked to genes for specific rust resistance and indeterminate growth habit in common bean. *Euphytica*, *105*(2), 133–141.

Park, S. O., Coyne, D. P., Jung, G., Skroch, P. W., Arnaud-Santana, E., Steadman, J. R., … Nienhuis, J. (2000). Mapping of QTL for Seed Size and Shape Traits in Common Bean. *J. Amer. Soc. Hort. Sci.*, *125*(4), 466–475.

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

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Robbins, M. D., Sim, S.-C., Yang, W., Van Deynze, A., van der Knaap, E., Joobeur, T., & Francis, D. M. (2011). Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato. *J. Exp. Bot.*, *62*(6), 1831–45.

Rosales-Serna, R., Hernández-Delgado, S., González-Paz, M., Acosta-Gallegos, J. A., & Mayek-Pérez, N. (2005). Genetic relationships and diversity revealed by AFLP markers in mexican common bean bred cultivars. *Crop Sci*, *45*(5), 1951.

Schlueter, J. A., Goicoechea, J. L., Collura, K., Gill, N., Lin, J.-Y., Yu, Y., Kudrna, D., Zuccolo, A., Vallejos, C. E., Muñoz-Torres, M., Blair, M. W., Tohme, J., Tomkins, J., McClean, P., Wing, R. A., Jackson, S. A. (2008). BAC-end Sequence Analysis and a Draft Physical Map of the Common Bean (*Phaseolus vulgaris* L.) Genome. *Trop. Plant Biol.*, *1*(1), 40–48.

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., Jenkins, J., Shu, S., Song, Q., Chavarro, C., Torres-Torres, M., Geffroy, V., Moghaddam, S. M., Gao, D., Abernathy, B., Barry, K., Blair, M., Brick, M. A., Chovatia, M., Gepts, P., Goodstein, D. M., Gonzales, M., Hellsten, U., Hyten, D.L., Jia, G., Kelly, J. D., Kudrna, D., Lee, R., Richard, M. M. S., Miklas, P. N., Osorno, J M., Rodrigues, J., Thareau, V., Urrea, C. A., Wang, M., Yu, Y., Zhang, M., Wing, R. A., Cregan, P. B., Rokhsar D. S., & Jackson, S. A., (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.*, *46*(7), 707–13.

Schneider, K. A., Grafton, K. F., & Kelly, J. D. (2001). QTL analysis of resistance to fusarium root rot in bean. *Crop Sci*, *41*(2), 535.

Schneider, K. A., Rosales-Serna, R., Ibarra-Perez, F., Cazares-Enriquez, B., Acosta-Gallegos, J. A., Ramirez-Vallejo, P., Wassimi, N., Kelly, J. D. (1997). Improving common bean performance under drought stress. *Crop Sci*, *37*(1), 43.

Shi, C., Navabi, A., & Yu, K. (2011). Association mapping of common bacterial blight resistance QTL in Ontario bean breeding populations. *BMC Plant Biology*, *11*(1), 52.

Singh, S. P., Gepts, P., & Debouck, D. G. (1991). Races of common bean (*Phaseolus vulgaris*, Fabaceae). *Econ. Bot.*, *45*(3), 379–396.

Skroch, P. W., & Nienhuis, J. (1995). Qualitative and quantitative characterization of RAPD variation among snap bean (*Phaseolus vulgaris*) genotypes. *Theor. Appl. Genet.* , *91*(6-7), 1078–85.

Smartt, J. (1990). *Grain Legumes: Evolution and Genetics Resources* (p. 379). Cambridge: Cambridge University Press.

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., & Buckler, E. S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nat. Genet.*, *28*(3), 286–9.

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T. R., McMullen, M. D., Holland, J. B., Buckler, E. S. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.*, *43*(2), 159–62.

Tohme, J., Gonzalez, D. O., Beebe, S., & Duque, M. C. (1996). AFLP analysis of gene pools of a wild bean core collection. *Crop Sci*, *36*(5), 1375.

USDA. (2014). Crop production 2013 summary. Retrieved from http://usda01.library.cornell.edu/usda/current/CropProdSu/CropProdSu-01-10-2014.pdf

USDA-AMS. (2013). Bean market news 2013 summary. Retrieved from http://www.ams.usda.gov/mnreports/lsaba.pdf

USDA-ERS. (2012). Dry beans. *Vegetables & Pulses*. Retrieved from http://www.ers.usda.gov/topics/crops/vegetables-pulses/dry-beans.aspx#.VDSLMvldVS2

Väli, U., Brandström, M., Johansson, M., & Ellegren, H. (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genetics*, *9*(1), 8.

Velasquez, V. L. B., & Gepts, P. (1994). RFLP diversity of common bean (*Phaseolus vulgaris*) in its centres of origin. *Genome*, *37*(2), 256–263.

Walters, K. J., Hosfield, G. L., Uebersax, M. A., & Kelly, J. D. (1997). Navy bean canning quality: correlations, heritability estimates, and randomly amplified polymorphic dna markers associated with component traits. *J. Amer. Soc. Hort. Sci.*, *122*(3), 338–343.

Wang, H., Smith, K. P., Combs, E., Blake, T., Horsley, R. D., & Muehlbauer, G. J. (2012). Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. *Theor. Appl. Genet.* , *124*(1), 111–24.

Wang, M., Jiang, N., Jia, T., Leach, L., Cockram, J., Comadran, J., Waugh, R., Ramsay, L., Thomas, B., Luo, Z. (2012). Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. *Theor. Appl. Genet.* , *124*(2), 233–46.

Young, R. A., & Kelly, J. D. (1997). RAPD markers linked to three major anthracnose resistance genes in common bean. *Crop Sci*, *37*(3), 940.

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

Yu, K., Park, S. J., & Poysa, V. (1999). Abundance and variation of microsatellite DNA sequences in beans (*Phaseolus* and *Vigna*). *Genome*, *42*(1), 27–34.

Zahniser, S., & Farah Wells, H. (2014). Commodity highlight: dry beans. Retrieved from http://www.ers.usda.gov/media/1680744/vgs-354-sa1.pdf

CHAPTER 2. GENOME WIDE ASSOCIATION STUDY OF AGRONOMIC TRAITS IN COMMON BEAN (*PHASEOLUS VULGARIS* L.)

Abstract

Genome wide association study (GWAS) is a method that utilizes the natural variation that exists in a population of unrelated individuals to discover genomic regions underlying a trait. Discovering genomic regions associated with important agronomic traits provides valuable information for genetic and genomics studies and facilitates breeding through marker assisted selection (MAS). The best combination of agronomic and seed characteristics are the main criteria for selection in common bean breeding. However, the genetic architecture of many important agronomic traits are not completely dissected and GWAS can serve to discover the genomic regions associated with these traits. We conducted a GWAS using a panel of 280 diverse bean genotypes from the Middle American gene pool (MDP) and about 15,000 SNPs with minor allele frequency of 5%. The phenotypic data were provided by breeders from Colorado, Michigan, Nebraska, and North Dakota. The study led to the discovery of known and new genomic regions with strong candidate genes association with the agronomic traits.

Introduction

Common bean (*Phaseolus vulgaris* L.) is one of the most important legumes for human consumption (Broughton et al., 2003). It has a relatively small diploid genome (2n=22; 521.1 Mb; Schmutz et al., 2014). Common bean consists of two main gene pools: Middle American and Andean. Domestication has occurred independently in each gene pool (Gepts and Bliss, 1986; Koenig and Gepts, 1989; Khairallah et al., 1990; Koinange and Gepts, 1992; Freyre et al., 1996; Mamidi et al., 2013). The two gene pools are strongly differentiated, and the Middle American gene pool has greater sequence diversity (Mamidi et al., 2013; Schmutz et al., 2014).

Selection under domestication in each gene pool has generated distinct eco-geographical races and market classes (Singh et al., 1991; Beebe et al., 2000; Díaz and Blair, 2006; Mamidi et al., 2011b). The Middle American gene pool consists of races Durango/Jalisco and Mesoamerican with important market classes in the United States such as pinto, great northern, small red and pink in race Durango/Jalisco, and navy, and black in race Mesoamerican (MA) (Mensack et al., 2010). Many studies have investigated the population structure in the gene pools of common bean and confirmed the above classifications (Blair et al., 2009; Kwak & Gepts, 2009).

Cultivated forms of common bean show considerable morphological variation (Hedrick et al., 1931) for growth habit, and seed size and color (Leakey, 1988; Singh, 1989). Domestication and breeding has selected for agro-morphological traits important for production and altered the genetic architecture underling those traits. Breeding for higher yield in bean is affected by many interdependent traits including growth habit, seed size, and maturity (Kornegay et al., 1992). Genetic factors controlling important agronomic traits have been mapped to common bean linkage groups using bi-parental populations (Beattie, Larsen, Michaels, & Pauls, 2003; Blair, Iriarte, & Beebe, 2006; Blair et al., 2012a; Checa & Blair, 2012; Pérez-Vega et al., 2010; Tar'an, Michaels, & Pauls, 2002; Wright & Kelly, 2011). Nevertheless, our knowledge of genes controlling agronomic traits is limited. QTL analysis usually has low resolution due to the limited number of recombination events in a bi-parental population (Balasubramanian et al., 2009). QTL intervals can span a few centimorgans (cM) which can indeed be megabases (Mb) long in physical distance and contain hundreds of gene models.

In contrast, GWAS has the advantage of capturing more recombination events using an association panel of unrelated individuals. This means higher mapping resolution due to shorter linkage disequilibrium (LD) blocks compared to a bi-parental population. Thus, higher marker

saturation is necessary to cover the whole genome. Higher mapping resolution, though, has only recently been possible with the development of high throughput genotyping in common bean (Hyten et al., 2010, Cregan, unpublished data). The availability of a reference genome sequence (Schmutz et al., 2014) will lead to even higher mapping resolution in GWAS as resequencing of key GWAS mapping parents is completed. Many studies have demonstrated the usefulness of GWAS by their discovery of candidate genes affecting trait expression (Zhao et al., 2011; Korte and Farlow, 2013; Li et al., 2013; Appels et al., 2013; Verslues et al., 2014).

In this study, we conducted GWAS followed by candidate gene identification for seven important agronomic traits that affect common bean production: days to flower, days to maturity, growth habit, canopy height, lodging, seed weight, and seed yield. Over 35,000 SNPs were obtained by combining two Illumina iSelect 6K Gene Chip and genotype-by-sequencing (GBS) data over a collection of 280 diverse genotypes from a population of Middle American genotypes. These are referred to as the Middle American Diversity Panel. We also investigated the population structure and LD at the race, market class, chromosome and genome-wide level.

Materials and Methods

Middle American diversity panel and SNP data set

The Middle American Diversity Panel (MDP) encompasses 280 Middle American dry bean genotypes from the 507 genotypes of the entire BeanCAP[1] diversity panel. This subpopulation was chosen to reduce the effect of population structure during the GWAS. The MDP itself consists of 100 race Mesoamerican (MA) and 180 race Durango/Jalisco (DJ) genotypes, respectively.

[1] http://www.beancap.org/

The SNP data set were obtained by genotyping the MDP with two Illumina iSelect 6K Gene Chip (BARCBEAN6K_1 and BARCBEAN6K_2) sets and by genotype-by-sequencing (GBS). The Gene Chip SNPs were developed at the USDA/ARS Beltsville Agriculture Research Center by aligning 19.6 billion bases of sequence data from 19 genotypes (Cregan, unpublished data). GBS libraries were prepared and analyzed at the Institute for Genomic Diversity (IGD), according to Elshire et al., (2011), using the enzyme *Ape*KI for digestion. Based on the IGD GBS report, the GBS analysis pipeline 3.0.147, an extension to the Java program TASSEL (Bradbury et al., 2007), was used to call SNPs from the sequenced GBS library. Tags were aligned to the reference genome G19833 V1.0 (Schmutz et al., 2014). VCFtools (v0.1.10) (Danecek et al., 2011) was used to summarize and filter the data. Burrows-Wheeler Aligner (BWA) version 0.7.3a-r367 (Li and Durbin, 2009) was used to align the sequences to the reference genome. A total of 5,836,050 tags were generated out of which 1,390,081 were aligned to unique positions (23.8%), 175,125 were aligned to multiple positions (3%), and 4,270,844 could not be aligned (73.2%). A total of 64,478 SNPs were obtained which was reduced to a final number of 25,618 SNPs after filtering for missing data.

A total of 10,783 SNPs from the two 6K Gene Chip were mapped to the common bean reference genome (Schmutz et al., 2014). The 10,783 SNPs from the two 6K Gene Chips and 25,618 SNPs from the GBS platform were merged based on their position in the reference genome. None of the SNPs overlapped. Both genotypes and SNPs were filtered for missing data, and those with 50% or greater missing data were excluded. None of the genotypes met this criterion but 1,365 SNPs were discarded leading to 280 genotypes and 35,036 potential SNPs for the GWAS. Missing data were imputed using the likelihood based method implemented in

fastPHASE 1.3 (Scheet and Stephens, 2006) with default settings. The polymorphic information content (PIC) was calculated in PowerMarker (Liu and Muse, 2005).

Phenotypic analysis

Seven agronomic traits were collected from field trials grown by BeanCAP collaborators in Colorado, Michigan, Nebraska and North Dakota. The traits are: days to flower, days to maturity, growth habit, canopy height, lodging, seed weight, and seed yield. Days to flower was measured as the number of days from planting to when approximately 50% of the plants in a plot have at least one opened flower. Days to maturity was measured as the number of days from planting until harvest maturity. Growth habit was recorded during flowering and verified at senescence as type I= determinate erect or upright; type II= indeterminate erect; and type III= indeterminate prostrate. Canopy height was measured at harvest and was recorded in centimeters from the base of the plant (soil surface) to the top of the canopy. Lodging was scored at harvest on a 1 to 5 scale, where 1 =100% plants standing erect, and 5= 100% plants flat on the ground. Seed yield was recorded in kg/ha at 16% moisture and rounded up to the nearest whole number. Seed weight was measured as the weight of 100 randomly selected undamaged seeds recorded in grams at 16% moisture. Growth habit data for Nebraska were not available, and the phenotypic data for growth habit and lodging in North Dakota were not used for the analysis. The analysis were conducted separately for growth habit with determinate genotypes included and excluded. The experimental design was an incomplete block design with two replicates. The trait heritability was calculated for each subpopulation (DJ and MA) using PROC MIXED in SAS 9.3(SAS institute, 2011) based on the method proposed by Holland et al., (2003). Correlation among traits was calculated and ploted in R 3.0 (Team, 2013) using cor.matrix() and corrplot() from corrplot package (Wei, Taiyun, 2013). The histograms and boxplots were created in R 3.0

(Team, 2013) using the hist() and boxplot() functions, respectively. The North Dakota breeding program provided the adjusted phenotypes, which were calculated using LS means in SAS 9.3 (SAS institute, 2011) per location and across all locations. This led to a total of 32 phenotypes to be analyzed by GWAS for the entire population.

Population structure and kinship

Both population structure and kinship were calculated using markers with pairwise R^2 of less than 0.5 for all pairwise comparisons. STRUCTURE.2.3(Pritchard et al., 2000), a model based method, was used to assign the subpopulation membership to each individual. We used an admixture model with independent allele frequencies, a burn-in of 100,000 and an MCMC replication of 500,000 for K = 1 to 10 with five replications. The Wilcoxon two sample *t*-test implemented in SAS 9.3 (SAS institute, 2011) was used to select the optimum number of sub-populations (Rosenberg et al., 2001). The optimum number of subpopulations was the smallest K in the first non-significant Wilcoxon test. Distruct1.1 (Rosenberg, 2003) was used for graphical display of the STRUCTURE output. The Principal Component Analysis (PCA) using PRINCOMP in SAS 9.3 (SAS institute, 2011) was employed to control for population structure in GWAS (Price et al., 2006). To account for individual relatedness, an identity-by-state kinship matrix was generated using the EMMA algorithm (Kang et al., 2008) e Mbedded in GAPIT (Lipka et al., 2012).

Linkage disequilibrium

Pairwise linkage disequilibrium (LD) between markers in the null model was calculated as the squared allele frequency correlation in PLINK (Purcell et al., 2007) after filtering for minor allele frequency (MAF) ≥ 5%. R package LDcorSV (Mangin et al., 2012) was used to calculate the pairwise linkage disequilibrium when accounting for population structure,

relatedness or both. LD heat maps were generated for each chromosome in R 3.0 (Team, 2013) using the LDheatmap package (Shin et al., 2006). The pericentromeric borders were determined based on Supplementary Table 6 from Schmutz et al., (2014).

Genome-wide association study (GWAS)

A total of 15,343 markers with minor allele frequency ≥ 5% were used for the MDP GWAS analyses. These were performed using GAPIT, an R package developed by Lipka et al. (2012). Multiple models were tested per trait: null general linear model, general linear models with fixed effects to control for population structure, and univariate unified mixed linear model (Yu et al., 2006b) using the population parameters previously determined (P3D) (Zhang et al., 2010) protocol to control for relatedness (random effect), or both relatedness and population structure . We used principal components to account for population structure. Both, 7 PCs (controls for ~25% of variation), and the optimal number of PCs selected by BIC in GAPIT were tested. Finally the best model was selected based on the mean squared difference (MSD) as described by Mamidi et al. (2011). The phenotypic variation explained by significant markers was calculated using a likelihood-ratio-based R^2 (Sun et al., 2010) using the genABLE package in R (Aulchenko et al., 2007). GWAS were conducted for all traits with the complete genotypic panel as well as separately for each race. Analyses were performed using pooled data across all locations and for each location separately.

For traits with major genotypic effect, multi-locus mixed model (MLMM) (Segura et al., 2012) was used to evaluate the results for single marker tests. MLMM reduces the masking effect of population structure or selection on causative loci by forward and backward stepwise regression. Markers in complete LD (R^2=1) were excluded for this analysis. Multiple-Bonferroni

criterion (Mbonf) implemented in MLMM was used to select the best model (Chen and Chen, 2008).

Candidate gene identification

Markers mapped to scaffolds but not the main assembly were excluded for candidate gene detection. Two approaches were used to find the candidate genes: 1) *a priori* candidates based on the literature, and 2) evaluation of genes up to 200Kb upstream and downstream of the significant markers. Significant markers were defined as those falling within the 0.1 percentile tail of the empirical distribution of p-values after 10,000 bootstraps (Mamidi et al., 2014).

Results

Population structure

The admixture model in STRUCTURE (Pritchard et al., 2000) was used to assign subpopulation membership coefficients to the individual genotypes. Based on Wilcoxon two-sample *t*-test, the MDP consisted of seven subpopulations corresponding to the seven main market classes in the Middle American gene pool (Figure 2.1.A). K=2 subpopulations shows the split between races Mesoamerican and Durango, and for K=3 to K=7 the genotypes clustered according to the market classes. As expected, admixture was observed between the market classes. Figure 2.1.B shows the first three principal components from PCA that correspond to

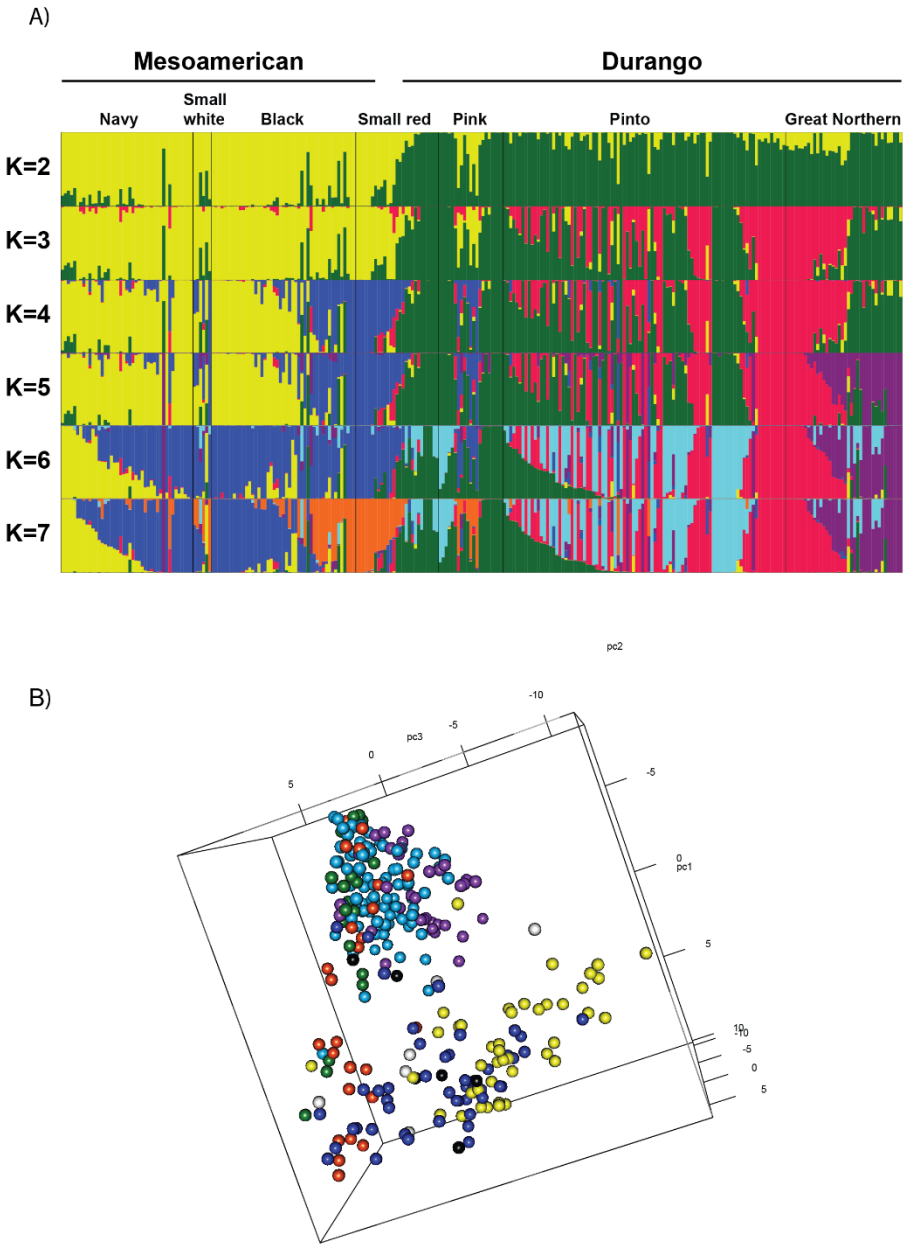


Figure 2.1. STRUCTURE and principal component analysis. **A**) STRUCTURE analysis of the MDP from *K*=2 to *K*=7. *K*=2 shows the split between two major races and from *K*=3 to *K*=7 each race subdivides into market classes. **B**) Three-dimensional principal component analysis of the MDP. The first dimension separates the two major races. The color coding of market classes correspond to that of part A.

the STRUCTURE results. The first component separates races Mesoamerican and Durango. Pinks and small red genotypes were distributed across both races, a result consistent with the admixture pattern in these market classes from the STRUCTURE analysis (Figure 2.1.A). PC2 clusters some of the pinto and great northern genotypes closer to pink and small red market classes. PC3 separates navy and black market classes in the race Mesoamerican and shows that small reds and some pinks are more closely related to the black market class than navy.

Linkage disequilibrium

LD heat maps were created for each chromosome in MDP, race MA, and race DJ to compare the extent of LD at different levels of population structure. The LD pattern not only varied between subpopulations but also varied by chromosome within a subpopulation even after controlling for population relatedness (Figure A.1. A and B). For example, Pv06 and Pv02 show high LD in their euchromatic regions compared to other chromosomes in both DJ and MA subpopulation. Pv06, Pv07 and Pv11 show dramatically different LD patterns among races. Generally, Pv01, Pv07 (except race MA), Pv08, Pv10 and Pv11 show large LD blocks in their pericentromeric region.

Figure 2.2 is a genome-wide LD heat maps that includes only LD values > 0.6. The triangle above the diagonal is based on the null model whereas the under the diagonal image is corrected for both population structure and relatedness. The majority of inter-chromosomal LD occurs among pericentromeric regions. Pv08 and Pv09 have the largest (38.5 Mb) and smallest (5.7 Mb) pericentromeric regions, respectively (Schmutz et al., 2014). However, the largest inter-chromosomal LD is between Pv07 and other chromosomes; an almost 30 Mb region, slightly larger than the pericentomeric region (27 Mb), is in LD with either a single SNP or narrow regions on other chromosomes. Only Pv05 and Pv09 do not exhibit any $r^2 \geq 0.6$ with

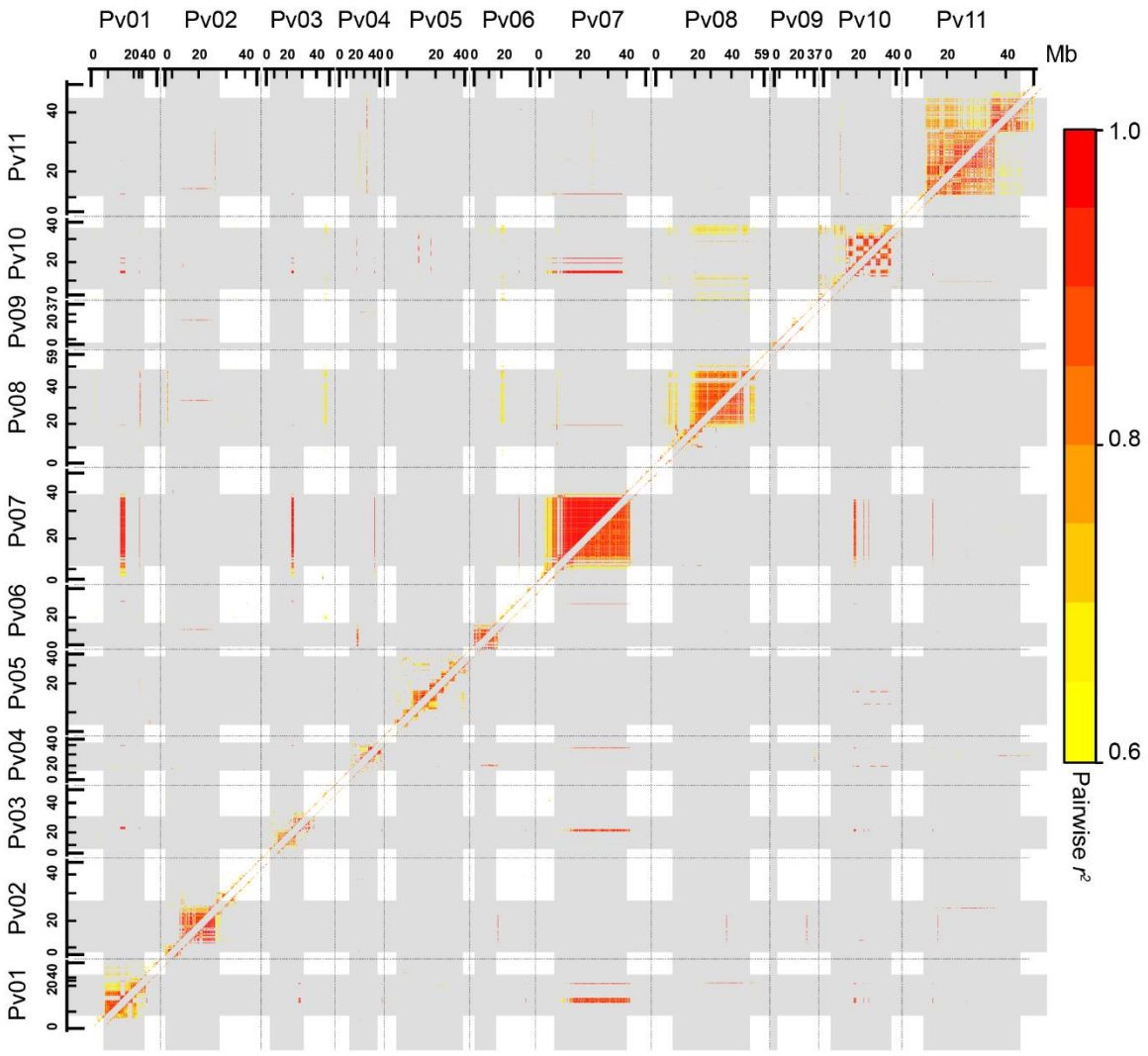


Figure 2.2. Genome-wide LD heat map. Data above the diagonal represent the null model, and data below the diagonal image represents the model that accounts for both population structure and relatedness. Markers every 50Kb were used and only pairwise $r^2 > 0.6$ are shown. The grey rectangular show the pericentromeric regions. Grey dashed lines define the chromosome boundaries.

Pv07. Pv08 and Pv11 have a large 30 Mb block in LD with small regions or single SNPs on Pv07. The other prominent inter-chromosomal LD could be between Pv08 and Pv11. Pv08 is in LD with Pv01, Pv02, Pv03, Pv06, Pv07, and Pv10 in the null model (Figure 2.2). After controlling for population structure and relatedness, Pv08 remains in high LD with Pv07, small

regions on Pv01 and Pv02. Pv11 shows LD blocks with Pv02, Pv04, Pv07, and Pv10. Although the mixed model significantly reduced the overall inter-chromosomal LD, some long range LD blocks persisted. The DJ subpopulation shows a long range LD pattern similar to that of MDP but very little inter-chromosomal LD is detected for the MA subpopulation (Figure A.2. A and B).

Phenotypic analysis

The phenotypic expression of all agronomic traits varied between locations and subpopulations. The longest and shortest days to flower occurred in North Dakota (49-68 d) and Michigan (35-55 d), respectively. The average days to flower across all the locations was 46 and 49 days in the MA and DJ subpopulations, respectively. Days to maturity ranged from 60-125 days. The plants matured earlier in Michigan (96-110 d) compared to North Dakota (75-125 d). Canopy height varied from 21 to 89 cm with the widest range in Colorado. Seed weight showed a bimodal distribution due to population structure with values ranging from 14.1 to 39.7g/100 seeds in race MA and 21.1-53.7g/100 seeds in race DJ. Seed yield mean ranged between 442 to 5,539 Kg/ha in different locations. North Dakota and Michigan trials trended toward smaller yields while Colorado and Nebraska trials trended towards higher yields. Figure 2.3 shows the distribution of each trait across different locations and combined. Trait values were highly correlated among locations except for seed yield, and days to maturity in North Dakota. Days to maturity was positively correlated ($r > 0.5$) with days to flower among different locations. Growth habit and lodging were positively correlated but canopy height in Michigan and Nebraska were negatively correlated with growth habit and lodging. Seed weight and days to flower had negative correlation ($r = -0.40$ to -0.54) except in North Dakota. Seed yield was not

correlated with any trait (Figure A.3). Table 2.1 shows the heritability values for each trait in races DJ and MA.

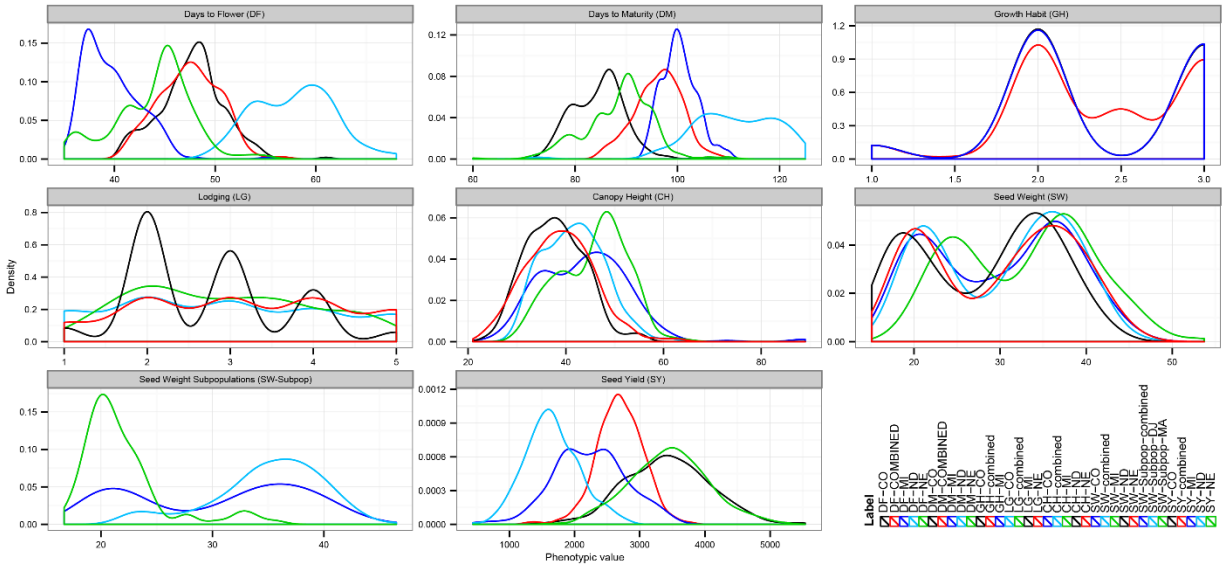


Figure 2.3. Density plot of phenotypic distribution of seven agronomic triats. Color coding is provided in the label section. CO: Colorado, MI: Michigan, NE: Nebraska, ND: North Dakota, Combined: across all the locations, DJ: Durango/Jalisco, MA: Mesoamerican.

Table 2.1. Heritability of agronomic traits across four locations.

Trait	Durango/Jalisco		Mesoamerican	
	Plot Basis	Family Mean Basis	Plot Basis	Family Mean Basis
Days to Flower	0.54	0.87	0.53	0.88
Days to maturity	0.35	0.71	0.16	0.53
Growth habit (with determinates)	0.68	0.88	0.65	0.86
Lodging	0.56	0.85	0.44	0.78
Canopy height	0.51	0.87	0.40	0.81
Seed weight	0.74	0.94	0.81	0.96
Seed yield	0.19	0.56	0.19	0.57

Genome wide association study (GWAS)

We identified genomic regions underlying seven agronomic traits using 15,284 SNPs with a MAF ≥ 5%. The number of SNPs dropped to 14,478 and 12,600 SNPs in DJ and MA

subpopulation analysis, respectively due to lower genetic diversity in each subpopulation compared to full population. The F_{ST} between the DJ and MA subpopulation was 0.21 which indicates a great degree of differentiation (Hartl and Clark, 2007). The polymorphism information content (PIC) and gene diversity was similar between the two subpopulations for markers with MAF ≥ 5%. PIC ranged from 0.09 to 0.37 in race DJ and from 0.10 to 0.37 in race MA with an average value of 0.26 and 0.25 in race DJ and MA, respectively. Gene diversity ranged from 0.09 to 0.5 in race DJ with an average of 0.32. These values were 0.11 to 0.50 and 0.31 for MA subpopulation.

The p-values for the best model varied among traits, locations and subpopulations. Thus the Bonferroni cutoff would not be a suitable approach to identify the significant SNPs. Rather, we bootstrapped 10,000 times the p-values for each trait, and SNPs falling in the top 0.1 percentile of the empirical distribution were considered significant. This led to a different cutoff threshold for each trait and sometimes was even more stringent than the Bonferroni. Table A.1 shows the significant SNPs, and the candidate gene models within 200Kb of the significant marker.

Days to flower and maturity

The strongest signal for days to flower is on Pv01 in a region of extensive LD and includes significant markers in a large block from approximately 8 Mb to 20 Mb. The most significant GWAS peak on Pv01 differed among the races. The top GWAS peak for race DJ genotypes occurred at ~8 Mb while in race MA the top peak was at ~15 Mb (Figure B. A, B, C). It is difficult to determine whether both regions are equally important in both subpopulations or one of the regions is the main peak in each race and the other peak is due to extensive LD. Indeed, the 8 Mb region remains in LD (r^2 = ~0.6) with the 15 Mb and 20 Mb regions after

controlling for both population structure and relatedness in the best GWAS model. To further investigate this region, we used the multi-locus mixed-mode (MLMM) method (Segura et al., 2012) for the whole population and each subpopulation. The most significant SNP controlled all the variation on Pv01 when used as a cofactor in the MLMM analysis. The same result was observed at the subpopulation level. We also used the most significant SNP in the MA subpopulation (m32210/15.82 Mb) as a cofactor in MLMM analysis for race DJ. This controlled for all the variation on Pv01. In the next step when the second top SNP, which now appears on Pv03 (m2535/48.03 Mb), was used as a cofactor, once again the most significant region in DJ (8 Mb) appeared. The optimum model based on Mbonf included both m32210 (Pv01) and m32210 (Pv03) as cofactors (Table 2.2). It is noteworthy to mention that m2535 on Pv03, although not found by linear regression but using the stepwise regression approach, is inside the gene model *Phvul.003G252400* which encodes a C2H2 zinc finger protein. The *Arabidopsis* and maize homolog to this gene is *INDETERMINATE DOMAIN 1* (*IDD1)* which was first found in maize as a key regulator of the transition to flowering (Colasanti et al., 2006). When the most significant marker in race DJ (m32381/8 Mb) was used as a cofactor in a MLMM analysis for the MA subpopulation, the significant peak in MA remained significant (m32211 which is adjacent to m32210/15.82 Mb). Only after using m32211/15 Mb as a cofactor, was all the variation on Pv01 controlled. This result might imply that the 15 Mb region has an important role in the MA subpopulation, and its signal in MA is independent from the signal on 8 Mb. The candidate gene models on Pv01 can be found in Table A.1. We also conducted GWAS after removing 19 genotypes with determinant phenotype. The results were similar to the original analysis (data not shown).

For days to maturity, GWAS peaks were noted across the genome and different peaks in each location and subpopulation. The most noticeable peak is at the beginning of Pv11. For North Dakota, the most significant peak was found on Pv08, and for Nebraska the significant peaks were located on Pv01 and Pv11 (Figure B. D, E, F). Most of the candidate genes in Table A.1 are either involved in flowering time or senescence and nutrient remobilization.

Growth habit, lodging, and canopy height

We conducted separate GWAS including and excluding determinate genotypes because determinate genotypes might have a different underling genetics for their architecture due to their determinate growth habit. Types II and III are both indeterminate but differ in their architecture. GWAS on the entire MDP population showed a major signal on Pv01 including the end of this chromosome where flowering time candidate genes *LWD1*, *SPY*, and *TFL-1* reside (Figure B. J, K, L). When the determinate genotypes were excluded from the analysis, the peak on Pv01 was lost and a major peak appeared on Pv11 (Figure B. G, H, I). Pv06/21.49 Mb was significant in Michigan with or without determinate genotypes.

Lodging showed a strong consistent Pv07/ ~46 Mb peak across all the locations as well within each location (Figure B. M, N, O). The peak is 68.6Kb wide and consists of multiple SNPs, with pairwise r^2 values of greater than 0.5. In the null model, the 46 Mb region shows $r^2 \approx$ 0.6 with two other single significant SNPs (m10969; Pv07/48.63 Mb and m10611; Pv07/45.18 Mb) and the 47 Mb region on the same chromosome, but the LD drops to $r^2 < 0.5$ after controlling for population structure and relatedness in the best model. When m10689 /46.13 Mb is included as a cofactor in MLMM analysis, it controls for all the variation on Pv07 (Table 2.2). However, when SNPs at 48 Mb or 45 Mb are used as cofactors, m10689/46.13 Mb still remains the most significant peak which might imply that the main signal on Pv07 is at 46 Mb and the

other two regions might or might not be true associations. A single significant SNP (m19820) on Pv01 passes the significance cutoff only across all the locations and in Colorado. The SNP is polymorphic only in the DJ subpopulation. When the Pv07/46 Mb peak is used as a cofactor in a MLMM analysis, this region becomes the strongest signal (Table 2.2). The beginning of Pv05 in Michigan shows a major peak for which we were not able to find any candidate genes.

The same peaks on Pv07 at positions 46 Mb and 45 Mb were significant for canopy height across all the locations and in each location except for North Dakota (Figure B. P, Q, R). SNP m16972 on Pv10 was only significant in Michigan and is 38.7Kb upstream of *Phvul.010G021200* encoding the Arabidopsis *RIN4* homolog. In addition to the significant regions on Pv07 that are shared between lodging and canopy height, the 43 Mb region on Pv11 was significant for canopy height in Nebraska. This region was significant also for growth habit and is close to the *Phvul.011G164800* (*SPL4*) gene model (Table A.1).

Seed weight and seed yield

The major peaks for seed weight reside on Pv10 which appears in all locations and subpopulations. The next prominent peaks are located on Pv03 and Pv06 (Figure B. S, T, U).

GWAS for seed yield revealed multiple peaks on different chromosomes (Figure B. V, W, X). The end of Pv03 and Pv06 are significant for both seed yield and seed weight. *ASN1* (*Phvul.006G069300*) on Pv06 is a candidate gene for both seed weight and yield although it is only significant in Michigan for seed yield (Table A.1). This gene affects the seed protein content and seed weight (Lam et al., 2003). Unique peaks on Pv05 and Pv11 were detected in Colorado and North Dakota, respectively, but no candidate genes were identified.

We did not perform MLMM for seed weight and seed yield because the best model only includes seven PCs to control for population structure, and a kinship matrix is required for MLMM analysis.

Table 2.2. Significant markers in the optimal step of MLMM analysis based on Mbon.

	Optimal step (Mbonf)	
Trait	**Cofactor1**	**Cofatcor2**
Days to flower	m32210 (Pv01)	-
Days to maturity	-	-
Growth habit (with determinates)	m17978(Pv01)	m31418(Pv01)
Growth habit (No determinates)	m20650(Pv11)	m10689(Pv07)
Lodging	m10685(Pv07)	m19820(PV01)
Canopy height	m10611(Pv07)	-

Discussion

Understanding the genetic architecture controlling important traits has both biological and breeding implications. The availability of whole genome sequence data (Schmutz et al., 2014) has provided common bean with a wealth of genomic information and makes it possible to further saturate the genome with genetic markers useful for genomics and population studies. Indeed, higher marker coverage enhances the accuracy of QTL and GWA studies that are important tools to analyze the genetic architecture of any trait (Varshney et al., 2009; Deschamps and Campbell, 2009; Davey et al., 2011). Utilizing imputation, we combined two common bean Illumina iSelect SNP arrays with GBS data to obtain 15,284 SNPs with a MAF≥5% that are distributed across the genome. The SNP data set was used to dissect the underlying genomic regions controlling important agronomic traits in a collection of varieties and important germplasm from the Middle American gene pool.

Linkage disequilibrium

Like many other species, common bean possesses both intra and inter-chromosomal LD. The extent of LD between unlinked markers can confound association studies. Thus, the results should be carefully interpreted. Based on the marker set used in this study, the pattern of intra-chromosomal LD varies among all chromosomes for the full population as well as for the same chromosome among subpopulations. The same observation is true for inter-chromosomal LD. SNP allele frequencies differed among the Mesoamerican and Durango/Jalisco subpopulations. This affected LD estimates in the two subpopulations. Many markers that caused long range inter-chromosomal LD in race DJ were monomorphic or had a MAF<5% in race MA, thus were not included in in the LD heat maps. This lead to a very low inter-chromosomal LD due to absence of informative markers. Therefore, the extent of LD depends on the population and the marker set used for the analyses. Due to low diversity among the race MA genotypes, we did not have enough informative markers to determine the extent of long range LD in that race. These inter- and intra-chromosomal LD patterns in common bean might define the characteristics of bean itself, its races, and market classes. They might also be a result of selecting for favorable allelic combinations at multiple loci on different chromosomes during the breeding process. Such haplotype blocks of LD have been observed in other species (Patil et al., 2001; Wiltshire et al., 2003; Lindblad-Toh et al., 2005; Tang et al., 2006; Li et al., 2009; Gore et al., 2009; Comadran et al., 2011; Robbins et al., 2011). Although the mixed model significantly reduced the overall inter-chromosomal LD, some long range LD blocks persisted.

Genome wide association study (GWAS)

Days to flower

Pv01 shows considerable intra-chromosomal LD after controlling for both population structure and relatedness. This makes it difficult to dissect the major Pv01 GWAS peak for days to flower at the proximal end of the chromosome. In fact, most significant SNPs mapping close to flowering time candidate genes are located in the low recombination pericentromeric region and have LD values ≥ 0.6. It remains to be determined whether these genes are part of a large functional network necessitating such extensive LD or whether only a few genes control flowering in common bean and the rest are false positives due to LD caused by random evolutionary processes. The first candidate gene on Pv01, *Phvul.001G064200* is an ortholog of *SEUSS* (*SEU*) which affects floral meristem identity in coordination with other flowering genes in *Arabidopsis*. It is a component of a complex that represses AGAMOUS (AG) expression (Gregis et al., 2006) in all four whorls in early stages of flower development where *AG* confers the final definition of floral meristem identity (Mizukami and Ma, 1997; Conner and Liu, 2000; Franks et al., 2002; Sridhar et al., 2004; Liu and Karmarkar, 2008). For bean, *SEU* gene expression is the lowest for flower and flower buds in G19833 (Schmutz et al., 2014). Further confirmation for the relevance of this region is provided by Blair et al. (2006) who reported the *df1.1* QTL for days to flower in a cross between an Andean and a wild bean. The closest marker to this QTL, *BMd32*, is physically located between 7,034,858 bp to 7,035,622 bp on Pv01, less than 1 Mb upstream of *SEU*. Since our analysis does not include Andean types, it is not possible to infer that the same region underlies the flowering time in both gene pools but this possibility cannot be ruled out.

The KNUCLE (*KNU*) ortholog, *Phvul.001G087900*, near the Pv01/15 Mb GWAS peak is another candidate and has its highest RNA expression in the flower buds (Schmutz et al., 2014). In *Arabidopsis*, *KNU* is induced by AG and in turn represses *WUS* to maintain the meristem proliferative phase. This process promotes the differentiation phase at a certain time and ensures floral determinacy (Payne et al., 2004).

At the Pv01/20.16 Mb GWAS peak is *Phvul.001G094300* an ortholog of *Arabidopsis BRG-1 ASSOCIATED FACTOR 60* (*AtBAF60*) which represses *FLOWERING LOCUS C* (*FLC*). *FLC* represses flowering, thus down regulation of *BAF60* in *Arabidopsis* results in late flowering under long days (Jégu et al., 2014).

The peak Pv01/45.8 Mb marker, m18033 is located in the *SPINDLY* (*SPY*) bean homolog (*Phvul.001G192300)*. *SPY* negatively regulates gibberellic acid (GA) production, a key hormone required for flowering. Several *SPY* mutants flower under short days. Immediately next to the bean SPY ortholog is *Phvul.001G192200*, the bean ortholog for *LIGHT-REGULATED WD1* (*LWD1*) with a significant SNP nearby (Table 3.2). *LWD1* is a recently discovered circadian gene that shows diurnal expression under short and long days and renders the plant photoperiod insensitive. *Arabidopsis* double mutant *lwd1/lwd2* shows early flowering under both short and long days but is more prominent under short days. In this mutant the expression of oscillator proteins increased and their period length decreased to 21 hours which implies the whole process advanced three hours. *LWD* is thought to control the expression of central oscillators which suppress flowering before twilight. The biological role of homologous bean genes are not clear at this point. If this pathway and homologous genes act similarly in common bean, *Phvul.001G192200* might be a candidate gene that confers photoperiod insensitivity to some cultivars of this species. *Phvul.001G189200*, an ortholog of Arabidopsis *TFL1* and known as the

fin locus candidate (Repinski, Kwak, & Gepts, 2012) is located 244Kb upstream of LWD1 and SPY.

In brief, the *Arabidopsis* candidate gene orthologs that map near flowering time GWAS peaks are involved in GA, autonomous/FLC and the photoperiod/FT pathway that all control the commitment to flowering. The exact function of most of these candidate genes remains unknown in common bean.

Days to maturity

The exact biological pathway(s) that control days to maturity is not clear. Thus we looked for candidate genes among the known or hypothesized pathways involved in the maturity process such as leaf senescence and nutrient remobilization, seed development, and flowering time due to its positive correlation with days to maturity in ours and other studies (Blair et al., 2012b). Plants that produce mature fruits undergo senescence as the final step of their development. Nutrients are then remobilized and stored in fruits. For example removing the flowers or restricting the pod growth in soybean, postponed leaf senescence. (Miceli et al., 1995; Noodén, 1988). Thus leaf senescence is known as an important maturation process. *AtNHL10*, an ortholog of *Phvul.007G268700* on Pv07, is a late embryogenesis abundant (LEA) gene whose transcripts appear in senescing leaves in a salicylic acid (SA) dependent manner (Zheng et al., 2004). However, its exact function in the senescence pathway is not known. *SAG12* (*Phvul.011G081100*) at Pv11/7.3 Mb GWAS peak encodes a cysteine protease (Lohman et al., 1994). This has served as a senescence marker, which is only activated by aging not by stress or hormones (Noh and Amasino, 1999). *SAG12* was found to be highly expressed in aged rapeseed and during increased nitrogen remobilization in nitrate deficient soils (Desclos et al., 2009). This might imply a role in nitrogen remobilization. Martínez et al., (2008) and Otegui et al., (2005),

found SAG12 as a component of the small lytic vacuole called the senescence-associated vacuole (SAV). SAVs have acidic components with proteolytic activity such as SAG12, and could be taken up by outolysoms. Avila-Ospina et al., 2014 proposed induced autophagy is associated with leaf senescence and nitrogen remobilization. SAG12 is exclusively associated with senescence and is usually used as a senescence marker.

The other days to maturity GWAS peak, Pv11/4.3 Mb, is inside the gene model *Phvul.011G050300*, an ortholog of *Arabidopsis TARGET of RAPAMYCIN* (*TOR*). TOR is a member of a signaling pathway that perceives the plant's nutrient and energy status(Kim & Guan, 2011; Wullschleger, Loewith, & Hall, 2006). *TOR* harnesses senescence and nutrient recycling by regulating *SERINE/THREONINE PROTEIN PHOSPHATASE 2A* (*PP2A*) activity (MacKintosh, 1992). Genes involved in nitrogen flow, such as glutamate dehydrogenase and glutamine synthetase 1, are induced in loss of function *TOR* mutants in *Arabidopsis*. On the other hand, enhanced *TOR* expression increases organ and cell size, seed production and osmotic stress tolerance (Deprost et al., 2007). The GWAS peak at this region extends about 145Kb. On the other end of this peak, is gene model *Phvul.011G052100* which belongs to the plant regulator *RWP-RK* family protein. Its *Arabidopsis* homolog, *AtNLP8,* is preferentially expressed in aged leaves and seeds (Chardin et al., 2014).

One candidate gene that might affect days to maturity through flowering is *Phvul.011G053300* in Pv11/4.46 Mb GWAS peak, an ortholog of *Arabidopsis FPA* gene. *FPA* is part of the autonomous flowering pathway and flowering is delayed in its mutant allele (Koornneef et al., 1991). Plants overexpressing *FPA* flower earlier in short days and are insensitive to photoperiod (Scho Mburg, 2001). Another flowering candidate gene is the ortholog of *FRIGIDA LIKE 1* (*FRL1*) which resides at the end of Pv08 (58.40 Mb). This gene delays

flowering in winter annual *Arabidopsis* (Michaels et al., 2004). *Phvul.007G267100*, a *NUCLEAR FACTOR Y*, *SUBUNIT A2* (*NF-YA2)* in Pv07/50.52 Mb GWAS peak, is a target of microRNA *miR169d*,and modifying the *miRNA169d* target site delays flowering (Xu et al., 2014).

We also found two homologs of *Arabidopsis REDUCED VERNALIZATION RESPONSE 1* (*VRN1*), *Phvul.007G022900* and *Phvul.011G050800*, both ~5Kb upstream of the Pv07/1.58 Mb and Pv11/4.38 Mb GWAS peaks, respectively. However, common bean is a tropical legume that doesn't undergo vernalization. Thus, if these are true positive associations, the role of vernalization genes in flowering or maturity time remains unknown in bean. Interestingly, Levy et al., (2002) reported that constitutive overexpression of *VRN1*, led to early flowering without vernalization in *Arabidopsis*.

Growth Habit

Upright growth habit is favorable because it facilitates direct harvest in dry bean. Moreover, it reduces the risk of disease by increasing ventilation in the canopy and minimizes pod contact with soil (Coyne & Schuster, 1974; Park, 1993; Schwartz et al., 1987). Our phenotypic data shows a negative correlation between growth habit and canopy height. Genotypes with type III growth habit might have shorter canopy height due to their prostrate growth habit. We do not have any information about the actual stem length, node number, and internode length in this study. Many studies on wild beans reported significantly higher stem length in type III growth habit (García, 1997).

When determinate genotypes are excluded from the analysis, the major GWAS peak is Pv11/43.2 Mb. This peak contains four significant SNPs located in *Phvul.011G164800*; three of these are in complete LD with each other and two of the SNPs cause non-synonymous substitutions (C145G and S146N). Interestingly, the best *Arabidopsis* homolog for this gene

model is *SQUAMOSA PROMOTER BINDING PROTEIN-LIKE* 4 (*SPL4)* (4E-33, 55% identity). The *Arabidopsis* and the bean homolog contain a *miR156* cleavage site in their 3' UTR. In *Arabidopsis*, *SPL3, SPL4*, and *SPL5* are reported to act redundantly to promote flowering and affect juvenile to adult transition (Schwarz et al., 2008; Wang et al., 2009; Wu & Poethig, 2006; Yamaguchi et al., 2009). Shikata et al., (2009), proposed that *SPL3, SPL4, SPL5, SPL9*, and *SPL15* take part in shoot development during reproductive stage. The *Arabidopsis* co-expression of *SPL11, SPL2, SPL3, SPL4, SPL5, SPL9*, and *SPL15* suggests that they are likely to be involved in the same or related pathways (Obayashi et al., 2007).

Another GWAS peak is the Pv06/21.4 Mb near *Phvul.006G097400.* This gene is homologous to *Arabidopsis HASTY* (*HST1*) that affects both flowering and shoot juvenile to adult phase transition in *Arabidopsis*. *HST1* is involved in miRNA transport into the cytoplasm, and its mutant form accumulates less cytoplasmic miRNA (Park et al., 2005). miR156 affects plant architecture by regulating *SPL* transcript level in grasses, legumes, and trees. In rice, control of *OsSPL14* via *miR156,* creates an "ideal" plant with less tillers, higher yield and improved lodging resistance (Jiao et al., 2010). Aung (2014) reported three *Medicago sativa SPL* orthologs (*SPL6, SPL12*, and *SPL13*) that are controlled by *miR156*. Overexpression of *miR156* in alfalfa, down-regulated the expression of three *SPL* genes leading to architectural changes such as decreased internode length, increased branching and higher biomass. In poplar, overexpression of maize *miR156* resulted in enhanced branching, reduced internode length and stem lignin (Rubinelli et al., 2013). Poethig (2009) proposed that the timing of the juvenile to adult phase transition might be the underlying reason of branching regulation in the *miR156-SPL* pathway. Thus, *HST1* and *SPL4* are logical candidate genes regulating branching, a key contributor to the Type III growth habit.

When determinate genotypes are included in the GWAS, the major peak for growth habit is located on Pv01 with the smallest p-values at ~45 Mb close to two flowering time candidate genes, *LWD1* and *SPY*. Also located near this GWAS peak is *TFL*-1, the ortholog of bean gene *Fin*, whose recessive allele controls the determinate phenotype. Therefore *Fin* and not LWD1 and *SPY* is the likely candidate here.

In addition to candidate genes related to flowering time on Pv01, we found candidate genes related to stem length. *Phvul.001G128800* encodes a cytokinin oxidase (*ATCKXX1*) which reduces cytokinin content. Cytokinin deficiency leads to increased root growth but decreased stem growth by affecting the cell proliferation in the meristem (Werner et al., 2003). Moreover, cytokynin mediates the effect of auxin on apical dominance. Cytokinins travel from root to break auxiliary buds dormancy (Palni et al., 1988; Cline, 1991; Bangerth, 1994; Nordström et al., 2004) which then leads to increased branching. This gene maps ~57kb from the peak marker but no other intervening genes are located in this low recombination, high LD region of Pv01.

Lodging and canopy height

Four Pv07 GWAS peaks spanning 3.5 Mb are shared between lodging and plant height. The single major peak for lodging contains a cluster of SNPs that cover five gene models at Pv07/46 Mb GWAS peak. Three of these SNPs map within genes that could be involved in establishing plant architecture. A SNP within *Phvul.007G221800*, a leucine-rich repeat receptor-like protein kinase, causes a non-synonymous substitution (S740T). This gene model is homologous to *BRASSINOSTEROID INSENSITIVE 1 PRECURSOR* (*29904.t000125*) in *Ricinus communis* with 76% similarity (Goodstein et al., 2012). *BRI1* is a membrane-localized LRR receptor-like kinase (RLK) that perceives and conveys the brassinosteroid signal (Clouse, 2002; Peng & Li, 2003; Thummel & Chory, 2002; Wang & He, 2004). Plants defective in *BRI1* express

dwarfism. The *Arabidopsis* homolog (*At2g33170*) affects root length through cytokinin (ten Hove et al., 2011), but there are no reports on its effect on shoot lenght. There is also a *RING/U-box* superfamily protein (*Phvul.007G221700*) in this region whose function has not been characterized. It contains a C3HC4 zinc finger domain and could act as a transcription factor that affects expression of its downstream genes.

Phvul.007G246700 at Pv07/48.6 Mb encodes a plant invertase/pectin methylesterase inhibitor. The significant marker m10689 is located within this gene. Pectin is produced in golgi and then secreted into the cell walls. The changes in their methylesterification level, controlled by pectin methylesterases (PMEs), can affect cell wall rigidity (Micheli, 2001; Castillejo et al., 2004). PMEs convert the pectin into more rigid pectate gel by de-esterificating the pectin which allows Ca^{2+} to cross-link the acidic pectins (Jarvis, 1984; Carpita and Gibeaut, 1993). Enhanced freezing tolerance due to cell wall rigidity has been reported in *Arabidopsis* where the wall rigidity is caused by elevated *AtPME41* activity (Qu et al., 2011). *AtPME41* is the *Arabidopsis* ortholog of *Phvul.007G246700*. Qu et al., (2011) suggested the PME activity under chilling stress might be modulated by brassinosteroids. We might speculate that in bean there might be an expression difference of brassinosteriods or PME encoding genes in lodging versus non-lodging genotypes.

The semi-dwarf phenotype in grasses has been favored since the Green Revolution because it results in higher yield and resistance to lodging (Khush, 2001; Van Camp, 2005). Mutations related to brassinisteroid and gibberellin metabolism and regulation can induce the semi-dwarf -lodging resistance phenotype (Szekeres et al., 1996; Li et al., 1996; Clouse, 1996; Choe, 1998, 1999; Noguchi, 1999; Chono et al., 2003; Hong et al., 2003; Martí et al., 2006; Nakamura et al., 2006; Nole-Wilson et al., 2010; Ordonio et al., 2014). *Phvul.007G12200* at

Pv07/45 Mb encodes a *GA2ox6* which inactivates the bioactive GAs and their precursors in *Arabidopsis* (Seo et al., 2006). Plants overexpressing *AtGA2ox6* show GA deficient phenotypes such as dwarfism and decreased apical dominance. *AGL15* is a MADS-box transcription factor that can induce *AtGA2ox6* expression (Wang et al., 2004; Lo et al., 2008). The significant marker in this region, m10611, is not located in the gene, but this gene is the closest gene to this marker. Interestingly, another *AtGA2ox6* homolog *(Phvul.001G215000*) resides close to the Pv01/47.93 lodging GWAS peak.

The 46 Mb region is not only a major peak in lodging and canopy height but also shows a considerably small p-value for growth habit when determinate individuals are excluded. Moreover, when a SNP in this region is used as a cofactor in MLMM analysis, it controls for all the variation on Pv07 for growth habit, lodging and canopy height. This might imply that the 46 Mb region on Pv07 plays a role in these interrelated traits.

Seed weight and seed yield

Determining seed weight gene candidates is more complicated because of the LD that exists across the significant regions. The peaks on the two arms of Pv10 are in LD with $r^2 = 0.8$ in the null model. However, after controlling for population structure using the PC matrix, LD is reduced to 0.2. Prior to controlling for population structure, an inter-chromosomal LD of 0.7 existed between Pv06 and Pv10, but after controlling for population structure it is reduced to almost zero and the regions remain significant in both the null and the 7PC model. This might imply that the peaks are not spurious results due to inter and intra-chromosomal LD.

Based on studies in *Arabidopsis*, seed size and weight are determined by endosperm, embryo and integument growth (Berger et al., 2006; Zhou et al., 2009). A candidate gene on Pv10, *Phvul.010G017600*, is homologous to *Arabidopsis ALPHA-AMYLASE LIKE GENE*

(*AMY1*). *AMY1* induction in globular stage chalaza endosperm is associated with the onset of suspensor and endosperm programmed cell death (PCD) and early nutrient mobilization to nourish the growing embryo. The starch content in the endosperm decreases rapidly from the heart to cotyledon stage. Thus, it has been suggested that alpha-amylase might play a role in starch degradation after cell death (Mansfield & Briarty, 1994; Sreenivasulu & Wobus, 2013). This gene has the highest expression level in flower buds and green mature pods in bean (Schmutz et al., 2014). The candidate gene on the other arm of Pv10 is homologous to *Arabidopsis shrunken seed1* (*SSE1)* which encodes a peroxisome assembly factor that affects protein trafficking and the seed storage protein profile. *Arabidopsis* mutants of this gene have wrinkled seeds, contain less protein and oil bodies but have higher starch content. Moreover, the embryo remains smaller in the *sse1* mutants during seed filling and at maturity. Reduced reserve availability and growth deficiencies (smaller cotyledon areas and thinner hypocotyls), may inhibit embryo expansion (Lin, 1999; Lin, Cluette-Brown, & Goodman, 2004). These all will affect seed size and ultimately yield.

Phvul.003G232600 in the Pv03/45.34 Mb GAWS peak is an ortholog of *Arabidopsis AINTEGUMENTA* (*ANT*). *ANT,* an AP2 transcription factor, controls the size of many organs such as leaf, floral organs and seeds by affecting cell number rather than cell size (Elliott, 1996; Klucher, Chow, Reiser, & Fischer, 1996; Krizek, 1999). The gene was specifically evaluated for its effect on the seed size in *M. trancatula* by overexpressing it using a seed specific promoter (BäUmlein et al., 1991; Zakharov et al., 2004). In a greenhouse environment, these transgenic plants yielded larger seeds without any negative pleiotropic effects (Confalonieri et al., 2014).

An interesting candidate gene, *Phvul.006G069300* on Pv06, is homologous to *Arabidopsis ASN1*, a glutamine-dependent asparagine synthase 1. This gene is also a candidate

for seed yield and is part of the domestication event in the Mesoamerican gene pool (Schmutz et al., 2014). *ASN1* transfers an amino group from glutamine to aspartate to produce glutamate and asparagine, two key amino acids required for seed protein synthesis which in turn affect seed weight (Sanders et al., 2009). *Arabidopsis* transgenic plants overexpressing *ASN1* show increased nitrogen content, higher seed-soluble protein content, and a slight increase in their seed weight, These changes could partially be due to higher asparagine supply to the sink (Lam et al., 2003).

Another Pv06 seed weight candidate gene is *Phvul.006G77400*, an ortholog of *Arabidopsis AGL80/FEM111*. *AGL80* is a MADS box transcription factor that regulates central cell differentiation that leads to endosperm cells. It may control genes involved in endosperm development and/or viability. A functioning endosperm is necessary to provide nutrition to the embryo (Portereiko et al., 2006). The properties of endosperm extends a homeostatic environment for embryo development (Melkus et al., 2009). Indeed, the endosperm engulfs the embryo and is surrounded by ovule integument. The coordinated development of the endosperm, embryo and integument will determine the size of the mature seed. Studies on *Arabidopsis* and maize indicate that the endosperm plays a crucial role in determining seed size (Kermicle & Alleman, 1990; Lin, 1984; Scott, Spielman, Bailey, & Dickinson, 1998). The QTL SW6.6, genetically positioned near the *BM187* marker in common bean (Checa and Blair, 2012), is physically located at 20.2 Mb on Pv06, very close to our candidate *ASN1* and *AGL80*genes.

Phvul.007G088200 on Pv07, a sucrose-proton symporter 2, is homologous to *ATSUC2/SUT1* (*AT1G22710*), *VfSUT1* (NCBI accession number Z93774) and *BnA7.SUT1* (NCBI accession number AY065839). Due to lack of plasmodesmata between the maternal and filial cells, assimilated carbon is transported to the seed as sucrose through the apoplast. In *Vicia*

faba, cotyledon transfer cells specifically express *VfSUT1* which leads to sucrose accumulation. In turn, high levels of sucrose induces maturation (Weber et al., 1998; Borisjuk et al., 2002). In *Brassica napus*, *BnA7.SUT1* alleles are associated with seed yield traits. The *BnA7.SUT1.a* allele is linked with higher seed weight, and the *BnA7.SUT1.b* allele is associated with increased seed yield (Li et al., 2011).

Seed weight, size, and quantity determines the final yield. Thus, it is not unexpected to find similar genomic regions affecting these traits. *ASN1* is one example; we found another candidate gene in the same region of *ASN1* for seed yield in Michigan: *Phvul.006G066900* encodes an acetyl-CoA carboxylase carboxyl transferase subunit beta. Transgenic tobacco expressing the plastidic subunit of acetyl-CoA carboxylase in chloroplasts, showed an increase in the fatty acid content of leaves (not seeds) resulting in leaf longevity and a two fold increase in the seed yield (Madoka, 2002). Another region that seems to affect yield related traits is located at the end of Pv03. Although we did not find overlapping gene models between seed weight and yield, the end of Pv03 seems to encompass a number of genes affecting the yield components. Three candidate genes are located within 1 Mb at ~50 Mb. One candidate, *CDF1*, is a DOF transcription factor (*Phvul.003G275800*). The members of this family are known to function by regulating and activating the expression of storage protein genes in both monocots and dicots (Vicente-Carbajosa and Carbonero, 2005). *AtCDF1* (*AT5G62430*) and its peanut homolog was found to be expressed during seed development along with other DOF genes (Gaur et al., 2011; Yan, 2012). *Phvul.03G288500* is located 1.2 Mb downstream of *CDF1*.This is an *AINTEGUMENTA-like 6* (*AIL6, PLT3*) homolog, encoding an AP2/ERF type transcription factor and is known to regulate cell proliferation in flowers. A transgenic line over-expressing *AIL6* under the constitutive 35S promoter, produced altered floral organ size, morphology, and

sometimes larger seeds (Krizek & Eaddy, 2012). Less than 1 Mb from *AIL6* is *Phvul.003G289100*, an ortholog of *Arabidopsis PASTICCINO* (*PAS2*) gene. Based on *in silico* analysis of differentially expressed ESTs in *Phaseolus vulgaris*, the bean *PAS2* homolog is only expressed in seed (Abid et al., 2011). This gene is involved in cell division and differentiation. Its null mutant in *Arabidopsis* was lethal to embryo but partial loss of function resulted in decreased production of very long chain fatty acids in general (Bach and Faure, 2010). The Pv03 QTL for yield related traits were previously reported by multiple authors (Blair et al., 2006; Checa & Blair, 2012; Pérez-Vega et al., 2010; Wright & Kelly, 2011). Linares-Ramirez (2013) discovered a yield QTL on Pv03 spanning from 51.72 Mb to 51.86 Mb with the largest effect on seed yield in one of the environments. The *MCCB* gene model (*Phvul.003G291600*) falls in this region and is 242Kb downstream of a significant SNP in our study. *MCCB* encodes a subunit of the 3-methylcrotonyl CoA carboxylase (MCCase) that is located in the mitochondria. Homozygous *mccb* mutants show larger and heavier seeds with an overall lower seed yield compared to wild types (Ding et al., 2012). For the Michigan data, a significant region at Pv03/46.3 Mb, is ~ 200Kb upstream of *Phvul.003G241900*, an APETALA2 (AP2) DNA binding protein. *AP2* is not only a flower homeotic gene but also plays an important role in determining seed size, weight and reserve (such as oil and protein) accumulation in seed. *AP2* affects these traits through the maternal sporophyte and endosperm (Jofuku et al., 2005). A seed yield QTL ranging from 45.5 Mb to 47.8 Mb on Pv03 was also found by Linares-Ramirez (2013).

In addition to Pv03 and Pv06, we observed scattered signals on Pv01 near three candidate genes. *Phvul001G071500* encodes an AMINO ACID PERMEASE 2 (AAP2) protein, a phloem loader for amino acids (Hirner et al., 1998). *AAP2* plays a role in nitrogen, carbon, and storage protein transport from source to sink. It is expressed in the stem phloem and silique vein system.

Arabidopsis aap2 mutant seeds show an increase in the fatty acid content but no change in the carbon levels. Mutants had higher number of branches and siliques per plant and elevated seed yield (Zhang et al., 2010). In the DJ subpopulation we also found other amino acid transporters such as *AAP1*, *AAP6*, and *AAP8* clustered near each other on Pv01. *AAP8* is another phloem loader for amino acids (Okumoto et al., 2002).The GUS (promoter-β-glucoronidase) expression pattern for *AAP8* indicated that it is located in the veins of young flowers, peduncles, siliques and very young seeds (Schmidt et al., 2007). *AAP6* on the other hand is a xylem parenchyma amino acid transporter (Okumoto et al., 2002). *AAP2* and *AAP6* differ in their affinity for various amino acids (Fischer et al., 2002). *AAP6* and *AAP8* are the only members of this family capable of transporting aspartate. *AAP1* regulates amino acid trafficking into the developing embryo in *Arabidopsis*. It has a different expression pattern and is expressed in the endosperm and cotyledons instead of the vascular system (Hirner et al., 1998). In the *aap1 Arabidopsis* mutant, the total nitrogen and carbon content decreased in the seed while the free amino acid levels increased in the seed coat/endosperm. The fatty acid content remained the same but the seed storage content decreased. This change in the seed storage protein content was also observed in *Vicia narbonensis* mutants for this gene. The *Arabidopsis* mutants showed lower seed weight, total silique and seed number. In brief, *AAP1*controls the amino acid uptake into the embryo and consequently the nitrogen availability which affects the seed weight and yield (Sanders et al., 2009). In *Arabidopsis*, *AAP1*, *AAP6*, and *AAP8* are all paralogous, and *AAP1* and *AAP8* are the result of a regional duplication on chromosome one but have diverged in their function (Okumoto et al., 2002). Interestingly in common bean, *AAP8* and *AAP1* are tandem duplicates on Pv01/~11 Mb, and only 170Kb upstream of *AAP6* which might suggest this regional duplication has been conserved in *P. vulgaris*. Moreover, significant markers next to *AAP1* (m28775/Pv01,

11.1 Mb) and *AAP2* (m29921/Pv01, 9.4 Mb) are in LD of 0.74 in DJ subpopulation (EMMA model) and 0.62 in the MDP population (7PC model). This might suggest that functional dependency or coordination between genes could result in intra-chromosomal LD.

Another candidate gene on Pv01 is *FUSCA3* (*FUS3*) which is involved in seed maturation. *FUS3* is part of B3-domain family transcription factors (Luerssen et al., 1998), and the *fus3* mutant shows reduced levels of storage compounds, higher anthocyanin, and decreased tolerance to desiccation (To et al., 2006).

Other candidates

It is important to be cautious when interpreting GWAS data. Candidate genes need to be validated in different populations and eventually functional analyses needs to be performed. GWAS analyses can produce both false positive and false negatives. False negatives might not only be due to the nature of regression analysis but also our choice of selecting the significant cutoff value to control for experiment-wide error rate. Such adjustment helps to decrease the false positives but might also increase false negatives. There are different ways to determine the significance level in GWAS. The threshold we used (0.1 percentile tail of the empirical distribution of p-values) is stringent which is generally favorable but does this mean if a peak does not pass this criteria it has no importance? The study of Atwell et al., 2010, based on *a priori* information, demonstrates that not all the SNPs close to known genes possess the highest rank in that population. In fact, since we used *a priori* knowledge while evaluating the 200Kb surrounding significant GWAS peaks for candidate genes, some candidate genes that might be functionally significant, fall in the one percentile tail of the empirical distribution of p-values which is less stringent than the 0.1 percentile cutoff that we originally used. These candidate

gene models are highlighted in gray in Table A.1 and are described below because the interpretation of the data should not be limited to statistical or arbitrary cutoffs.

Phvul.001G071900, the *Arabidopsis SW13C* ortholog, is a flowering time candidate gene on Pv01. *SWI3C* plays an important role in regulating gibberellin biosynthesis which is responsible for flowering time and many other basic functions of the plant. The lack of SWI3C leads to defects in GA signaling via *Gibberellin-Insensitive Dwarf 1* (*GID1*). It also interacts with other DELLA proteins such as RGA like2 (RGL2), RGL3 and SPY (Sarnowska et al., 2013). Another such candidate gene is *Phvul.011G082200,* annotated as *BLH8*/*PNF* on Pv011 which is ~67Kb away from a SNP with p-value of 8.34E-06. In *Arabidopsis* this gene is necessary for the shoot apical meristem (SAM) to respond to floral induction cues to start reproductive development (Smith et al., 2004).

For seed weight, the Pv03/~48 Mb GWAS peak is prominent in the DJ subpopulation (across all the locations and in ND, CO, and NE). Near this region resides *Phvul.003G253100*, an ortholog of *Arabidopsis DWF4* (*AT3G50660*). *DWF4*, a member of *CYP90B* gene family, encodes a 22α-hydroxylase which acts as a rate limiting factor in brassinisteroid (BR) biosynthesis. Jiang et al. (2013) showed that homozygous *dwf4* mutants produces seeds that are 17% lighter than those of wild types. BR treatment after blocking the *DWF4* gene with triazole, increased the expression of some of the genes controlling the seed size such as *SHB1, IKU1, MINI3, IKU2, HSF15* and *KLU*. It was concluded that BR increases seed size by enhancing the expression of many genes that increase seed size and repressing genes that negatively regulate seed size. A 2 Mb long GWAS peak at Pv08/43.3 Mb falls in a domestication sweep window (Schmutz et al., 2014). Nitrate reductase (*Phvul.008G168000*) is located in this region and has been genetically mapped to the SW8.1 seed weight QTL in common bean (Pérez-Vega et al.,

2010). We found peaks that pass the 0.1 percentile tail at the beginning and the end of Pv08 in different locations and subpopulation but currently have not found any candidate genes in these regions. Similarly, Linares-Ramirez (2013) found a seed weight QTL at the beginning of Pv08 flanked by markers at 0.86 Mb and 1.47 Mb. Pv08 has extended intra-chromosomal LD that makes the interpretation and candidate gene identification difficult. For example Linares-Ramirez (2013), found a seed weight QTL on Pv08 whose flanking markers are 44.42 Mb far apart, each close to one of the pericentromeric region boundaries (Schmutz et al., 2014). Many QTL analysis in common bean have located seed related QTL on the same chromosomes we discovered with this GWAS analysis. (Beattie et al., 2003; Blair et al., 2006; Blair et al., 2012b; Linares-Ramirez, 2013; Pérez-Vega et al., 2010; Schneider et al., 1997; Wright & Kelly, 2011). However, due to the lack of sequence data for the markers, most of these cannot be physically mapped on the genome and collocated with specific candidate genes.

For seed yield, a SNP at the Pv07/20.2 Mb GWAS peak, is significant at 0.1 percentile in the DJ subpopulation but it is 285.8Kb upstream of *AMT1;2* (*Phvul.007G120200*) beyond our 200kb cutoff for candidate gene selection. Markers within the 200kb region fall in the less stringent one percentile tail cutoff. *AMT1* is responsible for ammonium (NH4+) uptake in roots. Overexpression of this gene in rice enhanced the expression of genes involved in the nitrogen assimilation which led to higher nitrogen assimilates, starch, and sugar, and increased seed yield by 20% (Ranathunge et al., 2014). Another similar example is the *Phvul.003G285600*, an ortholog of *Arabidopsis IKU1*, a VQ motif-containing protein (Wang et al., 2010). *IKU1* and *IKU2* are known to control the seed size in *Arabidopsis* by affecting the endosperm and integument development (Garcia et al., 2003).

References

Abid, G., Muhovski, Y., Jacquemin, J.-M., Mingeot, D., Sassi, K., Toussaint, A., and Baudoin, J.-P. (2011). In silico identification and characterization of putative differentially expressed genes involved in common bean (*Phaseolus vulgaris* L.) seed development. *Plant Cell Tiss. Org.* 107(2): 341–353.

Appels, R., Barrero, R., and Bellgard, M. (2013). Advances in biotechnology and informatics to link variation in the genome to phenotypes in plants and animals. *Funct. Integr. Genomics* 13(1): 1–9.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., Jiang, R., Muliyati, N.W., Zhang, X., Amer, M.A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J.R., Faure, N., Kniskern, J.M., Jones, J.D.G., Michael, T., Nemri, A., Roux, F., Salt, D.E., Tang, C., Todesco, M., Traw, M.B., Weigel, D., Marjoram, P., Borevitz, J.O., Bergelson, J., and Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* 465(7298): 627–31.

Aulchenko, Y.S., Ripke, S., Isaacs, A., and van Duijn, C.M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23(10): 1294–6.

Aung, B. (2014). Effects of microRNA156 on flowering time and plant architecture in medicago sativa. *Univ. West. Ontario - Electron. Thesis Diss. Repos.*

Avila-Ospina, L., Moison, M., Yoshimoto, K., and Masclaux-Daubresse, C. (2014). Autophagy, plant senescence, and nutrient recycling. *J. Exp. Bot.* 65(14): 3799–811.

Bach, L., and Faure, J.-D. (2010). Role of very-long-chain fatty acids in plant development, when chain length does matter. *C. R. Biol.* 333(4): 361–70.

Balasubramanian, S., Schwartz, C., Singh, A., Warthmann, N., Kim, M.C., Maloof, J.N., Loudet, O., Trainer, G.T., Dabi, T., Borevitz, J.O., Chory, J., and Weigel, D. (2009). QTL mapping in new Arabidopsis thaliana advanced intercross-recombinant inbred lines. *PLoS One* 4(2): e4318.

Bangerth, F. (1994). Response of cytokinin concentration in the xylem exudate of bean (*Phaseolus vulgaris* L.) plants to decapitation and auxin treatment, and relationship to apical dominance. *Planta* 194(3).

BäUmlein, H., Boerjan, W., Nagy, I., Bassfüner, R., Van Montagu, M., Inzé, D., and Wobus, U. (1991). A novel seed protein gene from Vicia faba is developmentally regulated in transgenic tobacco and Arabidopsis plants. *Mol. Gen. Genet. MGG* 225(3): 459–467.

Beattie, A.D., Larsen, J., Michaels, T.E., and Pauls, K.P. (2003). Mapping quantitative trait loci for a common bean (*Phaseolus vulgaris* L.) ideotype. *Genome* 46(3): 411–22.

Beebe, S., Skroch, P.W., Tohme, J., Duque, M.C., Pedraza, F., and Nienhuis, J. (2000). Structure of genetic diversity among common bean landraces of Middle American origin based on correspondence analysis of RAPD. *Crop Sci.* 40(1): 264.

Berger, F., Grini, P.E., and Schnittger, A. (2006). Endosperm: an integrator of seed growth and development. *Curr. Opin. Plant Biol.* 9(6): 664–70.

Blair, M.W., Díaz, L.M., Buendía, H.F., and Duque, M.C. (2009). Genetic diversity, seed size associations and population structure of a core collection of common beans (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* 119(6): 955–72.

Blair, M.W., Galeano, C.H., Tovar, E., Muñoz Torres, M.C., Castrillón, A.V., Beebe, S.E., and Rao, I.M. (2012a). Development of a Mesoamerican intra-genepool genetic map for quantitative trait loci detection in a drought tolerant × susceptible common bean (*Phaseolus vulgaris* L.) cross. *Mol. Breed.* 29(1): 71–88.

Blair, M.W., Iriarte, G., and Beebe, S. (2006). QTL analysis of yield traits in an advanced backcross population derived from a cultivated Andean x wild common bean (*Phaseolus vulgaris* L.) cross. *Theor. Appl. Genet.* 112(6): 1149–63.

Borisjuk, L., Walenta, S., Rolletschek, H., Mueller-Klieser, W., Wobus, U., and Weber, H. (2002). Spatial analysis of plant metabolism: Sucrose imaging within *Vicia faba* cotyledons reveals specific developmental patterns. *Plant J.* 29(4): 521–530.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23: 2633–2635.

Broughton, W.J., Hernández, G., Blair, M., Beebe, S., Gepts, P., and Vanderleyden, J. (2003). Beans (*Phaseolus* spp.) – model food legumes. *Plant Soil* 252(1): 55–128.

Van Camp, W. (2005). Yield enhancement genes: seeds for growth. *Curr. Opin. Biotechnol.* 16(2): 147–53.

Carpita, N.C., and Gibeaut, D.M. (1993). Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *Plant J.* 3(1): 1–30.

Castillejo, C., de la Fuente, J.I., Iannetta, P., Botella, M.A., and Valpuesta, V. (2004). Pectin esterase gene family in strawberry fruit: study of FaPE1, a ripening-specific isoform. *J. Exp. Bot.* 55(398): 909–18.

Chardin, C., Girin, T., Roudier, F., Meyer, C., and Krapp, A. (2014). The plant RWP-RK transcription factors: Key regulators of nitrogen responses and of gametophyte development. *J. Exp. Bot.*: eru261–.

Checa, O.E., and Blair, M.W. (2012). Inheritance of yield-related traits in climbing beans (L.). *Crop Sci.* 52(5): 1998.

Chen, J., and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 95: 759–771.

Choe, S. (1998). The DWF4 gene of Arabidopsis encodes a cytochrome P450 that mediates multiple 22alpha-Hydroxylation steps in brassinosteroid biosynthesis. *Plant Cell Online* 10(2): 231–244.

Choe, S. (1999). The Arabidopsis dw f 7/ste1 mutant is defective in the Delta7 Sterol C-5 Desaturation step leading to brassinosteroid biosynthesis. *Plant Cell Online* 11(2): 207–222.

Chono, M., Honda, I., Zeniya, H., Yoneyama, K., Saisho, D., Takeda, K., Takatsuto, S., Hoshino, T., and Watanabe, Y. (2003). A semidwarf phenotype of barley uzu results from a nucleotide substitution in the gene encoding a putative brassinosteroid receptor. *Plant Physiol.* 133(3): 1209–19.

Cline, M.G. (1991). Apical dominance. *Bot. Rev.* 57(4): 318–358.

Clouse, S. (1996). A brassinosteroid-insensitive mutant in Arabidopsis thaliana exhibits multiple defects in growth and development. *Plant Physiol.* 111(3): 671–678.

Clouse, S.D. (2002). Brassinosteroid Signal Transduction. *Mol. Cell* 10(5): 973–982.

Colasanti, J., Tremblay, R., Wong, A.Y.M., Coneva, V., Kozaki, A., and Mable, B.K. (2006). The maize INDETERMINATE1 flowering time regulator defines a highly conserved zinc finger protein family in higher plants. *BMC Genomics* 7(1): 158.

Comadran, J., Ramsay, L., MacKenzie, K., Hayes, P., Close, T.J., Muehlbauer, G., Stein, N., and Waugh, R. (2011). Patterns of polymorphism and linkage disequilibrium in cultivated barley. *Theor. Appl. Genet.* 122(3): 523–31.

Confalonieri, M., Carelli, M., Galimberti, V., Macovei, A., Panara, F., Biggiogera, M., Scotti, C., and Calderini, O. (2014). Seed-Specific Expression of AINTEGUMENTA in Medicago truncatula Led to the Production of Larger Seeds and Improved Seed Germination. *Plant Mol. Biol. Report.* 32(5): 957–970.

Conner, J., and Liu, Z. (2000). LEUNIG, a putative transcriptional corepressor that regulates AGAMOUS expression during flower development. *Proc. Natl. Acad. Sci. U. S. A.* 97(23): 12902–7.

Coyne, D.P., and Schuster, M.L. (1974). Inheritance and linkage relations of reaction to *Xanthomonas phaseoli* (E. F. Smith) Dowson (common blight), stage of plant development and plant habit in *Phaseolus vulgaris* L. *Euphytica* 23(2): 195–204.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., and Blaxter, M.L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12(7): 499–510.

Deprost, D., Yao, L., Sormani, R., Moreau, M., Leterreux, G., Nicolaï, M., Bedu, M., Robaglia, C., and Meyer, C. (2007). The Arabidopsis TOR kinase links plant growth, yield, stress resistance and mRNA translation. *EMBO Rep.* 8(9): 864–70.

Deschamps, S., and Campbell, M.A. (2009). Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Mol. Breed.* 25(4): 553–570.

Desclos, M., Etienne, P., Coquet, L., Jouenne, T., Bonnefoy, J., Segura, R., Reze, S., Ourry, A., and Avice, J.-C. (2009). A combined 15N tracing/proteomics study in *Brassica napus* reveals the chronology of proteomics events associated with N remobilisation during leaf senescence induced by nitrate limitation or starvation. *Proteomics* 9(13): 3580–608.

Díaz, L.M., and Blair, M.W. (2006). Race structure within the Mesoamerican gene pool of common bean (*Phaseolus vulgaris* L.) as determined by microsatellite markers. *Theor. Appl. Genet.* 114(1): 143–54.

Ding, G., Che, P., Ilarslan, H., Wurtele, E.S., and Nikolau, B.J. (2012). Genetic dissection of methylcrotonyl CoA carboxylase indicates a complex role for mitochondrial leucine catabolism during seed development and germination. *Plant J.* 70(4): 562–77.

Elliott, R.C. (1996). AINTEGUMENTA, an APETALA2-like gene of Arabidopsis with pleiotropic roles in ovule development and floral organ growth. *Plant Cell Online* 8(2): 155–168.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6(5): e19379.

Fischer, W.-N., Loo, D.D.F., Koch, . Wolfgang, Ludewig, U., Boorer, K.J., Tegeder, M., Rentsch, D., Wright, E.M., and Frommer, W.B. (2002). Low and high affinity amino acid H + -cotransporters for cellular import of neutral and charged amino acids. *Plant J.* 29(6): 717–731.

Franks, R.G., Wang, C., Levin, J.Z., and Liu, Z. (2002). SEUSS, a member of a novel family of plant regulatory proteins, represses floral homeotic gene expression with LEUNIG. *Development* 129(1): 253–263.

Freyre, R., Ríos, R., Guzmán, L., Debouck, D.G., and Gepts, P. (1996). Ecogeographic distribution of *Phaseolus* spp. (*Fabaceae*) in Bolivia. *Econ. Bot.* 50(2): 195–215.

García, E. (1997). Morphological and agronomic traits of a wild population and an improved cultivar of common bean (*Phaseolus vulgaris* L.). *Ann. Bot.* 79(2): 207–213.

Garcia, D., Saingery, V., Chambrier, P., Mayer, U., Jürgens, G., and Berger, F. (2003). Arabidopsis haiku mutants reveal new controls of seed size by endosperm. *Plant Physiol.* 131(4): 1661–70.

Gaur, V.S., Singh, U.S., and Kumar, A. (2011). Transcriptional profiling and in silico analysis of Dof transcription factor gene family for understanding their regulation during seed development of rice *Oryza sativa* L. *Mol. Biol. Rep.* 38(4): 2827–48.

Gepts, P., and Bliss, F.A. (1986). Phaseolin variability among wild and cultivated common beans (*Phaseolus vulgaris*) from Colombia. *Econ. Bot.* 40(4): 469–478.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40 (Database issue): D1178–86.

Gore, M.A., Chia, J.-M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurwitz, B.L., Peiffer, J.A., McMullen, M.D., Grills, G.S., Ross-Ibarra, J., Ware, D.H., and Buckler, E.S. (2009). A first-generation haplotype map of maize. *Science* 326(5956): 1115–7.

Gregis, V., Sessa, A., Colombo, L., and Kater, M.M. (2006). AGL24, SHORT VEGETATIVE PHASE, and APETALA1 redundantly control AGAMOUS during early stages of flower development in Arabidopsis. *Plant Cell* 18(6): 1373–82.

Hartl, D.L., and Clark, A.G. (2007). Principles of population genetics. Sinauer Associates.

Hedrick, U.P., Tapley, W.T., Eseltine, G.P. van., and Enzie, W.D. (1931). The vegetables of New York. Vol. 1, Part II. Beans of New York.

Hirner, B., Fischer, W.N., Rentsch, D., Kwart, M., and Frommer, W.B. (1998). Developmental control of H+/amino acid permease gene expression during seed development of Arabidopsis. *Plant J.* 14(5): 535–544.

Hong, Z., Ueguchi-Tanaka, M., Umemura, K., Uozu, S., Fujioka, S., Takatsuto, S., Yoshida, S., Ashikari, M., Kitano, H., and Matsuoka, M. (2003). A rice brassinosteroid-deficient mutant, ebisu dwarf (d2), is caused by a loss of function of a new member of cytochrome P450. *Plant Cell* 15(12): 2900–10.

Ten Hove, C.A., Bochdanovits, Z., Jansweijer, V.M.A., Koning, F.G., Berke, L., Sanchez-Perez, G.F., Scheres, B., and Heidstra, R. (2011). Probing the roles of LRR RLK genes in Arabidopsis thaliana roots using a custom T-DNA insertion set. *Plant Mol. Biol.* 76(1-2): 69–83.

Hyten, D.L., Song, Q., Fickus, E.W., Quigley, C. V, Lim, J.-S., Choi, I.-Y., Hwang, E.-Y., Pastor-Corrales, M., and Cregan, P.B. (2010). High-throughput SNP discovery and assay development in common bean. *BMC Genomics* 11(1): 475.

Jarvis, M.C. (1984). Structure and properties of pectin gels in plant cell walls. *Plant, Cell Environ.* 7(3): 153–164.

Jégu, T., Latrasse, D., Delarue, M., Hirt, H., Domenichini, S., Ariel, F., Crespi, M., Bergounioux, C., Raynaud, C., and Benhamed, M. (2014). The BAF60 subunit of the SWI/SNF chromatin-remodeling complex directly controls the formation of a gene loop at FLOWERING LOCUS C in Arabidopsis. *Plant Cell* 26(2): 538–51.

Jiang, W.-B., Huang, H.-Y., Hu, Y.-W., Zhu, S.-W., Wang, Z.-Y., and Lin, W.-H. (2013). Brassinosteroid regulates seed size and shape in Arabidopsis. *Plant Physiol.* 162(4): 1965–77.

Jiao, Y., Wang, Y., Xue, D., Wang, J., Yan, M., Liu, G., Dong, G., Zeng, D., Lu, Z., Zhu, X., Qian, Q., and Li, J. (2010). Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. *Nat. Genet.* 42(6): 541–4.

Jofuku, K.D., Omidyar, P.K., Gee, Z., and Okamuro, J.K. (2005). Control of seed mass and seed yield by the floral homeotic gene APETALA2. *Proc. Natl. Acad. Sci. U. S. A.* 102(8): 3117–22.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178: 1709–1723.

Kermicle, J.L., and Alleman, M. (1990). Gametic imprinting in maize in relation to the angiosperm life cycle. *Development* 108(Supplement): 9–14.

Khairallah, M.M., Adams, M.W., and Sears, B.B. (1990). Mitochondrial DNA polymorphisms of Malawian bean lines: further evidence for two major gene pools. *Theor. Appl. Genet.* 80(6): 753–61.

Khush, G.S. (2001). Green revolution: the way forward. *Nat. Rev. Genet.* 2(10): 815–22.

Kim, J., and Guan, K.-L. (2011). Amino acid signaling in TOR activation. *Annu. Rev. Biochem.* 80: 1001–32.

Klucher, K.M., Chow, H., Reiser, L., and Fischer, R.L. (1996). The AINTEGUMENTA gene of Arabidopsis required for ovule and female gametophyte development is related to the floral homeotic gene APETALA2. *Plant Cell* 8(2): 137–53.

Koenig, R., and Gepts, P. (1989). Allozyme diversity in wild *Phaseolus vulgaris*: further evidence for two major centers of genetic diversity. *Theor. Appl. Genet.* 78(6): 809–17.

Koinange, E.M.K., and Gepts, P. (1992). Hybrid weakness in wild *Phaseolus vulgaris* L. *J. Hered.* 83(2): 135–139.

Koornneef, M., Hanhart, C.J., and van der Veen, J.H. (1991). A genetic and physiological analysis of late flowering mutants in Arabidopsis thaliana. *Mol. Gen. Genet. MGG* 229(1): 57–66.

Kornegay, J., White, J.W., and de la Cruz, O.O. (1992). Growth habit and gene pool effects on inheritance of yield in common bean. *Euphytica* 62(3): 171–180.

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9(1): 29.

Krizek, B.A. (1999). Ectopic expression of AINTEGUMENTA in Arabidopsis plants results in increased growth of floral organs. *Dev. Genet.* 25(3): 224–36.

Krizek, B.A., and Eaddy, M. (2012). AINTEGUMENTA-LIKE6 regulates cellular differentiation in flowers. *Plant Mol. Biol.* 78(3): 199–209.

Kwak, M., and Gepts, P. (2009). Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., *Fabaceae*). *Theor. Appl. Genet.* 118(5): 979–92.

Lam, H.-M., Wong, P., Chan, H.-K., Yam, K.-M., Chen, L., Chow, C.-M., and Coruzzi, G.M. (2003). Overexpression of the *ASN1* gene enhances nitrogen status in seeds of Arabidopsis. *Plant Physiol.* 132(2): 926–35.

Leakey, C.L.A. (1988). Genetic Resources of Phaseolus Beans. p. 245–327. *In* Gepts, P. (ed.), Genetic resources of *Phaseolus* beans. Current Plant Science and Biotechnology in Agriculture. Springer Netherlands, Dordrecht.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li, F., Ma, C., Wang, X., Gao, C., Zhang, J., Wang, Y., Cong, N., Li, X., Wen, J., Yi, B., Shen, J., Tu, J., and Fu, T. (2011). Characterization of Sucrose transporter alleles and their association with seed yield-related traits in Brassica napus L. *BMC Plant Biol.* 11(1): 168.

Li, J., Nagpal, P., Vitart, V., McMorris, T.C., and Chory, J. (1996). A Role for Brassinosteroids in Light-Dependent Development of Arabidopsis. *Science (80-.).* 272(5260): 398–401.

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., Han, Y., Chai, Y., Guo, T., Yang, N., Liu, J., Warburton, M.L., Cheng, Y., Hao, X., Zhang, P., Zhao, J., Liu, Y., Wang, G., Li, J., and Yan, J. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45(1): 43–50.

Li, X., Tan, L., Zhu, Z., Huang, H., Liu, Y., Hu, S., and Sun, C. (2009). Patterns of nucleotide diversity in wild and cultivated rice. *Plant Syst. Evol.* 281(1-4): 97–106.

Lin, B.-Y. (1984). Ploidy barrier to endosperm development in maize. *Genetics* 107(1): 103–115.

Lin, Y. (1999). The Pex16p Homolog SSE1 and Storage Organelle Formation in Arabidopsis Seeds. *Science (80-.).* 284(5412): 328–330.

Lin, Y., Cluette-Brown, J.E., and Goodman, H.M. (2004). The peroxisome deficient Arabidopsis mutant sse1 exhibits impaired fatty acid synthesis. *Plant Physiol.* 135(2): 814–27.

Linares-Ramirez, A. (2013). Selection of dry bean genotypes adapted for drought tolerance in the northern Great Plains. : 125.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C., Mauceli, E., Xie, X., Breen, M., Wayne, R.K., Ostrander, E.A., Ponting, C.P., Galibert, F., Smith, D.R., DeJong, P.J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.-W., Cook, A., Cuff, J., Daly, M.J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.-P., Parker, H.G., Pollinger, J.P., Searle, S.M.J., Sutter, N.B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-Zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Graham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Kieu, A.C., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J.-P., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nguyen, T., Nicol, R., Norbu, N., Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S., Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov,

P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone, S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A., and Lander, E.S. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069): 803–19.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., and Zhang, Z. (2012). GAPIT: Genome association and prediction integrated tool. *Bioinformatics* 28: 2397–2399.

Liu, Z., and Karmarkar, V. (2008). Groucho/Tup1 family co-repressors in plant development. *Trends Plant Sci.* 13(3): 137–44.

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

Lo, S.-F., Yang, S.-Y., Chen, K.-T., Hsing, Y.-I., Zeevaart, J.A.D., Chen, L.-J., and Yu, S.-M. (2008). A novel class of gibberellin 2-oxidases control semidwarfism, tillering, and root development in rice. *Plant Cell* 20(10): 2603–18.

Lohman, K.N., Gan, S., John, M.C., and Amasino, R.M. (1994). Molecular analysis of natural leaf senescence in Arabidopsis thaliana. *Physiol. Plant.* 92(2): 322–328.

Luerssen, H., Kirik, V., Herrmann, P., and Misera, S. (1998). *FUSCA3* encodes a protein with a conserved VP1/ABI3-like B3 domain which is of functional importance for the regulation of seed maturation inArabidopsis thaliana. *Plant J.* 15(6): 755–764.

MacKintosh, C. (1992). Regulation of spinach-leaf nitrate reductase by reversible phosphorylation. *Biochim. Biophys. Acta - Mol. Cell Res.* 1137(1): 121–126.

Madoka, Y. (2002). Chloroplast transformation with modified accd operon increases Acetyl-CoA Carboxylase and causes extension of leaf longevity and increase in seed yield in tobacco. *Plant Cell Physiol.* 43(12): 1518–1525.

Mamidi, S., Chikara, S., Goos, R.J., Hyten, D.L., Annam, D., Moghaddam, S.M., Lee, R.K., Cregan, P.B., and McClean, P.E. (2011a). Genome-wide association analysis identifies candidate genes associated with iron deficiency chlorosis in soybean. *Plant Genome J.* 4(3): 154.

Mamidi, S., Lee, R.K., Goos, J.R., and McClean, P.E. (2014). Genome-wide association studies identifies seven major regions responsible for iron deficiency chlorosis in soybean (*Glycine max*). (SK Parida, Ed.). *PLoS One* 9(9): e107469.

Mamidi, S., Rossi, M., Annam, D., Moghaddam, S., Lee, R., Papa, R., and McClean, P. (2011b). Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. *Funct. Plant Biol.* 38(12): 953.

Mamidi, S., Rossi, M., Moghaddam, S.M., Annam, D., Lee, R., Papa, R., and McClean, P.E. (2013). Demographic factors shaped diversity in the two gene pools of wild common bean Phaseolus vulgaris L. *Heredity (Edinb).* 110(3): 267–76.

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., and Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity (Edinb).* 108(3): 285–91.

Mansfield, S. G., & Briarty, L.G. (1994). Endosperm development. p. 385–397. *In* Bowman, J. (ed.), Arabidopsis, An Atlas of Morphology and Development. Springer.

Martí, E., Gisbert, C., Bishop, G.J., Dixon, M.S., and García-Martínez, J.L. (2006). Genetic and physiological characterization of tomato cv. Micro-Tom. *J. Exp. Bot.* 57(9): 2037–47.

Martínez, D.E., Costa, M.L., and Guiamet, J.J. (2008). Senescence-associated degradation of chloroplast proteins inside and outside the organelle. *Plant Biol. (Stuttg).* 10 Suppl 1: 15–22.

Melkus, G., Rolletschek, H., Radchuk, R., Fuchs, J., Rutten, T., Wobus, U., Altmann, T., Jakob, P., and Borisjuk, L. (2009). The metabolic role of the legume endosperm: a noninvasive imaging study. *Plant Physiol.* 151(3): 1139–54.

Mensack, M.M., Fitzgerald, V.K., Ryan, E.P., Lewis, M.R., Thompson, H.J., and Brick, M.A. (2010). Evaluation of diversity among common beans (*Phaseolus vulgaris* L.) from two centers of domestication using "omics" technologies. *BMC Genomics* 11(1): 686.

Miceli, F., Crafts-Brandner, S.J., and Egli, D.B. (1995). Physical restriction of pod growth alters development of soybean plants. *Crop Sci.* 35(4): 1080.

Michaels, S.D., Bezerra, I.C., and Amasino, R.M. (2004). FRIGIDA-related genes are required for the winter-annual habit in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 101(9): 3281–5.

Micheli, F. (2001). Pectin methylesterases: cell wall enzymes with important roles in plant physiology. *Trends Plant Sci.* 6(9): 414–419.

Mizukami, Y., and Ma, H. (1997). Determination of Arabidopsis floral meristem identity by AGAMOUS. *Plant Cell* 9(3): 393–408.

Nakamura, A., Fujioka, S., Sunohara, H., Kamiya, N., Hong, Z., Inukai, Y., Miura, K., Takatsuto, S., Yoshida, S., Ueguchi-Tanaka, M., Hasegawa, Y., Kitano, H., and Matsuoka, M. (2006). The role of *OsBRI1* and its homologous genes, *OsBRL1* and *OsBRL3*, in rice. *Plant Physiol.* 140(2): 580–90.

Noguchi, T. (1999). Arabidopsis det2 is defective in the conversion of (24R)-24-Methylcholest-4-En-3-One to (24R)-24-Methyl-5alpha -Cholestan-3-One in brassinosteroid biosynthesis. *Plant Physiol.* 120(3): 833–840.

Noh, Y.-S., and Amasino, R.M. (1999). Identification of a promoter region responsible for the senescence-specific expression of *SAG12*. *Plant Mol. Biol.* 41(2): 181–194.

Nole-Wilson, S., Rueschhoff, E.E., Bhatti, H., and Franks, R.G. (2010). Synergistic disruptions in seuss cyp85A2 double mutants reveal a role for brassinolide synthesis during gynoecium and ovule development. *BMC Plant Biol.* 10(1): 198.

Noodén, L.D. (1988). Whole plant senescencee. p. 392–439. *In* Noodén, L.D., Leopold, A.C. (eds.), Senescence and aging in plants. Academic Press, San Diego.

Nordström, A., Tarkowski, P., Tarkowska, D., Norbaek, R., Astot, C., Dolezal, K., and Sandberg, G. (2004). Auxin regulation of cytokinin biosynthesis in Arabidopsis thaliana: a factor of potential importance for auxin-cytokinin-regulated development. *Proc. Natl. Acad. Sci. U. S. A.* 101(21): 8039–44.

Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K., and Ohta, H. (2007). ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Res.* 35(Database issue): D863–9.

Okumoto, S., Schmidt, R., Tegeder, M., Fischer, W.N., Rentsch, D., Frommer, W.B., and Koch, W. (2002). High affinity amino acid transporters specifically expressed in xylem parenchyma and developing seeds of Arabidopsis. *J. Biol. Chem.* 277(47): 45338–46.

Ordonio, R.L., Ito, Y., Hatakeyama, A., Ohmae-Shinohara, K., Kasuga, S., Tokunaga, T., Mizuno, H., Kitano, H., Matsuoka, M., and Sazuka, T. (2014). Gibberellin deficiency pleiotropically induces culm bending in sorghum: an insight into sorghum semi-dwarf breeding. *Sci. Rep.* 4: 5287.

Otegui, M.S., Noh, Y.-S., Martínez, D.E., Vila Petroff, M.G., Staehelin, L.A., Amasino, R.M., and Guiamet, J.J. (2005). Senescence-associated vacuoles with intense proteolytic activity develop in leaves of Arabidopsis and soybean. *Plant J.* 41(6): 831–44.

Palni, L.M., Burch, L., and Horgan, R. (1988). The effect of auxin concentration on cytokinin stability and metabolism. *Planta* 174(2): 231–4.

Park, S.J. (1993). Response of bush and upright plant type selections to white mold and seed yield of common beans grown in various row widths in southern Ontario. *Can. J. Plant Sci.* 73(1): 265–272.

Park, M.Y., Wu, G., Gonzalez-Sulser, A., Vaucheret, H., and Poethig, R.S. (2005). Nuclear processing and export of microRNAs in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 102(10): 3691–6.

Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P., Nguyen, B.T., Norris, M.C., Sheehan, J.B., Shen, N., Stern, D., Stokowski, R.P., Thomas, D.J., Trulson, M.O., Vyas, K.R., Frazer, K.A., Fodor, S.P., and Cox, D.R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294(5547): 1719–23.

Payne, T., Johnson, S.D., and Koltunow, A.M. (2004). KNUCKLES (KNU) encodes a C2H2 zinc-finger protein that regulates development of basal pattern elements of the Arabidopsis gynoecium. *Development* 131(15): 3737–49.

Peng, P., and Li, J. (2003). Brassinosteroid signal transduction: a mix of conservation and novelty. *J. Plant Growth Regul.* 22(4): 298–312.

Pérez-Vega, E., Pañeda, A., Rodríguez-Suárez, C., Campa, A., Giraldez, R., and Ferreira, J.J. (2010). Mapping of QTLs for morpho-agronomic and seed quality traits in a RIL population of common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* 120(7): 1367–80.

Poethig, R.S. (2009). Small RNAs and developmental timing in plants. *Curr. Opin. Genet. Dev.* 19(4): 374–8.

Portereiko, M.F., Lloyd, A., Steffen, J.G., Punwani, J.A., Otsuga, D., and Drews, G.N. (2006). AGL80 is required for central cell and endosperm development in Arabidopsis. *Plant Cell* 18(8): 1862–72.

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

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

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

Qu, T., Liu, R., Wang, W., An, L., Chen, T., Liu, G., and Zhao, Z. (2011). Brassinosteroids regulate pectin methylesterase activity and *AtPME41* expression in Arabidopsis under chilling stress. *Cryobiology* 63(2): 111–7.

Ranathunge, K., El-Kereamy, A., Gidda, S., Bi, Y.-M., and Rothstein, S.J. (2014). AMT1;1 transgenic rice plants with enhanced NH4(+) permeability show superior growth and higher yield under optimal and suboptimal NH4(+) conditions. *J. Exp. Bot.* 65(4): 965–79.

Robbins, M.D., Sim, S.-C., Yang, W., Van Deynze, A., van der Knaap, E., Joobeur, T., and Francis, D.M. (2011). Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato. *J. Exp. Bot.* 62(6): 1831–45.

Rosenberg, N.A. (2003). distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* 4(1): 137–138.

Rosenberg, N.A., Burke, T., Elo, K., Feldman, M.W., Freidlin, P.J., Groenen, M.A., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., Wimmers, K., and Weigend, S. (2001). Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* 159: 699–713.

Rubinelli, P.M., Chuck, G., Li, X., and Meilan, R. (2013). Constitutive expression of the Corngrass1 microRNA in poplar affects plant architecture and stem lignin content and composition. *Biomass and Bioenergy* 54: 312–321.

Sanders, A., Collier, R., Trethewy, A., Gould, G., Sieker, R., and Tegeder, M. (2009). AAP1 regulates import of amino acids into developing Arabidopsis embryos. *Plant J.* 59(4): 540–52.

Sarnowska, E.A., Rolicka, A.T., Bucior, E., Cwiek, P., Tohge, T., Fernie, A.R., Jikumaru, Y., Kamiya, Y., Franzen, R., Schmelzer, E., Porri, A., Sacharowski, S., Gratkowska, D.M., Zugaj, D.L., Taff, A., Zalewska, A., Archacki, R., Davis, S.J., Coupland, G., Koncz, C., Jerzmanowski, A., and Sarnowski, T.J. (2013). DELLA-interacting SWI3C core subunit of switch/sucrose nonfermenting chromatin remodeling complex modulates gibberellin responses and hormonal cross talk in Arabidopsis. *Plant Physiol.* 163(1): 305–17.

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

Schmidt, R., Stransky, H., and Koch, W. (2007). The amino acid permease AAP8 is important for early seed development in Arabidopsis thaliana. *Planta* 226(4): 805–13.

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., Jenkins, J., Shu, S., Song, Q., Chavarro, C., Torres-Torres, M., Geffroy, V., Moghaddam, S. M., Gao, D., Abernathy, B., Barry, K., Blair, M., Brick, M. A., Chovatia, M., Gepts, P., Goodstein, D. M., Gonzales, M., Hellsten, U., Hyten, D.L., Jia, G., Kelly, J. D., Kudrna, D., Lee, R., Richard, M. M. S., Miklas, P. N., Osorno, J M., Rodrigues, J., Thareau, V., Urrea, C. A., Wang, M., Yu, Y., Zhang, M., Wing, R. A., Cregan, P. B., Rokhsar D. S., & Jackson, S. A.,

(2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.*, *46*(7), 707–13.

Schneider, K.A., Rosales-Serna, R., Ibarra-Perez, F., Cazares-Enriquez, B., Acosta-Gallegos, J.A., Ramirez-Vallejo, P., Wassimi, N., and Kelly, J.D. (1997). Improving Common Bean Performance under Drought Stress. *Crop Sci.* 37(1): 43.

Schomburg, F.M. (2001). FPA, a gene involved in floral induction in arabidopsis, encodes a protein containing RNA-recognition motifs. *Plant Cell Online* 13(6): 1427–1436.

Schwartz, H.F., Casciano, D.H., Asenga, J.A., and Wood, D.R. (1987). Field measurement of white mold effects upon dry beans with genetic resistance or upright plant architecture1. *Crop Sci.* 27(4): 699.

Schwarz, S., Grande, A. V, Bujdoso, N., Saedler, H., and Huijser, P. (2008). The microRNA regulated SBP-box genes SPL9 and SPL15 control shoot maturation in Arabidopsis. *Plant Mol. Biol.* 67(1-2): 183–95.

Scott, R., Spielman, M., Bailey, J., and Dickinson, H. (1998). Parent-of-origin effects on seed development in Arabidopsis thaliana. *Development* 125(17): 3329–3341.

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44: 825–830.

Seo, M., Hanada, A., Kuwahara, A., Endo, A., Okamoto, M., Yamauchi, Y., North, H., Marion-Poll, A., Sun, T.-P., Koshiba, T., Kamiya, Y., Yamaguchi, S., and Nambara, E. (2006). Regulation of hormone metabolism in Arabidopsis seeds: phytochrome regulation of abscisic acid metabolism and abscisic acid regulation of gibberellin metabolism. *Plant J.* 48(3): 354–66.

Shikata, M., Koyama, T., Mitsuda, N., and Ohme-Takagi, M. (2009). Arabidopsis SBP-box genes SPL10, SPL11 and SPL2 control morphological change in association with shoot maturation in the reproductive phase. *Plant Cell Physiol.* 50(12): 2133–45.

Shin, J.-H., Blay, S., McNeney, B., and Graham, J. (2006). LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* 16: 1–10.

Singh, S.P. (1989). Patterns of variation in cultivated common bean (*Phaseolus vulgaris*, *Fabaceae*). *Econ. Bot.* 43(1): 39–57.

Singh, S.P., Gepts, P., and Debouck, D.G. (1991). Races of common bean (*Phaseolus vulgaris*, *Fabaceae*). *Econ. Bot.* 45(3): 379–396.

Smith, H.M.S., Campbell, B.C., and Hake, S. (2004). Competence to respond to floral inductive signals requires the homeobox genes *PENNYWISE* and *POUND-FOOLISH*. *Curr. Biol.* 14(9): 812–7.

Sreenivasulu, N., and Wobus, U. (2013). Seed-development programs: a systems biology-based comparison between dicots and monocots. *Annu. Rev. Plant Biol.* 64: 189–217.

Sridhar, V. V, Surendrarao, A., Gonzalez, D., Conlan, R.S., and Liu, Z. (2004). Transcriptional repression of target genes by LEUNIG and SEUSS, two interacting regulatory proteins for Arabidopsis flower development. *Proc. Natl. Acad. Sci. U. S. A.* 101(31): 11494–9.

Sun, G., Zhu, C., Kramer, M.H., Yang, S.-S., Song, W., Piepho, H.-P., and Yu, J. (2010). Variation explained in mixed-model association mapping. *Heredity (Edinb).* 105(4): 333–40.

Szekeres, M., Németh, K., Koncz-Kálmán, Z., Mathur, J., Kauschmann, A., Altmann, T., Rédei, G.P., Nagy, F., Schell, J., and Koncz, C. (1996). Brassinosteroids rescue the deficiency of CYP90, a cytochrome P450, controlling cell elongation and de-etiolation in arabidopsis. *Cell* 85(2): 171–182.

Tang, T., Lu, J., Huang, J., He, J., McCouch, S.R., Shen, Y., Kai, Z., Purugganan, M.D., Shi, S., and Wu, C.-I. (2006). Genomic variation in rice: genesis of highly polymorphic linkage blocks during domestication. (J Doebley, Ed.). *PLoS Genet.* 2(11): e199.

Tar'an, B., Michaels, T.E., and Pauls, K.P. (2002). Genetic mapping of agronomic traits in common bean. *Crop Sci.* 42(2): 544–556.

Team, R. (2013). R development core team. *R A Lang. Environ. Stat. Comput.*

Thummel, C.S., and Chory, J. (2002). Steroid signaling in plants and insects--common themes, different pathways. *Genes Dev.* 16(24): 3113–29.

To, A., Valon, C., Savino, G., Guilleminot, J., Devic, M., Giraudat, J., and Parcy, F. (2006). A network of local and redundant gene regulation governs Arabidopsis seed maturation. *Plant Cell* 18(7): 1642–51.

Varshney, R.K., Nayak, S.N., May, G.D., and Jackson, S.A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* 27(9): 522–30.

Verslues, P.E., Lasky, J.R., Juenger, T.E., Liu, T.-W., and Kumar, M.N. (2014). Genome-wide association mapping combined with reverse genetics identifies new effectors of low water potential-induced proline accumulation in Arabidopsis. *Plant Physiol.* 164(1): 144–59.

Vicente-Carbajosa, J., and Carbonero, P. (2005). Seed maturation: developing an intrusive phase to accomplish a quiescent state. *Int. J. Dev. Biol.* 49(5-6): 645–51.

Wang, H., Caruso, L. V, Downie, A.B., and Perry, S.E. (2004). The embryo MADS domain protein AGAMOUS-Like 15 directly regulates expression of a gene encoding an enzyme involved in gibberellin metabolism. *Plant Cell* 16(5): 1206–19.

Wang, R., Farrona, S., Vincent, C., Joecker, A., Schoof, H., Turck, F., Alonso-Blanco, C., Coupland, G., and Albani, M.C. (2009). PEP1 regulates perennial flowering in Arabis alpina. *Nature* 459(7245): 423–7.

Wang, A., Garcia, D., Zhang, H., Feng, K., Chaudhury, A., Berger, F., Peacock, W.J., Dennis, E.S., and Luo, M. (2010). The VQ motif protein IKU1 regulates endosperm growth and seed size in Arabidopsis. *Plant J.* 63(4): 670–9.

Wang, Z.-Y., and He, J.-X. (2004). Brassinosteroid signal transduction--choices of signals and receptors. *Trends Plant Sci.* 9(2): 91–6.

Weber, H., Heim, U., Golombek, S., Borisjuk, L., Manteuffel, R., and Wobus, U. (1998). Expression of a yeast-derived invertase in developing cotyledons ofVicia narbonensisalters the carbohydrate state and affects storage functions. *Plant J.* 16(2): 163–172.

Wei, Taiyun, and M.T.W. (2013). Package "corrplot." *Statistician* (56): 316–324.

Werner, T., Motyka, V., Laucou, V., Smets, R., Van Onckelen, H., and Schmülling, T. (2003). Cytokinin-deficient transgenic Arabidopsis plants show multiple developmental alterations indicating opposite functions of cytokinins in the regulation of shoot and root meristem activity. *Plant Cell* 15(11): 2532–50.

Wiltshire, T., Pletcher, M.T., Batalov, S., Barnes, S.W., Tarantino, L.M., Cooke, M.P., Wu, H., Smylie, K., Santrosyan, A., Copeland, N.G., Jenkins, N.A., Kalush, F., Mural, R.J., Glynne, R.J., Kay, S.A., Adams, M.D., and Fletcher, C.F. (2003). Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc. Natl. Acad. Sci. U. S. A.* 100(6): 3380–5.

Wright, E.M., and Kelly, J.D. (2011). Mapping QTL for seed yield and canning quality following processing of black bean (*Phaseolus vulgaris* L.). *Euphytica* 179(3): 471–484.

Wu, G., and Poethig, R.S. (2006). Temporal regulation of shoot development in Arabidopsis thaliana by miR156 and its target SPL3. *Development* 133(18): 3539–47.

Wullschleger, S., Loewith, R., and Hall, M.N. (2006). TOR signaling in growth and metabolism. *Cell* 124(3): 471–84.

Xu, M.Y., Zhang, L., Li, W.W., Hu, X.L., Wang, M.-B., Fan, Y.L., Zhang, C.Y., and Wang, L. (2014). Stress-induced early flowering is mediated by miR169 in Arabidopsis thaliana. *J. Exp. Bot.* 65(1): 89–101.

Yamaguchi, A., Wu, M.-F., Yang, L., Wu, G., Poethig, R.S., and Wagner, D. (2009). The microRNA-regulated SBP-Box transcription factor SPL3 is a direct upstream activator of LEAFY, FRUITFULL, and APETALA1. *Dev. Cell* 17(2): 268–78.

Yan, H. (2012). DOF transcription factors in developing peanut (*Arachis hypogaea*) seeds. *Am. J. Mol. Biol.* 02(01): 60–71.

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

Zakharov, A., Giersberg, M., Hosein, F., Melzer, M., Müntz, K., and Saalbach, I. (2004). Seed-specific promoters direct gene expression in non-seed tissue. *J. Exp. Bot.* 55(402): 1463–71.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., and Buckler, E.S. (2010a). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42: 355–360.

Zhang, L., Tan, Q., Lee, R., Trethewy, A., Lee, Y.-H., and Tegeder, M. (2010b). Altered xylem-phloem transfer of amino acids affects metabolism and leads to increased seed yield and oil content in Arabidopsis. *Plant Cell* 22(11): 3603–20.

Zhao, K., Tung, C.-W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., Norton, G.J., Islam, M.R., Reynolds, A., Mezey, J., McClung, A.M., Bustamante, C.D., and McCouch, S.R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nat. Commun.* 2: 467.

Zheng, M.S., Takahashi, H., Miyazaki, A., Hamamoto, H., Shah, J., Yamaguchi, I., and Kusano, T. (2004). Up-regulation of Arabidopsis thaliana NHL10 in the hypersensitive response to Cucumber mosaic virus infection and in senescing leaves is controlled by signalling pathways that differ in salicylate involvement. *Planta* 218(5): 740–50.

Zhou, Y., Zhang, X., Kang, X., Zhao, X., Zhang, X., and Ni, M. (2009). SHORT HYPOCOTYL UNDER BLUE1 associates with MINISEED3 and HAIKU2 promoters in vivo to regulate Arabidopsis seed development. *Plant Cell* 21(1): 106–17.

CHAPTER 3. DEVELOPING MARKET CLASS SPECIFIC INDEL MARKERS FROM NEXT GENERATION SEQUENCE DATA IN *PHASEOLUS VULGARIS* L.

Abstract

Next generation sequence data provides valuable information and tools for genetic and genomic research and offers new insights useful for marker development. This data is useful for the design of accurate and user-friendly molecular tools. Common bean (*Phaseolus vulgaris* L.) is a diverse crop in which separate domestication events happened in each gene pool followed by race and market class diversification that has resulted in different morphological characteristics in each commercial market class. This has led to essentially independent breeding programs within each market class which in turn has resulted in limited within market class sequence variation. Sequence data from selected genotypes of five bean market classes (pinto, black, navy, and light and dark red kidney) were used to develop InDel-based markers specific to each market class. Design of the InDel markers was conducted through a combination of assembly, alignment and primer design software using 1.6× to 5.1× coverage of Illumina GAII sequence data for each of the selected genotypes. The procedure we developed for primer design is fast, accurate, less error prone, and higher throughput than when they are designed manually. All InDel markers are easy to run and score with no need for PCR optimization. A total of 2687 InDel markers distributed across the genome were developed. To highlight their usefulness, they were employed to construct a phylogenetic tree and a genetic map, showing that InDel markers are reliable, simple, and accurate.

Introduction

Plant breeding embraces both art and science for crop improvement. Marker assisted selection (MAS) can boost the efficiency of breeding when markers linked to genes of interest are discovered (Yang et al., 2012). Marker development requires the comparison of genetic material of two or more genotypes to find the polymorphic regions that segregate in a breeding population. Thus, it is important to have adequate genetic variation among the genotypes of interest to develop markers that can be used for MAS and other genetic studies. In fact, using MAS to improve a trait of interest in a self-pollinating species like common bean is becoming more challenging today in the United States because of the narrow genetic diversity of this species (McClean et al., 1993; Sonnante et al., 1994).

Common bean is a diploid legume species with 11 chromosomes, a genome size of approximately 520 megabase pairs (Mbp), a few duplicated loci (Vallejos et al., 1992; Freyre et al., 1998) and 49% transposable elements (Schlueter et al., 2008). A reference genome sequence of genotype G19833 was recently released in August (Schmutz et al., in press). Common bean has two distinct gene pools, Middle American (from northern Mexico to Colombia) and Andean (from southern Peru to northwestern Argentina). Each gene pool underwent separate domestication events (Gepts and Bliss, 1986; Koenig and Gepts, 1989; Khairallah et al., 1990, 1992; Koinange and Gepts, 1992; Freyre et al., 1996) followed by the creation of ecogeographic races in each of the two gene pools due to further selection under domestication (Singh et al., 1991; Beebe et al., 2000; Diaz and Blair, 2006; Mamidi et al., 2011). Mamidi et al. (2011) estimated the duration and time of the single domestication event in each gene pool and suggested reciprocal migration between wild and landrace genotypes. There is a strong genetic differentiation between Middle American and Andean gene pools, and the Middle American

gene pool is more diverse compared to the Andean gene pool (Mamidi et al., 2013). According to Singh et al. (1991), the Middle American gene pool with the center of domestication in Central and North America consist of three races, Durango, Jalisco, and Mesoamerican. The typical commercial market classes in the United States from this gene pool are pinto, great northern (GN), small red and pink beans from race Durango, and navy, small white and black beans from race Mesoamerican. The Andean gene pool with its center of domestication in South America includes three races: Nueva Granada, Peru, and Chile. The commercial market classes of this gene pool in the United States are light red kidney (LRK), dark red kidney (DRK), white kidney, and cranberry beans which are all from the Nueva Granada race (Mensack et al., 2010).

Breeding for commercial varieties in common bean usually occurs within each market class in order to retain their preferred seed size, shape, color, and pattern. The narrow genetic diversity within a market class is due to the small size of bottleneck populations (Gepts and Bliss, 1986), the rigid quality required by processors and consumers (Ghaderi et al., 1984; Wang et al., 1988; Hosfield et al., 2000; Myers, 2000), the finite use of exotic germplasm (Silbernagel and Hannan, 1988, 1992;Miklas, 2000), and the restricted breeding strategies to meet consumer satisfaction regarding seed size, shape, and color (Singh, 1992). Although incorporation of exotic and unadapted germplasm is helpful in enhancement of genetic diversity, maintenance of the necessary phenotypic characteristics of each market class is challenging when using novel sources of variation due to linkage drag. Indeed, it has been documented in multiple plant species that quantitative traits are affected by genetic background (Tanksley and Hewitt, 1988; Doebley et al., 1995; Lark et al., 1995; Cockerham and Zeng, 1996; Li et al., 1997, 1998). This indicates the need for market class specific marker development to facilitate bean improvement by monitoring the variation that exists in each market class.

Most of the currently available markers for common bean are Sequence Characterized Amplified Region (SCAR) markers developed from Random Amplification of Polymorphic DNA (RAPD) markers through a slow and difficult process (Kelly et al., 2003). Other types of marker systems have been developed and used in different studies in common bean such as Inter Simple Sequence Repeats (ISSR) (Gonzalez et al., 2005), Simple Sequence Repeats (SSR) (Blair et al., 2003; Gomez et al., 2004; Buso et al., 2006; Galeano et al., 2009; Cordoba et al., 2010). Recently a high-throughput Golden Gate SNP assay was released by Hyten et al. (2010). However, most of the markers are based on polymorphism among a few genotypes from different market classes or even gene pools. Thus, the development and application of high throughput, user-friendly, market class-specific markers are indispensable.

Insertion-deletions (InDel) are one of the common sources of variation that are distributed widely throughout the genome. Mechanisms such as transposable elements, slippage in simple sequence replication, and unequal crossover can result in the formation of InDels (Britten et al., 2003). They can be converted to user-friendly markers that can be distinguished easily based on their size (Vali et al., 2008) with minimum laboratory equipment. Many genetic studies in plants and animals have successfully utilized InDels (Hayashi et al., 2006; Vali et al., 2008; Vasemagi et al., 2010; Ollitrault et al., 2012). InDels and SNPs are now the most widely used marker systems in Arabidopsis because they are abundant, PCR-based, and informative due to their co-dominant nature (Pacurar et al., 2012).

Next generation sequencing (NGS) provides inexpensive sequence data needed to develop genetic markers to be used in plant breeding and genetic studies. NGS technologies are efficient and offer new genomic information for minor crops for which a reference genome sequence is not available (Varshney et al., 2009) and accelerates the development of genomic

resources for crops with a reference genome. The objective of this study was to use Illumina sequence data from multiple genotypes within five bean market classes, which were selected from both the Andean and the Middle American gene pools, to develop user-friendly InDel markers that have wide applications for MAS and genomic studies.

Materials and Methods

Plant materials

Three diverse genotypes from pinto, navy, black, and LRK and two genotypes from DRK market classes where sequenced with the Illumina Genome Analyzer (GAII). Genotypes were chosen based on their divergence in a neighbor joining (NJ) tree that was created for 192 genotypes from nine different market classes using 1159 high quality SNP markers (Hyten et al., 2010). The sequenced genotypes in each market class were as follows: Stampede, Buckskin, and Sierra from the pinto market class; C20, Michelite, and Laker from the navy market class; Cornell 49242, T-39, and UI 906 from the black market class, California Early, Lark and, Kardinal from the LRK market class, and Red Hawk and Fiero from the DRK market class.

Marker development

The first step of marker design was a within genotype *de novo* assembly of the Illumina GAII DNA sequence data into contigs using Velvet 1.0 (Zerbino and Birney, 2008) with the default settings. BLAST+ (Camacho et al., 2009) was used to discover potential InDels. Using three genotypes within a market class, InDels were discovered by aligning contigs from one genotype as the query against a database consisting of the contigs of the two other genotypes. In addition, a pairwise blastn alignment was performed among all pairs of genotypes within a market class. An e-value cutoff of 1E-50 and a maximum hit of one and two were used in BLAST to obtain the best hit for pair-wise and three-way alignments, respectively. InDels were

discovered based on the size differences between the query and the database subject. Several filters were applied to potential InDels: A minimum InDel size of 8 bp was used to ensure an appropriate resolution using agarose gel electrophoresis; unique InDel fragments were assured by blasting InDel fragments against the *Phaseolus vulgaris* L. scaffold assembly ARRA-V0.9. ARRA-V0.9 was an intermediate scaffold assembly in the whole genome sequencing project. The e-value and maximum hit were set to 1E-50 and 20, respectively. Queries with multiple hits were excluded to decrease the probability of designing markers from repetitive regions. Contigs that contained more than four consecutive Ns were excluded because this could lead to false InDel discovery or false InDel size.

The contigs that contained InDels were aligned using Multalin (Corpet, 1988) to obtain the consensus region around the InDels for primer design. Primers were designed from the consensus sequence in BatchPrimer3 (You et al., 2008). The primer size parameters were set to 22, 26, and 32 bp as the minimum, optimum, and maximum size, respectively, and GC content was set to 35, 50, and 60% as the minimum, optimum, and maximum, respectively. Primer annealing temperature was set to a very narrow range of 67, 68, and 69°C as the minimum, optimum, and maximum temperature, respectively. Finally, the maximum Tm difference between forward and reverse primers was set to 2°C only. The PCR product length was approximatly10 times the InDel size to ensure the PCR products could be adequately separated on a 3% agarose gel for efficient scoring. The product size varied between InDel size (bp) × 10 and [InDel size (bp) × 10] + 10 (bp) for optimum and maximum values, respectively. The minimum product sizes were as listed in Table 3.1. These stringent criteria were intentional to avoid the need for PCR optimization for each primer set. All primer sets were optimized to amplify with a 55°C annealing temperature.

Table 3.1. InDel size and the corresponding minimum product size that was used by BatchPrimer3 for primer design.

InDel size (bp)	Minimum product size (bp)
8-9	70
10-11	80
12-14	90
15-17	100
18-22	110
23-26	120
27-29	130
30-36	150

Nomenclature

Including information on marker position on the reference genome in the marker names, provides the user with valuable information on marker distribution. Thus, the markers were named in the format NDSU-IND-NN-XX.XXXX, where NDSU stands for North Dakota State University, IND for InDels and NN for the chromosome number. The Xs represent the physical position on the reference genome (G19833- V1.0) in megabase pairs up to four decimals. For example InDel marker NDSU_IND_07_02.6485 is located on chromosome Pv07 at position 2.6485 Mbp in common bean V1.0 reference genome.

PCR amplification

The PCR protocol used to amplify all InDel markers was: 3 min at 95°C for one cycle, 20 s at 95°C, 30 s at 55°C, and 1 min at 72°C for 45 cycles, and 10 min at 72°C for one cycle. Each PCR reaction consisted of a final concentration of 1× PCR buffer including 0.15 mM $MgCl_2$, 0.5 mM dNTP mix, 0.25 mM forward/reverse primers and 1 unit of Taq polymerase with a 20 µl final volume.

Alignment of sequence data with the reference genome

The sequence data from 14 genotypes were mapped to the reference genome (V1.0) when the complete assembly became available. Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) with default settings was used to map the reads with the reference genome. SAMtools (Li et al., 2009) was used to convert the BWA output to a sorted bam file and obtain the mpileup file. The "pileup2indel" command with minimum coverage of five reads was used in VarScan (Koboldt et al., 2009) to find the number of InDels for each genotype based on the reference genome (G19833). The VarScan output was also filtered based on the frequency of the variant allele by read count. Only InDels with variant allele frequency of 80% and higher were considered.

Marker performance and application

To evaluate the performance of the designed markers, 219 pinto markers were screened on Stampede, Sierra, Buckskin, and G19833. Moreover, six markers were tested on a few random genotypes from nine market classes (Table 3.2) to evaluate the performance of InDel markers in the market classes other than the one from which they were originally designed.

Table 3.2. Genotypes used to test the performance of six markers in other market classes. Six random genotypes were selected from pinto, great northern, navy, black, pink, and small red market classes and four random genotypes were selected from dark and light red kidney and snap bean marker classes.

Order*	Genotype	Market class
1	Domino	Black
2	Raven	Black
3	T39	Black
4	Cornell49242	Black
5	Shania	Black
6	Black Knight	Black
7	BelMiNeb-RMR-3	Great northern
8	Matterhorn	Great northern

Table 3.2. Genotypes used to test the performance of six markers in other market classes (continued).

Order*	Genotype	Market class
9	Tara	Great northern
10	Coyne	Great northern
11	JM-24	Great northern
12	Gemini	Great northern
13	Michelite	Navy
14	Sanilac	Navy
15	Seafarer	Navy
16	Bunsi	Navy
17	C20	Navy
18	Laker	Navy
19	Pink Floyd	Pink
20	Victor	Pink
21	Viva	Pink
22	Roza	Pink
23	Gloria	Pink
24	PK915	Pink
25	AC Redbond	Small red
26	AC Earlired	Small red
27	Sapphire	Small red
28	E Mber	Small red
29	UI-3	Small red
30	NW-63	Small red
31	Sierra	Pinto
32	Buckskin	Pinto
33	Durango	Pinto
34	PT7-2	Pinto
35	Lariat	Pinto
36	Hatton	Pinto
37	Chinook 2000	Light red kidney
38	K-42	Light red kidney
39	Blush	Light red kidney
40	VA-19	Light red kidney
41	Montcalm	Dark red kidney
42	USDK-CBB-15	Dark red kidney
43	Fiero	Dark red kidney
44	CDRK	Dark red kidney
45	Benton	Snap bean

Table 3.2. Genotypes used to test the performance of six markers in other market classes (continued).

Order*	Genotype	Market class
46	91-G	Snap bean
47	Harvester	Snap bean
48	Cantare	Snap bean

*Indicates the order of the genotypes from left to right on the agarose gel in **Figure 3.3**.

To assess the InDel markers for applied genetic studies, 196 markers were used to screen 24 diverse pinto genotypes. The 24 pinto genotypes were as follow: Sierra, Aztec, Santa Fe, La Paz, Stampede, ND-307, Medicine Hat, Lariat, BelDakMe-RR-5, Sequoia, Remington, Durango, Max, PT7-2, Ouray, JM-126, Olathe, Hatton, Apache, UI-114, Nodak, Buckskin, Flint, and UI-196. These genotypes were chosen based on their diversity in a NJ tree which was constructed in ClustalX 2.1 (Larkin et al., 2007) using 1300 SNP markers (Hyten et al., 2010). The number of bootstraps used in ClustalX was 1000 and the genotypes were chosen from clusters that were diverse and had high bootstrap values in the tree. The 24 pinto genotypes were screened using the 196 markers and the markers that showed polymorphism were used to evaluate the performance of the InDel markers for distinguishing the 24 genotype and to construct a NJ tree. PowerMarker version 3.25 (Liu and Muse, 2005) was used to construct the NJ tree with 1000 bootstraps. The F_{st} value was calculated in PowerMarker as well to evaluate the overall genetic divergence among this collection of genotypes.

In addition, an F_2 population was used to evaluate the InDel markers for mapping purposes. The F_2 population, NDZ-11002 was derived from a cross between Lariat × Medicine Hat and consisted of 87 F2 genotypes. Eighty two pinto markers that were polymorphic between the two parents were employed to screen the F_2 population, and CarthaGène (De Givry et al., 2005) was used to build the genetic map. In CarthaGène, the "group" command was used with a

distance and LD threshold of 20 cM and 3.00, respectively to group markers into linkage groups. The "Buildfw 2 2 {} 1" command was used to order the markers on each linkage group and to obtain the map with the highest likelihood value. Qgene (Nelson, 1997) was then used to visualize the map images.

Multiplexing

Multiplexing of InDel markers was conducted using two and four markers in the same reaction mix. A total of one duplex and six fourplex sets were tested on 96 Middle American genotypes. The protocol for four markers in a 20 µl reaction mix was as follow: 1× PCR Buffer including 0.15 mM $MgCl_2$, 0.8 mM dNTP, 0.125 mM of each forward and reverse primer [total of 0.5 mM (0.125 × 4) forward and reverse primers] and 2 units of Taq polymerase. The protocol for multiplexing two markers was the same as above with an exception that 0.25 mM of each forward and reverse primer [total of 0.5 mM (0.25 × 2)] was used. The PCR amplification cycle was the same as the cycle used to amply a single marker. The resulting amplification products were visualized on a 3% agarose gel.

Results

Illumina sequencing, *de novo* assembly and primer design

The DNA sequence data consisted of 19.6 billion bases in the form of 114 bp paired-end reads from 250 to 400 bp size selected fragments of 14 genotypes. The 114 bp paired-end reads did not include Illumina adaptor sequences. The sequencing coverage ranged from 1.6× to 5.1× with an average of 3.7×. Laker and Kardinal had the lowest and highest number of raw GAII reads, 3,711,450 and 11,748,671 reads, respectively. Cornell 49242 with 261,313 and Stampede with 752,015 contigs had the smallest and largest number of assembled contigs, respectively. Only contigs 120 bp or greater were used for the assembly statistics. The mean contig length

across all 14 genotypes was 322 bp. The N50 varied from 222 to 651 bp, with an average N50 value of 394 bp. The GC content ranged from 32.5 to 40.9% with an average of 34.4%. The minimum GC value of 35% was used to design primers in BatchPrimer3. The Illumina reads and assembly information are summarized in Table 3.3.

Table 3.3. Illumina paired-end reads information and contig information after the *de novo assembly*.

Genotype	Illumina GAII Reads			Assembly			
	Market class	Number of reads (In Millions)	Genome coverage (x)*	Total length of asse Mbled contigs (Mbp)+	Number of asse Mbled contigs	N50 (bp)	GC%
Cornell49242	Black	7.20	3.2	66.97	261,313	300	40.9
T39	Black	8.42	3.7	178.16	527,965	422	32.6
UI906	Black	6.44	2.8	107.99	400,652	320	33.3
Red Hawk	DRK	7.20	3.2	128.33	427,260	360	35.0
Fiero	DRK	7.61	3.3	146.58	470,693	373	34.0
Cal Early	LRK	7.52	3.3	145.26	461,881	387	34.2
Lark	LRK	10.30	4.5	199.12	531,149	501	33.8
Kardinal	LRK	11.74	5.1	238.71	534,298	651	32.5
C20	Navy	7.84	3.4	100.52	355,778	338	37.6
Michelite	Navy	9.47	4.1	170.02	504,441	426	34.3
Laker	Navy	3.71	1.6	81.10	381,399	222	33.8
Stampede	Pinto	10.20	4.5	214.71	752,015	330	33.2
Sierra	Pinto	9.58	4.2	178.36	532,937	409	32.7
Buckskin	Pinto	10.45	4.6	190.93	529,875	477	33.7

* The genome coverage was calculated based on a 521 Mbp genome size and 114bp paired-end Illumina reads.

+ Contigs equal or longer than 120bp were considered to report the assembly statistics.

Alignment of the reads with the V1.0 reference genome for 14 genotypes indicated a range of 2330 to 45,770 InDels of 1 bp or greater across the genome. The minimum and maximum number of InDels belonged to the Laker and Buckskin genotypes, respectively (Table 3.4).

Table 3.4. Number and distribution of InDels in each genotype when aligned with G19833. The numbers are based on InDel size of 1bp and greater, and 521 Mbp genome size. VarScan was used to discover the InDel polymorphisms.

Genotype	Market class	Genome Coverage (x)	Number of InDels in alignment with G19833	InDel frequency per Mbp	Number of InDels in 1x coverage
Cornell49242	Black	3.2	9,226	17.7	2,901
T39	Black	3.7	33,662	64.6	9,049
UI906	Black	2.8	13,681	26.3	4,817
Red Hawk	DRK	3.2	6,899	13.2	2,169
Fiero	DRK	3.3	8,961	17.2	2,667
Cal Early	LRK	3.3	7,017	13.5	2,114
Lark	LRK	4.5	17,056	32.7	3,749
Kardinal	LRK	5.1	27,122	52.1	5,226
C20	Navy	3.4	11,205	21.5	3,238
Michelite	Navy	4.1	31,455	60.4	7,525
Laker	Navy	1.6	2,330	4.5	1,421
Stampede	Pinto	4.5	37,902	72.8	8,404
Sierra	Pinto	4.2	32,199	61.8	7,612
Buckskin	Pinto	4.6	45,770	87.9	9,907

The number of aligned contigs (BLAST output) within a market class varied between 296,852 and 859,350 in the three-way alignments. The aligned contigs were filtered based on the InDel size, and those with a tentative size of 8 bp or greater were retained. This significantly reduced the number of sequences to analyze in the next step. For example, the number of contigs in the DRK market class dropped from 296,852 to 1378, and the reduction was from 859,350 to 9634 for pinto market class. Filtering for uniqueness of hits in the common bean reference scaffolds reduced the range of contigs from 450 in DRK to 6010 in pinto, with an average value of 2114 contigs across all market classes. The total number of consensus sequences submitted to BatchPrimer3 across all alignments (three-way and all pair-wise alignments) varied from 11,406 in the pinto to 324 in the DRK market class when fragments containing four consecutive Ns in

their sequence were removed (Table 3.5). The basic properties of 11,406 pinto contigs submitted to BatchPrimer3 are summarized in Table 3.6.

Table 3.5. Filtering criteria for contigs used for primer design. The numbers are provided only for three-way alignment in all market classes except the dark red kidney market class.

Market classes	Number of contigs with InDels in three-way alignment			Numbers across all alignments	
	Blast hits	InDel size ≥8	Unique scaffold hit/ InDel ≥8	Contigs submitted to Batch primer3	Number of primer pairs
Pinto	859,350	9,634	6,010	11,406	1,343
Black	342,085	1,955	708	1,913	292
Navy	507,147	1,515	650	1,867	323
LRK	834,726	5,524	2,754	5,456	669
DRK	296,852 (pair-wise)	1,378	450	324	60

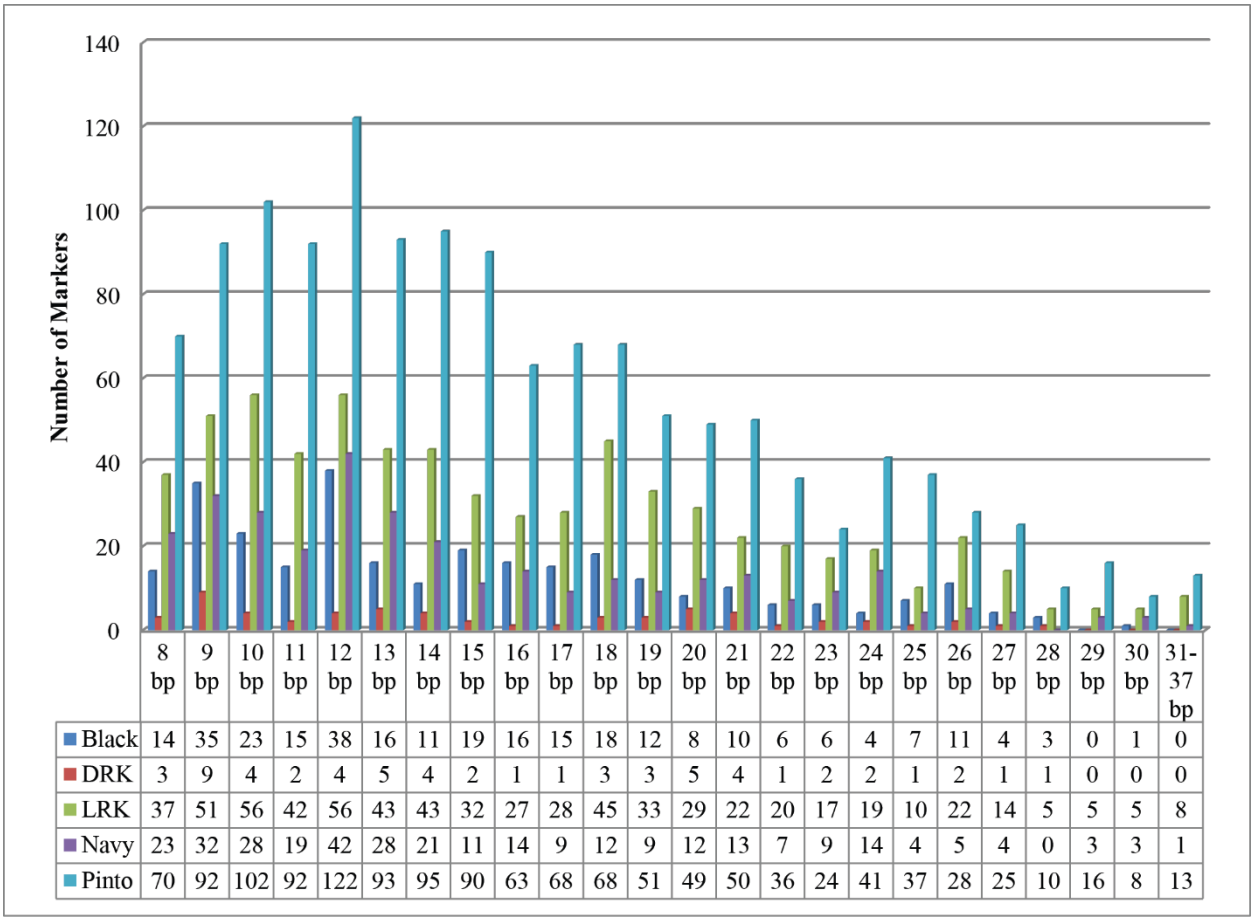
Table 3.6. Pre-analysis of 11,406 pinto contigs submitted to BatchPrimer3.

Item	Mean	Std. deviation	Min	Max	Coe. of Variation (%)
Sequence length (bp)	390.49	174.45	104	2,680	44.67
GC contents (%)	29.04	5.92	4.45	53.21	20.38

The final number of primer pairs ranged from 1343 in the pinto market class (highest) to 60 in the DRK market class (lowest) (Table 3.5). There were a total of 2687 primer pairs designed across all market classes (Supplementary Table[2]), all having single hits on the common bean reference genome V1.0 and only six homologs to common bean transposable elements (Scott Jackson, personal communication). The distribution of InDel sizes in each market class is illustrated in Figure 3.1. The average distribution of InDel markers varied from one per132 Kb

[2] http://www.frontiersin.org/journal/10.3389/fpls.2014.00185/abstract

on chromosome Pv05 to one per 314 Kb on chromosome Pv03 with an average of one InDel marker every 200 Kb across the genome. Although the euchromatic region forms less than half of the bean genome (44.1%), the marker density was higher in this region (65.8%) (Figure 3.2). Markers located in the pericentromeric region are highlighted in orange color in the Supplementary Table. The number of markers on the chromosomes varied from 144 to 333 on chromosomes Pv10 and Pv02, respectively, with an average value of 238 markers per chromosome.



	8 bp	9 bp	10 bp	11 bp	12 bp	13 bp	14 bp	15 bp	16 bp	17 bp	18 bp	19 bp	20 bp	21 bp	22 bp	23 bp	24 bp	25 bp	26 bp	27 bp	28 bp	29 bp	30 bp	31-37 bp
Black	14	35	23	15	38	16	11	19	16	15	18	12	8	10	6	6	4	7	11	4	3	0	1	0
DRK	3	9	4	2	4	5	4	2	1	1	3	3	5	4	1	2	2	1	2	1	1	0	0	0
LRK	37	51	56	42	56	43	43	32	27	28	45	33	29	22	20	17	19	10	22	14	5	5	5	8
Navy	23	32	28	19	42	28	21	11	14	9	12	9	12	13	7	9	14	4	5	4	0	3	3	1
Pinto	70	92	102	92	122	93	95	90	63	68	68	51	49	50	36	24	41	37	28	25	10	16	8	13

Figure 3.1. Distribution of 2687 InDel sizes in five market classes.

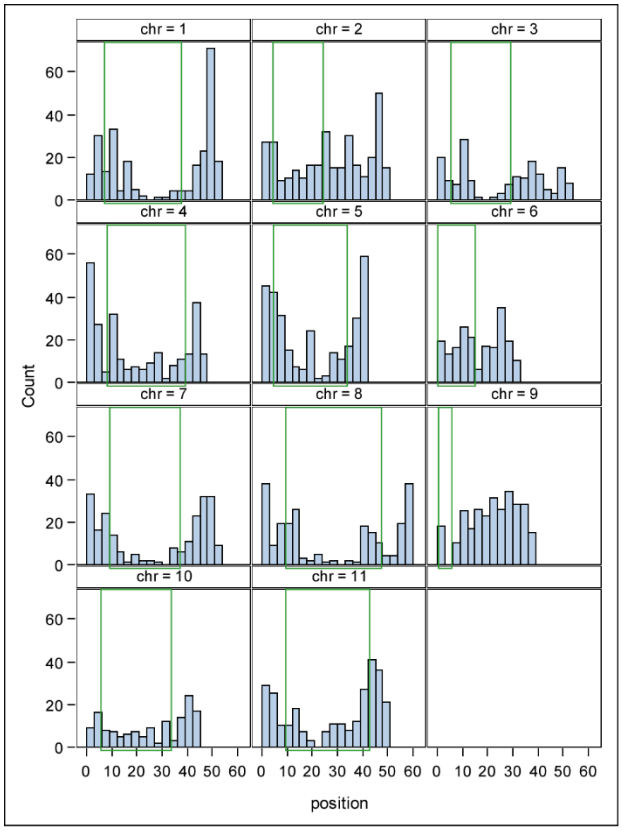


Figure 3.2. Physical distribution of 2687 InDel markers across 11 chromosomes of common bean. The x axis shows the chromosome length in Mbp and the y axis represents the frequency of InDel markers. The green rectangle indicates the pericentromeric region in each chromosome.

Marker performance

To evaluate the performance of the InDel markers, a total of 219 markers from the pinto market class were tested with Stampede, Sierra, Buckskin, and G19833. A total of 196 markers showed polymorphism (89.5%), and only 23 markers (10.5%) were either monomorphic or difficult to score among four genotypes. Six markers from four market classes were tested on 48 random genotypes from nine different market classes as well. Although the primers were originally designed for a specific market class, we observed polymorphism among the genotypes from other market classes. Based on a sample of genotypes, markers from the Middle American

gene pool did not show polymorphism in light and DRK market classes and the marker from the LRK market class showed polymorphism only among Andean genotypes (Figure 3.3).

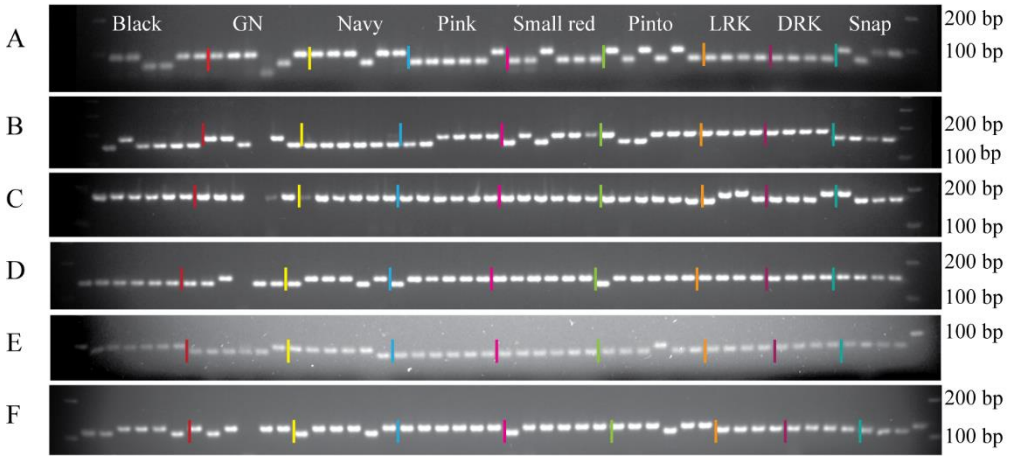


Figure 3.3. Six InDel markers tested on random genotypes from nine different market classes. The names of the genotypes, from left to right, are listed in Table 3.2. (A) Marker NDSU_IND_07_02.6485 from pinto market class. (B) Marker NDSU_IND_10_42.1355 from pinto market class. (C) Marker NDSU_IND_06_12.3324 from light red kidney market class. (D) Marker NDSU_IND_05_01.7405 from pinto market class. (E) Marker NDSU_IND_08_36.2119 from navy market class. (F) Marker NDSU_IND_09_07.6278 from black market class. First lane from right in all panels is the DNA Ladder.

Marker application

Phylogenetic analysis

A set of 196 InDel markers were used to screen 24 diverse pinto genotypes (Table 3.7) and 172 (87.7%) were polymorphic and used in a phylogenetic analysis. The NJ tree and the F_{st} value indicated two distinct clusters among the 24 pinto genotypes (Figure 3.4).

Table 3.7. Specifications of 24 pinto genotypes.

Genotype	Growth habit	Application date	Source
PT7-2	II to III	Not released	USDA-ARS-Washington
Sequoia	IIb	NA[1]	ISB[2]
Max	III	NA	ISB
Santa Fe	IIa	2010	MSU[3]
Stampede	IIa	2008	NDSU[4]
La Paz	IIb	2008	ProVita,Inc.
ND-307	IIb	2008	NDSU
Lariat	IIb	2008	NDSU
Medicine Hat	IIa	2007	Seminis
Durango	IIb	2007	ProVita,Inc.
Remington	IIb	1996	RogersSeedCo
Hatton	IIIa	1996	NDSU
Buckskin	IIIb	1995	Novartis Seed Inc.
Apache	IIIa	1995	ISB
Aztec	IIb	1993	MSU
BelDakMi-RR-5	II	1993	USDA-ARS-Beltsville-MD
Sierra	IIb	1990	MSU
UI-196	IIIb	1990	UI[5]
Flint	II to IIIb	1989	RogersSeedCo
JM-126	III to IIIa	1986	USDA-ARS/WSU[6]
Nodak	III	1985	USDA–ARS/NDSU
Olathe	III	1980	CSU[7]
Ouray	III to IIIa	1975	CSU
UI-114	III	1967	UI

[1] NA- Information not available
[2] ISB- Idaho Seed Bean Company
[3] MSU- Michigan State University
[4] NDSU- North Dakota State University
[5] UI- University of Idaho
[6] WSU-Washington State University
[7] CSU- Colorado State University

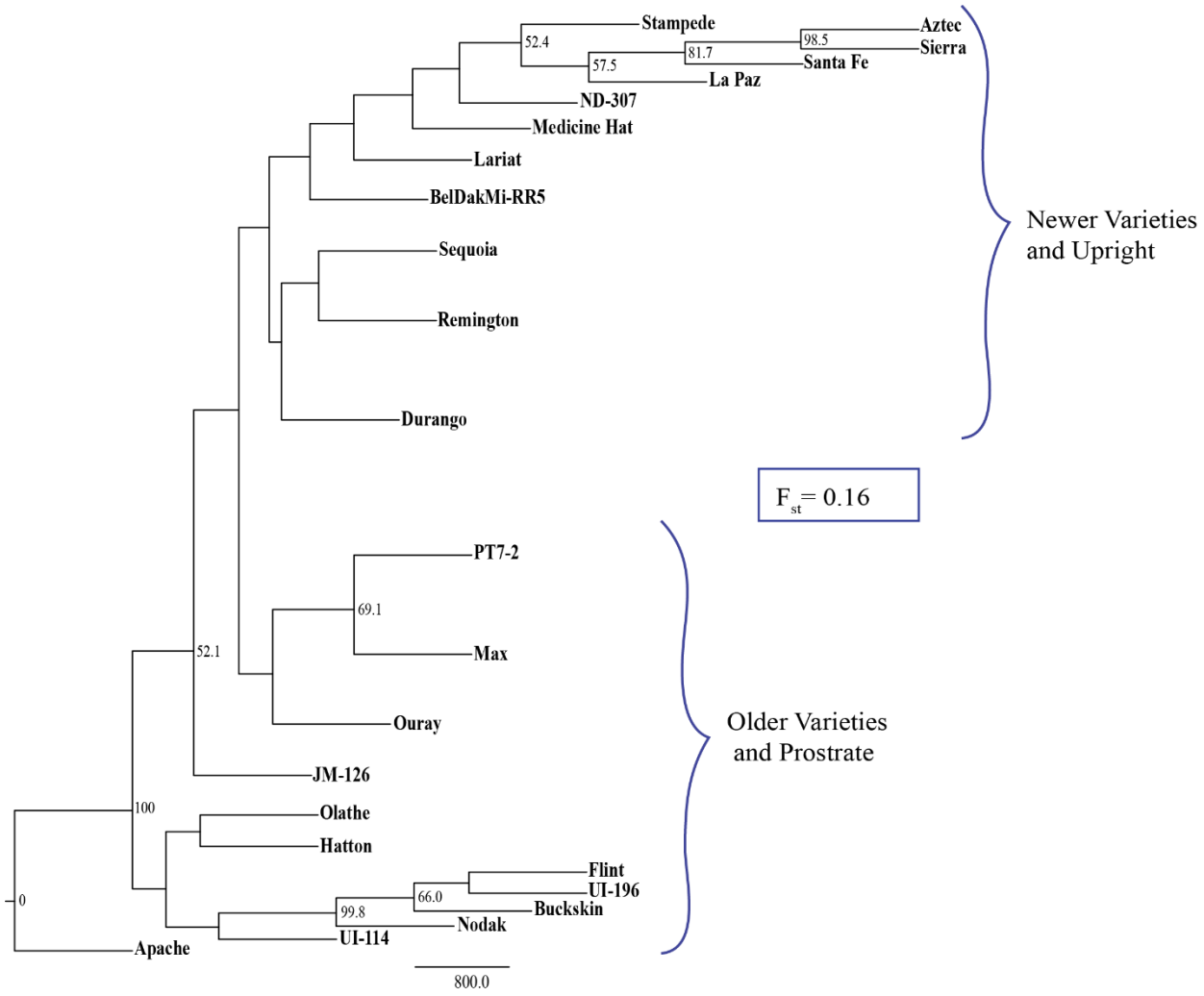


Figure 3.4. Neighbor joining tree of 24 pinto genotypes that cluster into two distinct groups (i) newer varieties with type II growth habit and (ii) older varieties with type III growth habit. The Fst value of 0.16 indicates the degree of variation between the two groups. Bootstrap values greater than 50% are shown on the nodes.

Genetic map

Eighty two polymorphic InDel markers where used to construct a genetic map. A total of nine linkage groups that correspond to nine chromosomes (all chromosomes except one and four) were built from the F_2 population with 87 genotypes. Five pairs of markers co-segregated in CarthaGène analysis. Among the 77 remaining markers, 18 were excluded from the genetic map when the "Buildfw" function of CarthaGène with a LOD threshold of 2.2 was used. The genetic and physical order was consistent for 54 of the 59 marker loci (Figure 3.5).

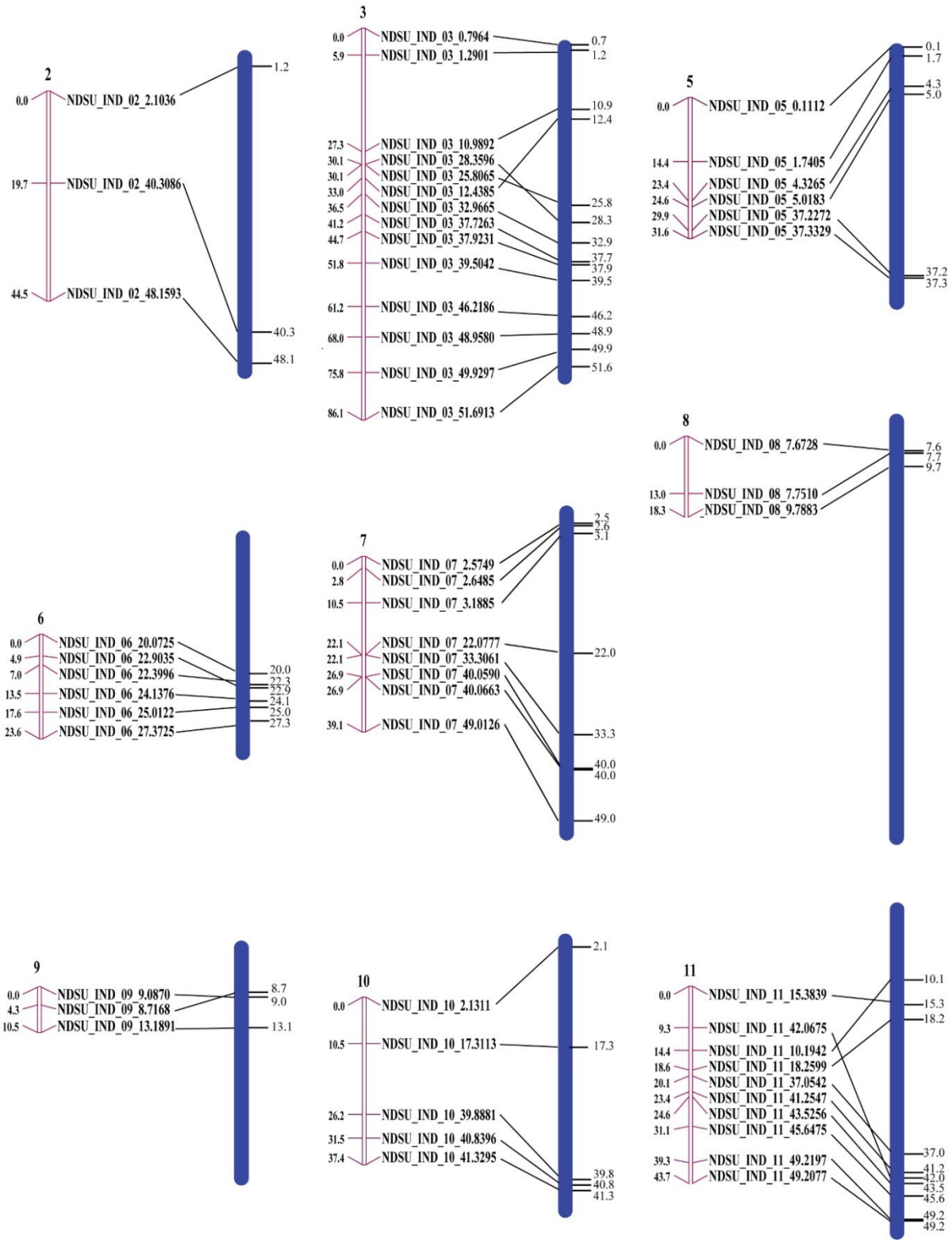


Figure 3.5. Correspondence between genetic and physical positions. The pink bars are linkage groups and the blue bars are the chromosomes with the physical positions of the InDel markers on their right side. The sizes of the chromosomes are proportional to their actual size.

Multiplexing

Multiplexing of tested InDel markers showed clear and scorable bands on the 3% agarose gel when one set of duplex and six sets of fourplex were used. Figure 3.6 illustrates the results of two fourplex sets on 48 Middle American genotypes on a 3% agarose gel as an example. The marker information is provided in Table 3.8.

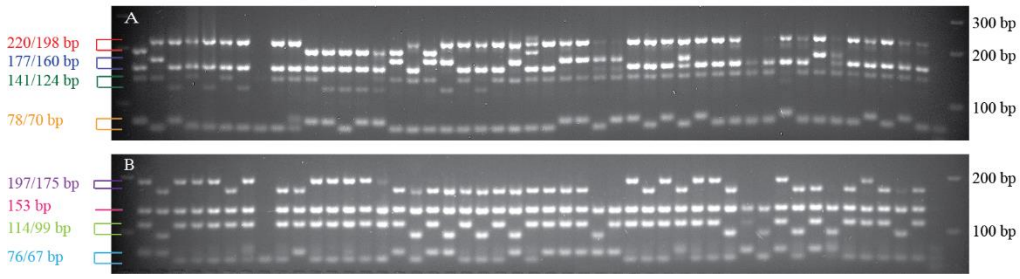


Figure 3.6. Multiplexing of markers on 48 Middle American bean genotypes showed distinct bands on the 3% agarose gel electrophoresis. (A) Amplification products using InDel markers NDSU_IND_05_37.2272, NDSU_IND_06_16.5002, NDSU_IND_11_30.9655, and NDSU_IND_06_31.8021 on 48 bean genotypes. All four markers showed polymorphism. (B) Amplification products using InDel markers NDSU_IND_11_33.0572, NDSU_IND_07_42.1709, NDSU_IND_07_25.1928, and NDSU_IND_10_19.1957 on the same 48 bean genotypes. Marker NDSU_IND_07_25.1928 was monomorphic and the other three were polymorphic. The first lane from the right in (A,B) are the DNA Ladders.

Table 3.8. Specifications of InDel markers used for multiplexing (two sets of fourplex). Marker shaded in grey was monomorphic.

Marker ID	Market class	Max-product size (bp)	Min-product size (bp)	InDel size (bp)
NDSU_IND_05_37.2272*	Pinto	78	70	8
NDSU_IND_06_16.5002*	Black	141	124	17
NDSU_IND_11_30.9655*	Pinto	177	160	17
NDSU_IND_06_31.8021*	Pinto	220	198	22
NDSU_IND_11_33.0572+	Pinto	76	67	9
NDSU_IND_07_42.1709+	Pinto	114	99	15
NDSU_IND_07_25.1928+	Black	153	138	15
NDSU_IND_10_19.1957+	Pinto	197	175	22

*One four-plex marker set mixed in one PCR reaction
+ The second four-plex marker set mixed in one PCR reaction

Discussion

Common bean is a diverse crop species with much variation in the seed color, shape, and many other phenotypic characteristics. The species includes wild types, landraces which are the domesticated forms, ecogeographical races which are the result of selection, and market classes within each of the ecogeographical races. Plant breeding is generally restricted to market classes to retain the specific characteristics of the market class. However, as indicated by the high polymorphism rate (87.7%) based upon our analysis of 24 genotypes of the pinto market class, InDel markers appear to be polymorphic even within a market class. InDel markers are easy to use co-dominant markers and are present throughout the genome. With the availability of abundant next generation sequence data, identification of InDels has become a simple process.

Marker development

We selected diverse genotypes in each market class based on the most comprehensive SNP dataset available. The Illumina GA II was used to generate paired end reads. The Illumina technology results in short reads but high coverage (Vera et al., 2008) as well as high quality data where 70% of base calls in 2 × 75 bp paired-end sequences have a quality score of Q30 or higher. The standard paired-end libraries of Illumina with a length between 200 and 500 bp can provide a platform to identify large and small InDels, inversions and other rearrangements. Paired-end reads boost the robustness of *de novo* assembly, SNP identification, and InDel discovery.

In this study we developed a genome wide collection of 2687 InDel markers that can be amplified without any PCR optimization and with minimum lab equipment. One of the filtering criteria that dramatically decreased the number of markers was the InDel size. There was a 215-fold decrease in the number of potential InDels in DRK when the contigs from the BLAST

output were filtered for a minimum InDel size of 8 bp, and this reduction was about 89-fold for the pinto market class in the three-way alignment. However, filtering for uniqueness of hits to the common bean reference scaffolds (ARRA-V0.9) did not cause a dramatic reduction. For example, the number of contigs in the DRK market class dropped to one third, and this reduction was only 1.6-fold for the pinto market class in the three-way alignment. Generally there were less InDels in the DRK market class possibly due to the presence of only two sequenced genotypes in this market class. Moreover, Andean types are reported to have narrower genetic diversity compared to the Mesoamerican genotypes (Beebe et al., 2001). The stringent primer design criteria also resulted in another huge drop in the number of primer pairs that were selected. These stringent criteria were necessary because the development of markers should be precise and cost effective with proper throughput (Jander et al., 2002).

Several factors affect the discovery of functional InDel markers. As observed in Arabidopsis, decreasing InDel size from 25 to 6 bp increased the number of markers from 277 to 1073 (Salathia et al., 2007; Hou et al., 2010). The phylogenetic relationship between the genotypes used for InDel discovery is important. Kardinal, an Andean genotype, is more closely related to the reference genome (G19833), another Andean genotype, than Buckskin, a Middle American genotype. Less Kardinal InDels were observed than Buckskin InDels (27,122 vs. 45,770) even thought it had greater read coverage (5.1× vs. 4.6×). This trend was observed for all genotypes: more InDels were discovered among Middle American genotypes than the Andean genotypes because the reference genome is of Andean origin (Table 3.4).

In total we discovered 2687 InDel markers for an average of one per 200 Kb. The fact that they are preferentially distributed in the highly recombinogenic region of the genome increases their utility for multiple genetic analyses.

Marker application

In our study, 87.7% of the 196 markers that were used in the phylogenetic analysis of 24 pinto genotypes were polymorphic. In the NJ tree, the pintos were separated based on plant architecture and the application/release date of the variety. Indeed, newer, upright pintos clearly clustered separately from older, prostrate pintos with fixation index (F_{st}) of 0.16 which indicates a great degree of genetic divergence among subpopulations (Hartl and Clark, 1997). This might be due to selection for growth habit in bean breeding programs where the newer breeding programs prefer the upright beans since this trait offers several advantages such as ease of management, higher grain yield, and reduced disease issues (Cunha et al., 2005). InDel markers have been used for phylogenetic studies. Steele et al. (2008) used InDel polymorphisms in rice to separate Basmati genotypes from other genotypes. Ollitrault et al. (2012) showed that citrus diversity and phylogenetics based on InDel data are consistent with those based on SSR markers.

The observation that InDel markers developed from one market class showed polymorphism in the other market classes indicates their broad utility. This denotes that although these InDel markers were designed to capture the variation within each market class, their performance and application can be expanded to the entire bean germplasm.

The InDel markers should be useful for genetic map construction because there are on average about 200 markers on each chromosome. We used a relatively small mapping population to illustrate their application for linkage analysis. Generally, recombination occurs more frequently in regions distal to the centromeric region (Curtis and Lukaszewski, 1991; Tanksley et al., 1992; Werner et al., 1992). Because of low marker density and a small number of recombination events, our map did not cover the centromeric blocks in the F2 mapping study, resulting in mapping only a portion of the chromosome or of two clusters of markers, one from

each arm. There were five discrepancies in our genetic map relative to the physical map. Marker order differences between the genetic and physical map or genetic maps from different populations or marker systems has been observed in other studies as well (Snelling et al., 2007; Wei et al., 2007; Xia et al., 2007). These differences could be a result of sequence assembly errors, inversions, and segregation distortion.

The possibility of conducting multiplex PCR is another indicator of the broad utility of InDel markers. With multiplexing, genotyping is even more cost effective due to reduced amount of reagents and DNA quantity needed for PCR amplification. Moreover, this method saves time when hundreds of markers are screened and broadens the coverage when DNA availability is limited (Edwards and Gibbs, 1994; Karaiskou and Primmer, 2008). Our InDel markers meet many of the criteria that Henegariu et al. (1997) mentioned as critical parameters in a multiplex PCR. According to their study, some of the basic principles include the appropriate primer length which should be 18–34 bp or higher and all of our primers are designed with a length of 26–32 bp. Henegariu et al. (1997) also reported that by increasing the primer length up to 28–30 bp, the annealing temperature could be increased resulting in a reduction of non-specific PCR products. GC content of 35–60% and annealing temperature between 55 and 58°C are other basic principles of multiplex PCR. The primers we designed have a minimum GC content of 35%, and all amplify at 55°C. Henegariu et al. (1997) indicated 54°C as the optimum temperature for co-amplification of loci in the multiplex PCR. Although the probability of non-specific product amplification increases at this temperature, simultaneous amplification of many specific loci greatly suppresses the yield of non-specific amplification products due to limited enzyme and nucleotide resources.

In conclusion, this study shows the usefulness of DNA sequence data as the raw material for primer development in the presence or absence of a reference genome. We show that contigs obtained from the *de novo* assembly of sequence data are sufficient for polymorphism discovery. However, without a completely assembled reference genome or a set of primary scaffolds, contigs cannot be filtered to eliminate the duplicate loci or transposable elements. The reference sequence could reduce the development of redundant markers and allow the determination of the exact physical position and order of the markers. The availability of high density markers affects the success of genetic map construction, map-based cloning (Pacurar et al., 2012) and diversity studies. The availability of 2687 InDel primers will enhance MAS and diversity studies in common bean.

References

Beebe, S., Rengifo, J., Gaitan, E., Duque, M. C., and Tohme, J. (2001). Diversity and origin of Andean landraces of common bean. *Crop Sci*. 41, 854–862.

Beebe, S., Skroch, P. W., Tohme, J., Duque, M. C., Pedraza, F., and Nienhuis, J. (2000). Structure of genetic diversity among common bean landraces of Middle American origin based on correspondence analysis of RAPD. *Crop Sci*. 40, 264–273.

Blair, M. W., Pedraza, F., Buendia, H. F., Gaitan-Solis, E., Beebe, S. E., Gepts, P., et al. (2003). Development of a genome-wide anchored microsatellite map for common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet*. 107, 1362–1374.

Britten, R. J., Rowen, L., Williams, J., and Cameron, R. A. (2003). Majority of divergence between closely related DNA samples is due to indels. *Proc. Natl. Acad. Sci. U.S.A.* 100, 4661–4665.

Buso, G. S. C., Amaral, Z. P. S., Brondani, R. P. V., and Ferreira, M. E. (2006). Microsatellite markers for the common bean*Phaseolus vulgaris*. *Mol. Ecol. Notes* 6, 252–254.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Cockerham, C. C., and Zeng, Z. B. (1996). Design III with marker loci. *Genetics* 143, 1437–1456.

Cordoba, J. M., Chavarro, C., Schlueter, J. A., Jackson, S. A., and Blair, M. W. (2010). Integration of physical and genetic maps of common bean through BAC-derived microsatellite markers. *BMC Genomics* 11:436.

Corpet, F. (1988). Multiple sequence alignment with hierarchical-clustering. *Nucleic Acids Res*. 16, 10881–10890.

Cunha, W. G. D., Ramalho, M. A. P., and Abreu, Â. D. F. B. (2005). Selection aiming at upright growth habit common bean with carioca type grains. *Crop Breed. Appl. Biotechnol*. 5, 379–386.

Curtis, C. A., and Lukaszewski, A. J. (1991). Genetic-linkage between C-bands and storage protein genes in chromosome-1b of tetraploid wheat. *Theor. Appl. Genet*. 81, 245–252.

De Givry, S., Bouchez, M., Chabrier, P., Milan, D., and Schiex, T. (2005). CAR(H)(T)AGene: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics* 21, 1703–1704.

Diaz, L. M., and Blair, M. W. (2006). Race structure within the Mesoamerican gene pool of common bean (*Phaseolus vulgaris* L.) as determined by microsatellite markers. *Theor. Appl. Genet*. 114, 143–154.

Doebley, J., Stec, A., and Gustus, C. (1995). Teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* 141, 333–346.

Edwards, M. C., and Gibbs, R. A. (1994). Multiplex PCR: advantages, development, and applications. *PCR Methods Appl*. 3, S65–S75.

Freyre, R., Rios, R., Guzman, L., Debouck, D. G., and Gepts, P. (1996). Ecogeographic distribution of *Phaseolus* spp (Fabaceae) in Bolivia. *Econ. Bot*. 50, 195–215.

Freyre, R., Skroch, P. W., Geffroy, V., Adam-Blondon, A. F., Shirmohamadali, A., Johnson, W. C., et al. (1998). Towards an integrated linkage map of common bean. 4. Development of a core linkage map and alignment of RFLP maps. *Theor. Appl. Genet*. 97, 847–856.

Galeano, C. H., Fernandez, A. C., Gomez, M., and Blair, M. W. (2009). Single strand conformation polymorphism based SNP and Indel markers for genetic mapping and synteny analysis of common bean (*Phaseolus vulgaris* L.). *BMC Genomics*10:629.

Gepts, P., and Bliss, F. A. (1986). Phaseolin variability among wild and cultivated common beans (*Phaseolus vulgaris*) from Colo Mbia. *Econ. Bot*. 40, 469–478.

Ghaderi, A., Hosfield, G. L., Adams, M. W., and Uebersax, M. A. (1984). Variability in culinary quality, component interrelationships, and breeding implications in navy and pinto beans. *J. Am. Soc. Horticult. Sci*. 109, 85–90.

Gomez, O. J., Blair, M. W., Frankow-Lindberg, B. E., and Gullberg, U. (2004). Molecular and phenotypic diversity of common bean landraces from Nicaragua. *Crop Sci*. 44, 1412–1418.

Gonzalez, A., Wong, A., Delgado-Salinas, A., Papa, R., and Gepts, P. (2005). Assessment of inter simple sequence repeat markers to differentiate sympatric wild and domesticated populations of common bean. *Crop Sci*. 45, 606–615.

Hartl, D. L., and Clark, A. G. (1997). *Principles of Population Genetics*. Sunderland, MA: Sinauer Associates.

Hayashi, K., Yoshida, H., and Ashikawa, I. (2006). Development of PCR-based allele-specific and InDel marker sets for nine rice blast resistance genes. *Theor. Appl. Genet*. 113, 251–260.

Henegariu, O., Heerema, N. A., Dlouhy, S. R., Vance, G. H., and Vogt, P. H. (1997). Multiplex PCR: Critical parameters and step-by-step protocol. *Biotechniques* 23, 504–511.

Hosfield, G. L., Uebersax, M. A., and Occena, L. G. (2000). "Technological and genetic improvements in dry bean quality and utilization," in *Bean Reseach, Production and Utilization. Proceeding of Idaho Bean Workshop*, ed S. P. Singh (Moscow: University of Idaho), 135–152.

Hou, X. H., Li, L. C., Peng, Z. Y., Wei, B. Y., Tang, S. J., Ding, M. Y., et al. (2010). A platform of high-density INDEL/CAPS markers for map-based cloning in Arabidopsis. *Plant J*. 63, 880–888.

Hyten, D. L., Song, Q. J., Fickus, E. W., Quigley, C. V., Lim, J. S., Choi, I. Y., et al. (2010). High-throughput SNP discovery and assay development in common bean. *BMC Genomics* 11:475.

Jander, G., Norris, S. R., Rounsley, S. D., Bush, D. F., Levin, I. M., and Last, R. L. (2002). Arabidopsis map-based cloning in the post-genome era. *Plant Physiol*. 129, 440–450.

Karaiskou, N., and Primmer, C. (2008). PCR multiplexing for maximising genetic analyses with limited DNA samples: an example in the collared flycatcher, *Ficedula albicollis*. *Ann. Zool. Fenn*. 45, 478–482.

Kelly, J. D., Gepts, P., Miklas, P. N., and Coyne, D. P. (2003). Tagging and mapping of genes and QTL and molecular marker-assisted selection for traits of economic importance in bean and cowpea. *Field Crops Res*. 82, 135–154.

Khairallah, M. M., Adams, M. W., and Sears, B. B. (1990). Mitochondrial DNA polymorphisms of Malawian bean lines: further evidence for 2 major gene pools. *Theor. Appl. Genet*. 80, 753–761.

Khairallah, M. M., Sears, B. B., and Adams, M. W. (1992). Mitochondrial restriction fragment length polymorphisms in wild *Phaseolus vulgaris* L: insights on the domestication of the common bean. *Theor. Appl. Genet*. 84, 915–922.

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., et al. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25, 2283–2285.

Koenig, R., and Gepts, P. (1989). Allozyme diversity in wild *Phaseolus vulgaris*: further evidence for 2 Major centers of genetic diversity. *Theor. Appl. Genet*. 78, 809–817.

Koinange, E. M. K., and Gepts, P. (1992). Hybrid weakness in wild *Phaseolus vulgaris* L. *J. Hered*. 83, 135–139.

Lark, K. G., Chase, K., Adler, F., Mansur, L. M., and Orf, J. H. (1995). Interactions between quantitative trait loci in soybean in which trait variation at one locus is conditional upon a specific allele at another. *Proc. Natl. Acad. Sci. U.S.A.* 92, 4656–4660.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Li, Z. K., Pinson, S. R. M., Park, W. D., Paterson, A. H., and Stansel, J. W. (1997). Epistasis for three grain yield components in rice (*Oryza sativa* L). *Genetics* 145, 453–465.

Li, Z. K., Pinson, S. R. M., Stansel, J. W., and Paterson, A. H. (1998). Genetic dissection of the source-sink relationship affecting fecundity and yield in rice (*Oryza sativa* L.). *Mol. Breed*. 4, 419–426.

Liu, K. J., and Muse, S. V. (2005). PowerMarker: an integrated analysis environment for genetic marker analysis.*Bioinformatics* 21, 2128–2129.

Mamidi, S., Rossi, M., Annam, D., Moghaddam, S., Lee, R., Papa, R., et al. (2011). Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. *Funct. Plant Biol*. 38, 953–967.

Mamidi, S., Rossi, M., Moghaddam, S. M., Annam, D., Lee, R., Papa, R., et al. (2013). Demographic factors shaped diversity in the two gene pools of wild common bean *Phaseolus vulgaris* L. *Heredity* 110, 267–276.

McClean, P. E., Myers, J. R., and Hammond, J. J. (1993). Coefficient of parentage and cluster analysis of North American dry bean cultivars. *Crop Sci*. 33, 190–197.

Mensack, M. M., Fitzgerald, V. K., Ryan, E. P., Lewis, M. R., Thompson, H. J., and Brick, M. A. (2010). Evaluation of diversity among common beans (*Phaseolus vulgaris* L.) from two centers of domestication using 'omics' technologies. *BMC Genomics* 11:686.

Miklas, P. N. (2000). "Use of *Phaseolus* germplasm in breeding pinto, great northern, pink, and red bean for the Pacific Northwest and intermountain region," in *Bean Research, Production and Utilization. Proceeding of the Idaho Bean Workshop*, ed S. P. Singh (Moscow: University of Idaho), 13–29.

Myers, J. R. (2000). "Tomorrow's snap bean cultivars," in *Bean Reseach, Production and Utilization. Proceeding Idaho Bean Workshop*, ed S. P. Singh (Moscow: University of Idaho), 39–51.

Nelson, J. C. (1997). QGENE: software for marker-based genomic analysis and breeding. *Mol. Breed*. 3, 239–245.

Ollitrault, F., Terol, J., Martin, A. A., Pina, J. A., Navarro, L., Talon, M., et al. (2012). Development of indel markers from*Citrus clementina* (Rutaceae) BAC-end sequences and interspecific transferability in *Citrus*. *Am. J. Bot*. 99, E268–E273.

Pacurar, D. I., Pacurar, M. L., Street, N., Bussell, J. D., Pop, T. I., Gutierrez, L., et al. (2012). A collection of INDEL markers for map-based cloning in seven Arabidopsis accessions. *J. Exp. Bot*. 63, 2491–2501.

Salathia, N., Lee, H. N., Sangster, T. A., Morneau, K., Landry, C. R., Schellenberg, K., et al. (2007). Indel arrays: an affordable alternative for genotyping. *Plant J*. 51, 727–737.

Schlueter, J. A., Goicoechea, J. L., Collura, K., Gill, N., Lin, J. Y., Yu, Y., et al. (2008). BAC-end sequence analysis and a draft physical map of the common bean (*Phaseolus vulgaris* L.) genome. *Trop. Plant Biol*. 1, 40–48.

Schmutz, J., McClean, P., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., et al. (in press). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet*.

Silbernagel, M. J., and Hannan, R. M. (1988). "Utilization of genetic resources in the development of commercial bean cultivars in the USA," in *Genetic Resources of Phaseolus Beans*, ed P. Gepts (Dordrecht: Kluwer), 561–596.

Silbernagel, M. J., and Hannan, R. M. (1992). "Use of plant introductions to develop U.S. bean cultivars," in *Use of Plant Introductions in Cultivar Development, Part 2*, eds H. Shands and L. E. Weisner (Madison, WI: CSSA Special Publication), 1–8.

Singh, S. P. (1992). "Common bean improvement in the tropics," in *Plant Breeding Reviews*, ed J. Janick (John Wiley and Sons, Inc.), 199–269.

Singh, S. P., Gepts, P., and Debouck, D. G. (1991). Races of common bean (*Phaseolus vulgaris*, Fabaceae). *Econ. Bot*. 45, 379–396.

Snelling, W. M., Chiu, R., Schein, J. E., Hobbs, M., Abbey, C. A., Adelson, D. L., et al. (2007). A physical map of the bovine genome. *Genome Biol*. 8:R165.

Sonnante, G., Stockton, T., Nodari, R. O., Velasquez, V. L. B., and Gepts, P. (1994). Evolution of genetic diversity during the domestication of common bean (*Phaseolus vulgaris* L). *Theor. Appl. Genet*. 89, 629–635.

Steele, K. A., Ogden, R., McEwing, R., Briggs, H., and Gorham, J. (2008). InDel markers distinguish Basmatis from other fragrant rice varieties. *Field Crops Res*. 105, 81–87.

Tanksley, S. D., Ganal, M. W., Prince, J. P., Devicente, M. C., Bonierbale, M. W., Broun, P., et al. (1992). High density molecular linkage maps of the tomato and potato genomes. *Genetics* 132, 1141–1160.

Tanksley, S. D., and Hewitt, J. (1988). Use of molecular markers in breeding for soluble solids content in tomato - a re-examination. *Theor. Appl. Genet*. 75, 811–823.

Vali, U., Brandstrom, M., Johansson, M., and Ellegren, H. (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genet*. 9:8.

Vallejos, C. E., Sakiyama, N. S., and Chase, C. D. (1992). A molecular marker-based linkage map of *Phaseolus Vulgaris* L.*Genetics* 131, 733–740.

Varshney, R. K., Close, T. J., Singh, N. K., Hoisington, D. A., and Cook, D. R. (2009). Orphan legume crops enter the genomics era! *Curr. Opin. Plant Biol*. 12, 202–210.

Vasemagi, A., Gross, R., Palm, D., Paaver, T., and Primmer, C. R. (2010). Discovery and application of insertion-deletion (INDEL) polymorphisms for QTL mapping of early life-history traits in Atlantic salmon. *BMC Genomics* 11:156.

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I., et al. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol*. 17, 1636–1647.

Wang, C. R., Chang, K. C., and Grafton, K. (1988). Canning quality evaluation of pinto and navy beans. *J. Food Sci*. 53, 772–776.

Wei, F., Coe, E., Nelson, W., Bharti, A. K., Engler, F., Butler, E., et al. (2007). Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet*. 3:e123.

Werner, J. E., Endo, T. R., and Gill, B. S. (1992). Toward a cytogenetically based physical map of the wheat genome. *Proc. Natl. Acad. Sci. U.S.A*. 89, 11307–11311.

Xia, Z., Tsubokura, Y., Hoshi, M., Hanawa, M., Yano, C., Okamura, K., et al. (2007). An integrated high-density linkage map of soybean with RFLP, SSR, STS, and AFLP markers using a single F-2 population. *DNA Res*. 14, 257–269.

Yang, H. A., Tao, Y., Zheng, Z. Q., Li, C. D., Sweetingham, M. W., and Howieson, J. G. (2012). Application of next-generation sequencing for rapid marker development in molecular plant breeding: a case study on anthracnose disease resistance in *Lupinus angustifolius* L. *BMC Genomics* 13:318.

You, F. M., Huo, N. X., Gu, Y. Q., Luo, M. C., Ma, Y. Q., Hane, D., et al. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* 9:253.

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*. 18, 821–829.

GENERAL CONCLUSION

Crop improvement requires continuous enhancements in the basic genomics, genetics, and biological knowledge as well as the technical tools necessary to implement the existing knowledge in applied programs. Knowing the population diversity and genetic architecture underlying traits of interest can facilitate and accelerate breeding. The availability of whole genome reference sequence of common bean has changed the direction of research towards high throughput analysis which leads to more accurate genomic discoveries, and facilitates the development of tools to generate new data or employ the existing data.

GWAS is a successful approach to discover new genomic factors associated with a trait. Our first objective was to use GWAS on a population of diverse bean genotypes, to dissect the genetic architecture of seven important agronomic traits in common bean. First, we evaluated the extent of linkage disequilibrium (LD) in the entire population and its constitutive subpopulations because GWAS is an approach based on LD. The results demonstrated that the extent of measurable LD depends on: the specific population (genotypes) itself, population structure and relatedness, the chromosome and the region on the chromosome, and the number of informative markers used to measure the LD. Inter- and intra-chromosomal LD play an important role in the interpretation of the GWAS since they might lead to spurious results. On the other hand, LD can indicate biological relevance between genomic regions identified for a single or multiple trait(s).

Interpretation of GWAS result is also affected by the significant p-value cutoff choice. Some GWAS peaks were biologically relevant to the trait based on *a priori* information but their p-value didn't pass our stringent cutoff of 0.1 percentile tail of the empirical distribution of the p-values but passed a less stringent criteria of 1 percentile. Therefore, a combination of evidences

such as *a priori* biological knowledge, previous discoveries using different methods like QTL analysis, and the results of current analyses were collectively used to interpret the data.

The phenotypic and genotypic analysis of agronomic traits revealed phenotypic correlation and overlapping genomic regions for inter-related traits. Moreover, location and subpopulation specific GWAS peaks denote the effect of environment on the genome expression, and the effect of race/market class specific breeding in common bean, respectively.

Market class specific breeding in bean has led to low genetic diversity in each marker class. Hence, as a second objective, we focused on assessing market class specific variation and developing a tool which helps to utilize this existing variation for genomic and breeding studies. Our InDel marker panel is a user-friendly, cost-effective medium throughput tool that can facilitate marker assisted selection (MAS) in breeding programs. Moreover, the InDel markers are distributed uniformly on the bean genome which makes them an appropriate tool for generating new information in any laboratory. For example, they can be employed for phylogenetic and diversity studies, and genetic map constructions.

In brief, this dissertation focuses on the discovering of new genomic knowledge as well as developing user-friendly, cost-effective tools that can generate new information and facilitate the implementation of new and existing genomic discoveries in applied bean improvement programs.

APPENDIX A. TABLES AND FIGURES

Tables and Figures begin on the following page.

Table A.1. Candidate genes in the 200Kb surrounding region of significant markers. The negative sign in column six indicates that the marker is downstream of the candidate gene and no sign indicates upstream positon of the marker. Markers that fall in the one percentile tail of the empirical distribution of p-value are highlighted in gray. The R^2_{LR} values are from the entire population and across all the locations analysis unless only a subpopulation/location is specified in the 10th column.

SNP	Pv	SNP position (bp)	-log10 (p-value)	Bean candidate gene	Marker distance from candidate gene (bp)	Arabidopsis gene model	Arabidopsis gene symbol	Arabidopsis annotation	Population/ location in which is significant	R^2_{LR} in the best model
Days to flower										
m23249	1	8,117,979	7.5	Phvul.001G064200	-96,502	AT1G43850	SEU	SEUSS transcriptional co-regulator	All/DJ/CO/NE	0.12
m24109	1	9,569,437	5.8	Phvul.001G071900	0	AT1G21700	ATSWI3C,CHB4,SWIC	SWITCH/sucrose non fermenting 3C	All/DJ/CO/MI/NE	0.09
m32210	1	15,827,862	8.4	Phvul.001G087900	5,006	AT5G14010	KNU	C2H2 and C2HC zinc fingers superfamily protein	All/MA/CO/MI/ND/MI -MA/ND-MA	0.14
m3673	1	16,748,236	5.6	Phvul.001G089900	32,915	AT1G66350	RGL,RGL1	RGA-like 1	ND-MA	0.19
m4911	1	20,164,029	7.7	Phvul.001G094300	163,704	AT5G14170	CHC1 ,AtBAF60, AtSWP73B	SWIB/MDM2 domain superfamily protein	All/DJ/MI/ND/CO	0.13
m18033	1	45,815,561	5	Phvul.001G192200	-6,809	AT1G12910	ATAN11,LWD1	Transducin/WD40 repeat-like superfamily protein	MI/ MI-DJ	0.08
m18033	1	45,815,561	5	Phvul.001G192300	0	AT3G11540	SPY	Tetratricopeptide repeat (TPR)-like superfamily protein	MI/MI-DJ	0.08
Days to maturity										
m8942	7	1,584,124	5.2	Phvul.007G022900	-4,351	AT3G18990	REM39,VRN1	AP2/B3-like transcriptional factor family protein	CO	0.07
m11196	7	50,527,356	4.1	Phvul.007G267100	0	AT3G05690.1	ATHAP2B,HAP2B,NF-YA2,UNE8	nuclear factor Y, subunit A2	All	0.09

Table A.1. Candidate genes in the 200Kb surrounding region of significant markers (continued).

SNP	Pv	SNP position (bp)	-log10 (p-value)	Bean candidate gene	Marker distance from candidate gene (bp)	Arabidopsis gene model	Arabidopsis gene symbol	Arabidopsis annotation	Population/ location in which is significant	R^2_{LR} in the best model
m14184	8	58,403,271	4.5	Phvul.008G275200	24,472	AT5G16320	FRL1	FRIGIDA like 1	ND	0.07
m17078	10	4,792,292	4.3	Phvul.010G032600	-3,625	AT2G43280	-	Far-red impaired responsive (FAR1) family protein	All/ND	0.10
m19177	11	4,315,986	3.8	Phvul.011G050300	0	AT1G50030.1	TOR	target of rapamycin	All/NE	0.10
m19189	11	4,358,407	4	Phvul.011G050800	-5,073	AT3G18990	REM39,VRN1	AP2/B3-like transcriptional factor family protein	All	0.04
m19214	11	4,461,471	3.8	Phvul.011G053300	64,564	AT2G43410	FPA	RNA binding	All	0.09
m19532	11	7,396,865	3.6	Phvul.011G081100	198,718	AT5G45890.1	SAG12	senescence-associated gene 12	DJ/DJ-NE	0.07
Growth habit: with determinates										
m31418	1	36,575,161	7.4	Phvul.001G128800	57,195	AT2G41510	ATCKX1,CKX1	cytokinin oxidase/dehydrogenase 1	All/DJ/CO/MI/DJ-CO/DJ-MI	0.12
m17218	1	42,857,105	8.6	Phvul.001G167200	0	AT1G61040	VIP5	plus-3 domain-containing protein	All/DJ/CO/MI/DJ-MI	0.15
m17978	1	45,734,467	10	Phvul.001G192200	71,407	AT1G12910	ATAN11,LWD1	Transducin/WD40 repeat-like superfamily protein	All/MA/CO/MI/DJ-MI/MA-CO/MA-MI	0.18
m17978	1	45,734,467	10	Phvul.001G192300	80,722	AT3G11540	SPY	Tetratricopeptide repeat (TPR)-like superfamily protein	All/MA/CO/MI/MA-CO/MA-MI	0.18
m17978	1	45,734,467	10	Phvul.001G189200	-171,141	AT5G03840	TFL-1,TFL1	PEBP (phosphatidylethanolamine-binding protein) family protein	All/MA/CO/MI/MA-CO/MA-MI	0.18
m17978	1	45,734,467	10	Phvul.001G191500	0	AT2G36200	-	P-loop containing nucleoside triphosphate hydrolases superfamily protein	All/MA/CO/MI/MA-CO/MA-MI	0.18

Table A.1. Candidate genes in the 200Kb surrounding region of significant markers (continued).

SNP	Pv	SNP position (bp)	-log10 (p-value)	Bean candidate gene	Marker distance from candidate gene (bp)	Arabidopsis gene model	Arabidopsis gene symbol	Arabidopsis annotation	Population/ location in which is significant	R^2_{LR} in the best model
Growth habit: no determinates										
m7418	6	21,498,732	5	Phvul.006G097400	-4,164	AT3G05040	HST,HST1	ARM repeat superfamily protein	MI	0.07
m20650	11	43,284,551	7.3	Phvul.011G164800	0	AT1G53160	SPL4	squamosa promoter-binding protein-like 4	All/CO/MI/DJ/DJ-CO/DJ-MI	0.13
Lodging										
m19820	1	47,939,346	4.6	Phvul.001G215000	-120,836	AT1G02400	ATGA2OX4,ATGA2OX6,DTA1,GA2OX6	gibberellin 2-oxidase 6	All/DJ/ CO-DJ	0.07
m21909	2	1,693,424	4.6	Phvul.002G015500	31,332	AT5G67260	CYCD3;2	CYCLIN D3;2	All/CO	0.07
m10611	7	45,183,105	5.7	Phvul.007G212200	-87,908	AT1G02400	ATGA2OX4,ATGA2OX6,DTA1,GA2OX6	gibberellin 2-oxidase 6	All/CO/NE	0.11
m10685	7	46,112,355	10.1	Phvul.007G221700	0	AT2G37150	-	RING/U-box superfamily protein	All/DJ/CO/MI/NE/CO-DJ/MI-DJ/NE-DJ	0.21
m10689	7	46,131,994	8.7	Phvul.007G221800	0	AT2G33170	-	Leucine-rich repeat receptor-like protein kinase family protein	All/DJ/CO/MI/NE/CO-DJ/MI-DJ/NE-DJ	0.18
m10969	7	48,630,699	4.7	Phvul.007G246700	0	AT4G02330	ATPMEPCRB,AtPME41	Plant invertase/pectin methylesterase inhibitor superfamily	All/DJ/CO/NE/CO-DJ/NE-DJ	0.08
m26399	8	12,646,413	3.9	Phvul.008G108600	-22,921	AT2G38050	ATDET2,DET2,DWF6	3-oxo-5-alpha-steroid 4-dehydrogenase family protein	MA/NE-MA/CO-MA	-
Canopy height										
m10611	7	45,183,105	7.3	Phvul.007G212200	-87,908	AT1G02400	ATGA2OX4,ATGA2OX6,DTA1,GA2OX6	gibberellin 2-oxidase 6	All/CO/MI/NE/NE-MA	0.13

Table A.1. Candidate genes in the 200Kb surrounding region of significant markers (continued).

SNP	Pv	SNP position (bp)	-log10 (p-value)	Bean candidate gene	Marker distance from candidate gene (bp)	Arabidopsis gene model	Arabidopsis gene symbol	Arabidopsis annotation	Population/ location in which is significant	R^2_{LR} in the best model
m10689	7	46,131,994	6.3	Phvul.007G221800	0	AT2G33170	-	Leucine-rich repeat receptor-like protein kinase family protein	All/ CO/MI/NE/MI-DJ/MI-MA/NE-DJ/NE-MA	0.11
m10969	7	48,630,699	5.5	Phvul.007G246700	0	AT4G02330	ATPMEPCRB, AtPME41	Plant invertase/pectin methylesterase inhibitor superfamily	MI/MI-DJ	-
m16972	10	3,143,074	5.3	Phvul.010G021200	38,070	AT3G25070	RIN4	RPM1 interacting protein 4	MI	0.09
m27790	11	43,312,966	4.3	Phvul.011G164800	-28,019	AT3G60030	SPL4	squamosa promoter-binding protein-like 4	NE	0.06
					Seed weight					
m2359	3	45,349,067	4.8	Phvul.003G232600	149,698	AT4G37750	ANT,CKC,CKC1,DRG	Integrase-type DNA-binding superfamily protein	ND-MA	0.25
m2535	3	48,030,868	5.71	Phvul.003G253100	79,755	AT3G50660	CLM,CYP90B1,DWF4,PSC1,SAV1,SNP2	Cytochrome P450 superfamily protein	DJ/DJ-CO/DJ-ND/DJ-NE	-
m7138	6	18,954,461	5.8	Phvul.006G069300	-130,265	AT3G47340	ASN1,AT-ASN1,DIN6	glutamine-dependent asparagine synthase 1	NE	0.10
m7216	6	19,528,741	6.1	Phvul.006G077400	80,558	AT5G48670	AGL80,FEM111	AGAMOUS-like 80	ND	0.12
m9513	7	8,931,664	5.2	Phvul.007G088200	-33,685	AT1G22710	ATSUC2,SUC2,SUT1	sucrose-proton symporter 2	ND-MA	0.29
m13132	8	43,536,446	5	Phvul.008G168000	0	AT1G77760	GNR1,NIA1,NR1	nitrate reductase 1	DJ/ND-DJ	0.11
m16957	10	2,756,229	7.4	Phvul.010G017600	22,158	AT4G25000	AMY1,ATAMY1	alpha-amylase-like	All/DJ/NE/ND-DJ	0.18
m17852	10	33,410,692	6.3	Phvul.010G089200	-188021	AT2G45690	ATPEX16,PEX16,SSE,SSE1	shrunken seed protein (SSE1)	All/DJ/NE/ND-DJ/NE-DJ	0.15

Table A.1. Candidate genes in the 200Kb surrounding region of significant markers (continued).

SNP	Pv	SNP position (bp)	-log10 (p-value)	Bean candidate gene	Marker distance from candidate gene (bp)	Arabidopsis gene model	Arabidopsis gene symbol	Arabidopsis annotation	Population/ location in which is significant	R^2_{LR} in the best model
					Seed yield					
m29921	1	9,355,723	4.7	Phvul.001G071500	136,574	AT5G09220	AAP2	amino acid permease 2	All/DJ/NE-DJ	0.11
m28775	1	11,112,061	4.8	Phvul.001G076600	-52,955	AT1G58360	AAP1,NAT2	amino acid permease 1	DJ/NE-DJ	0.09
m28775	1	11,112,061	4.8	Phvul.001G076500	-65,810	AT1G10010	AAP8,ATAAP8	amino acid permease 8	DJ/NE-DJ	0.09
m28775	1	11,112,061	4.8	Phvul.001G077000	117,415	AT5G49630	AAP6	amino acid permease 6	DJ/NE-DJ	0.09
m26475	1	15,530,118	3	Phvul.001G086800	368	AT3G26790	FUS3	AP2/B3-like transcriptional factor family protein	MA	0.03
m2423	3	46,344,501	3.7	Phvul.003G241900	199,576	AT4G36920	AP2,FL1,FLO2	Integrase-type DNA-binding superfamily protein	MI	0.05
m2757	3	50,187,048	4.5	Phvul.003G275800	7,753	AT5G62430	CDF1	cycling DOF factor 1	All/NE	0.06
m2872	3	51,168,538	3.4	Phvul.003G285600	0	AT2G35230	IKU1	VQ motif-containing protein	All	0.03
m2903	3	51,435,224	4.9	Phvul.003G288500	11,868	AT5G10510	AIL6,PLT3	AINTEGUMENTA-like 6	All	0.06
m2912	3	51,503,017	4.6	Phvul.003G289100	-10,829	AT5G10480	PAS2,PEP	Protein-tyrosine phosphatase-like, PTPLA	All	0.05
m7118	6	18,689,315	3.5	Phvul.006G066900	-139,162	ATCG00500	-	acetyl-CoA carboxylase carboxyl transferase subunit beta	MI	0.04
m7118	6	18,954,461	3.5	Phvul.006G069300	130,475	AT3G47340	ASN1,AT-ASN1,DIN6	glutamine-dependent asparagine synthase 1	MI	0.04

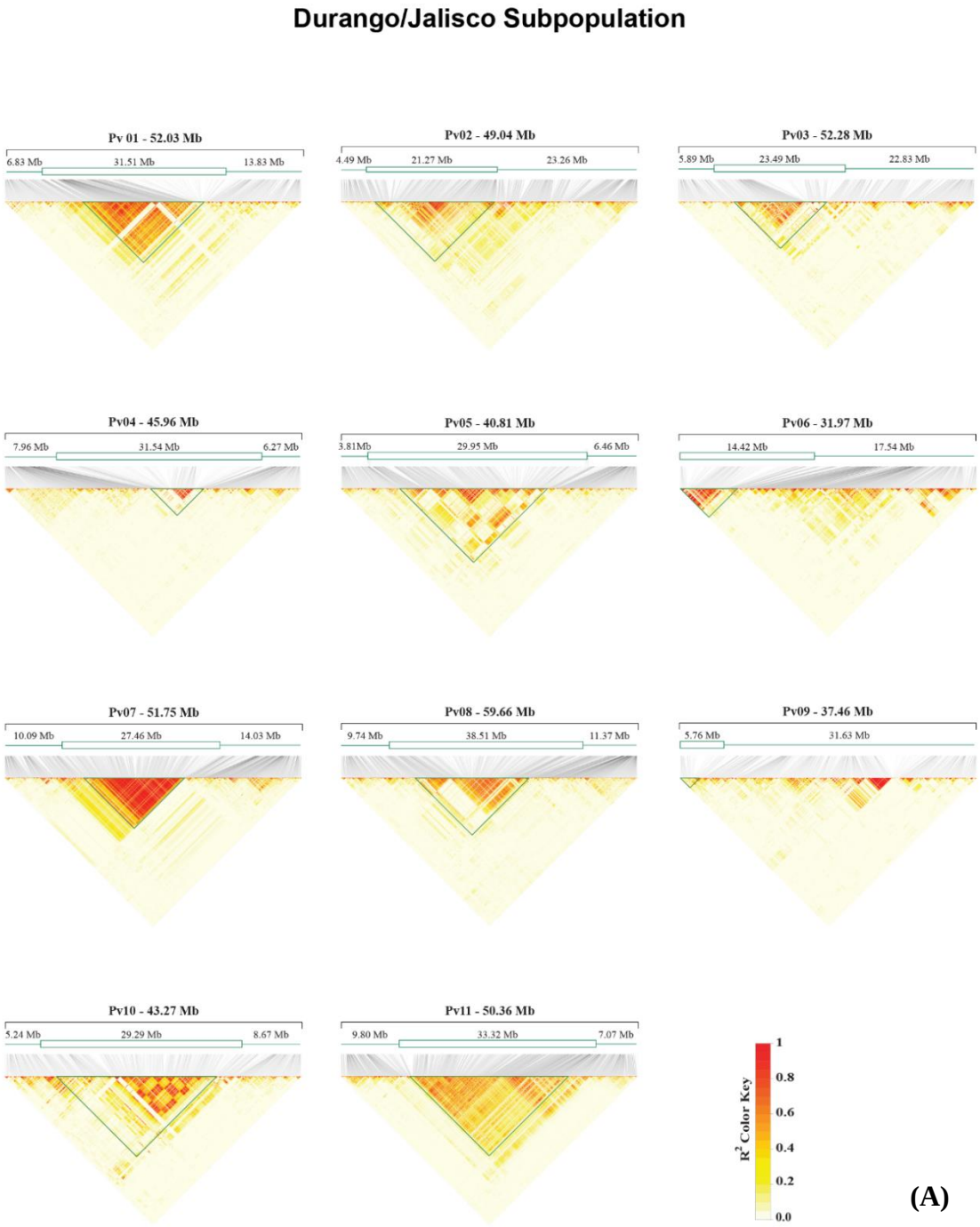


Figure A.1. LD heat maps of 11 chromosomes in race DJ (A) and race MA (B) after controlling for population relatedness. The green lines denote the boundaries of the pericentromeric region.

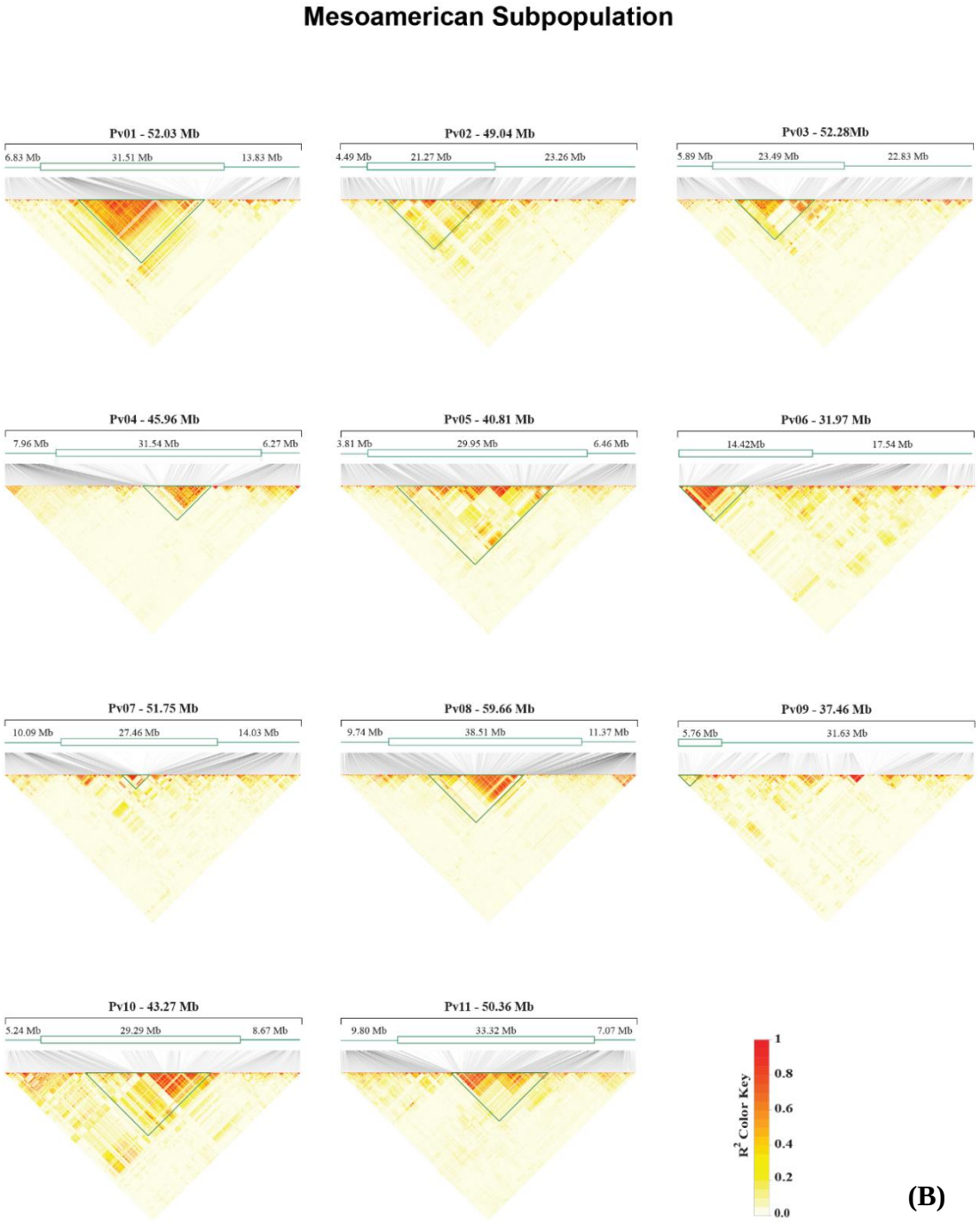


Figure A.1. LD heat maps of 11 chromosomes in race DJ (A) and race MA (B) after controlling for population relatedness (continued)

(A)

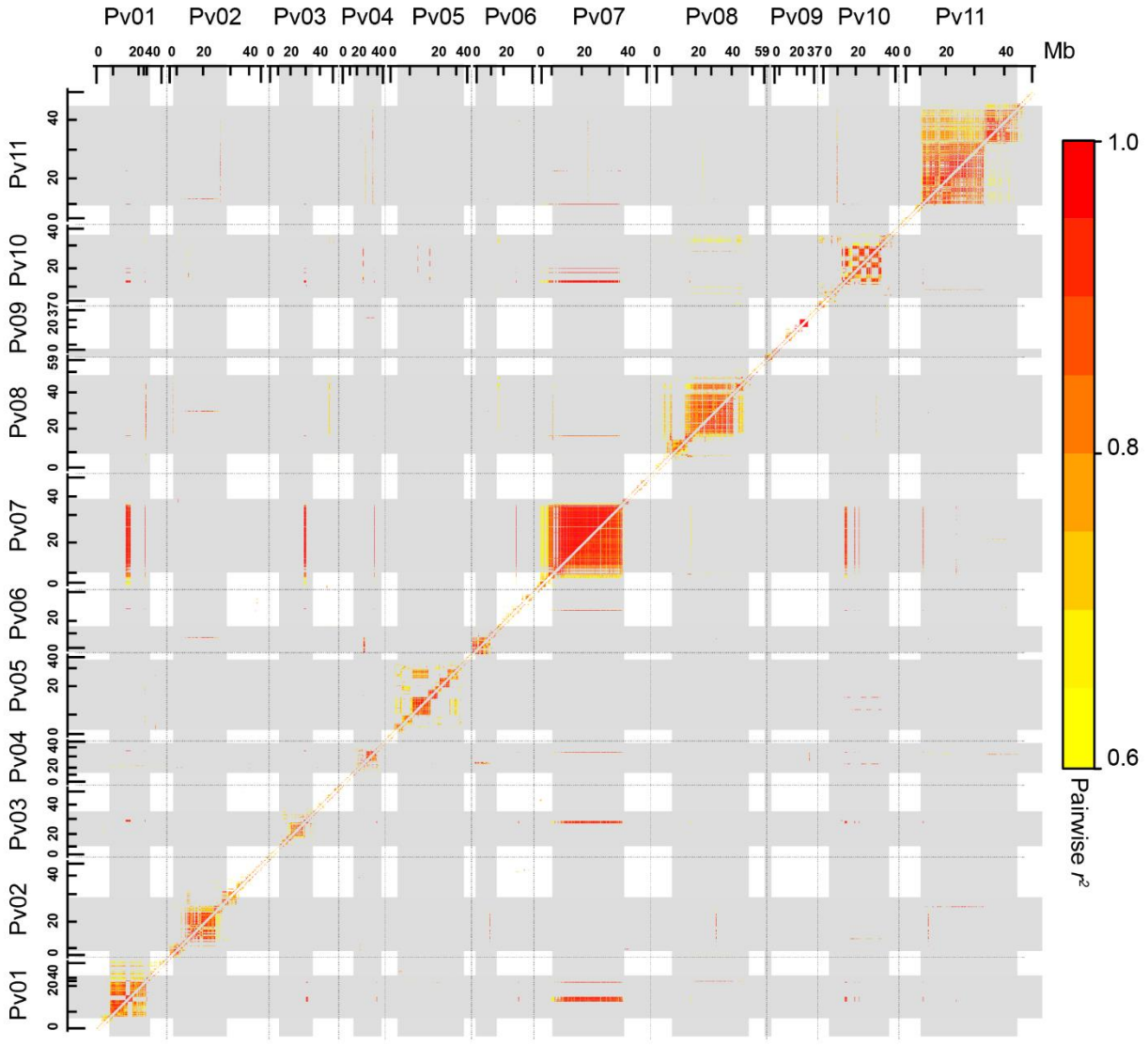


Figure A.2. Genome wide LD heat map in races A) Durango/Jalisco and B) Mesoamerican. Data above the diagonal represent the null model, and data below the diagonal image represents the model that accounts for population relatedness. Markers every 50Kb were used and only pairwise r2 > 0.6 are shown. The grey rectangular show the pericentromeric regions. Grey dashed lines define the chromosome boundaries.

(B)

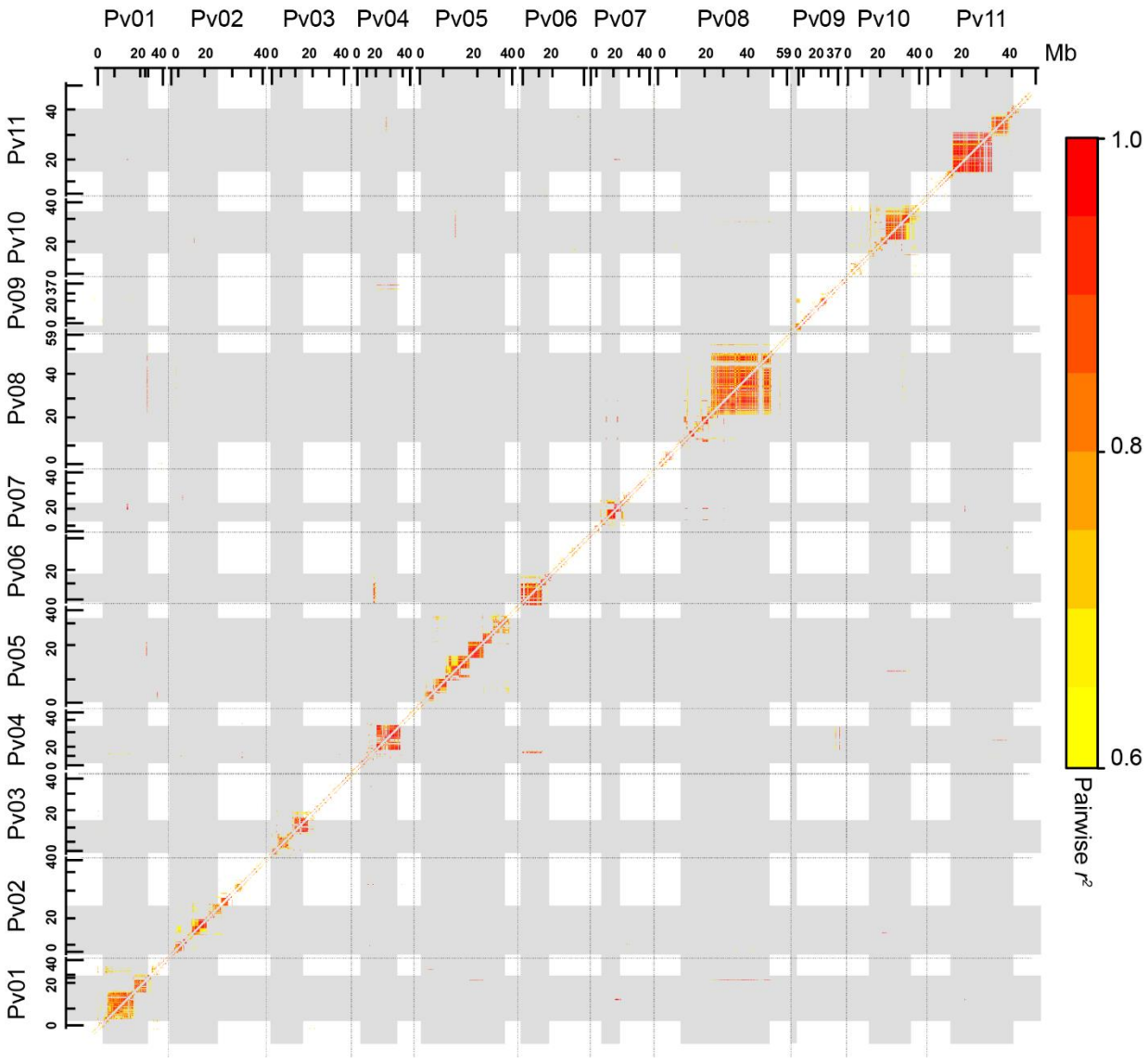


Figure A.2. Genome wide LD heat map in races A) Durango/Jalisco and B) Mesoamerican (continued).

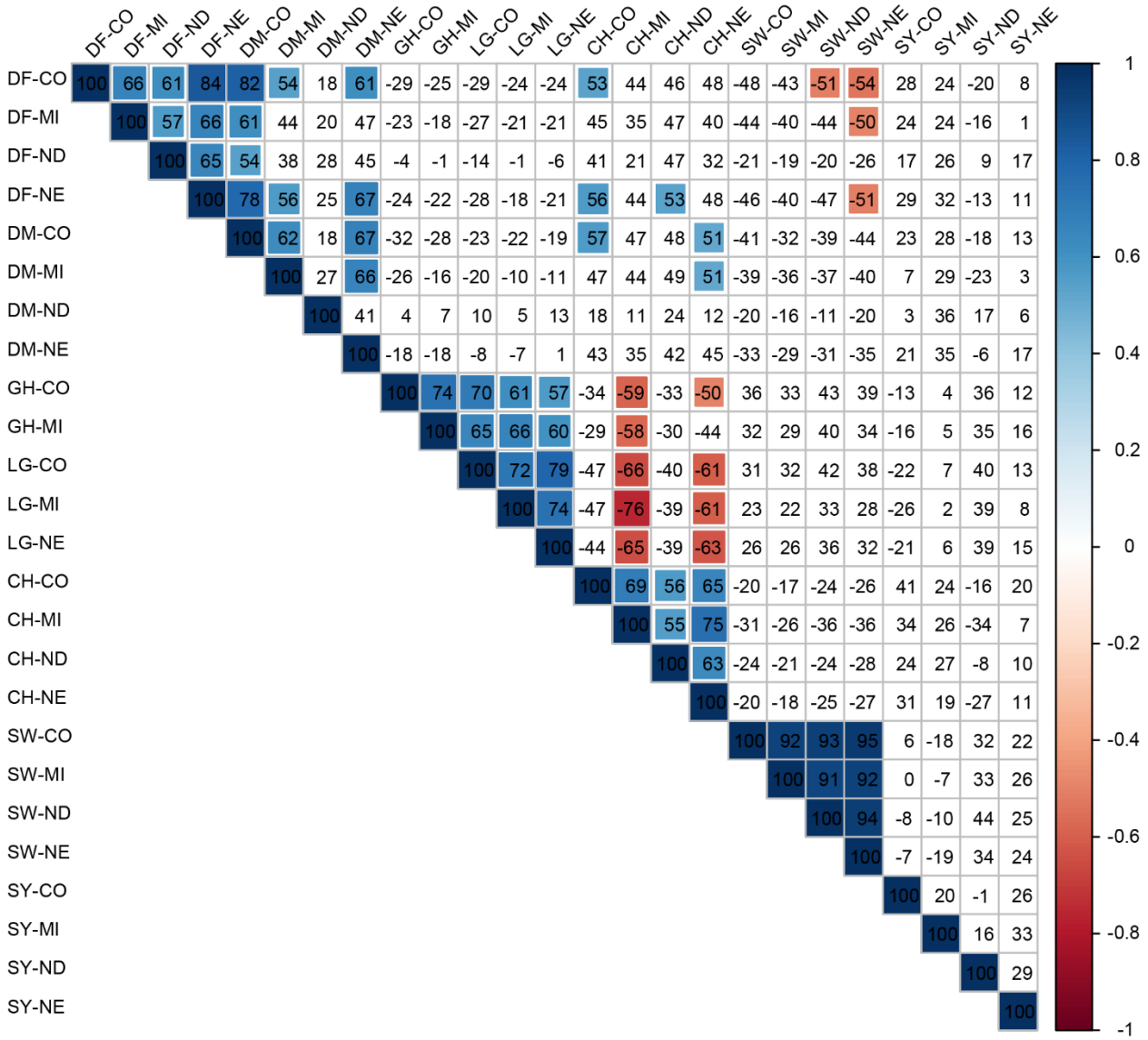
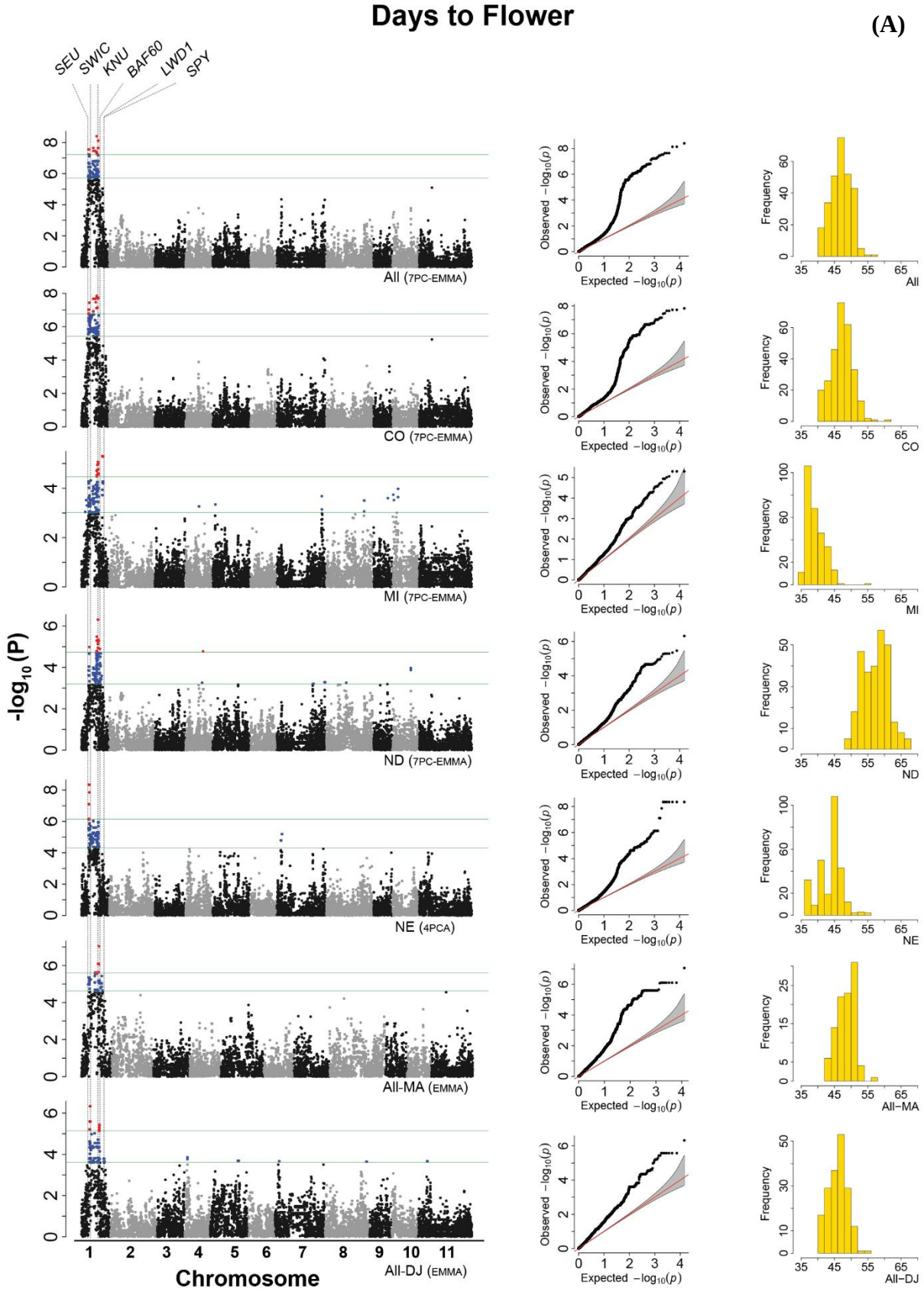


Figure A.3. Correlation between traits and locations based on adjusted means. Abbreviations: DF (days to flower), DM (days to maturity), GH (growth habit-with determinants), LG (lodging), CH (canopy height), SW (seed weight), SY (seed yield), CO (Colorado), MI (Michigan), ND (North Dakota), NE (Nebraska).

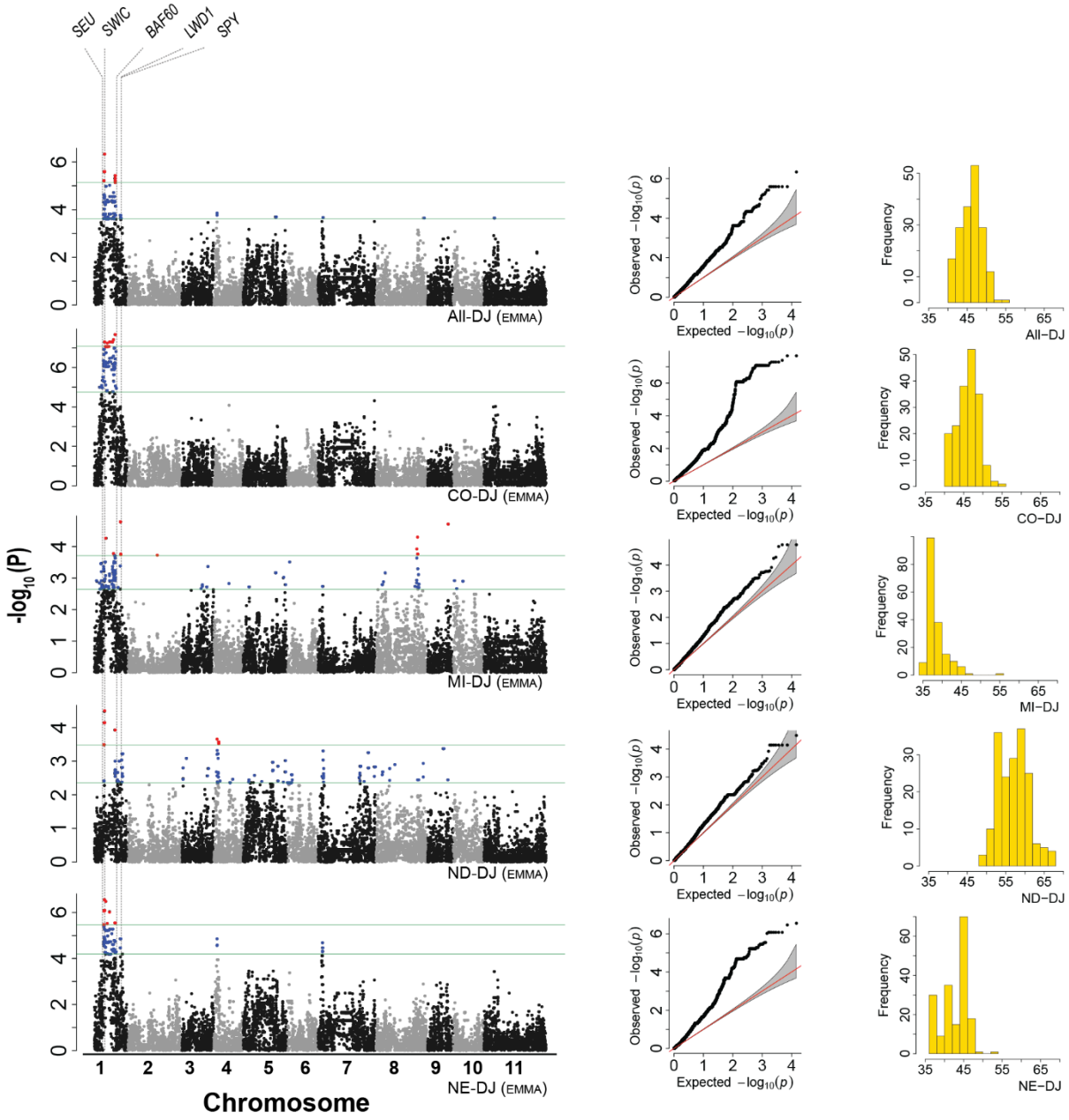
APPENDIX B. PHENOTYPE HISTOGRAM, MANHATTAN PLOTS, AND QQ PLOTS OF THE BEST MODELS FOR SEVEN AGRONOMIC TRAITS

Item begins on the following page. Candidate genes in the 200Kb surrounding region a GWAS peak are shown with vertical dashed lines. SNPs highlighted in red fall in the 0.1% tail of the empirical distribution of the p-values. SNPs highlighted in blue fall in the one percentile tail of the empirical distribution of the p-values. The two horizontal green lines in the Manhattan plots represent the –log10 of cut-off value for these two significant levels. The best model for each analysis is written in parenthesis under the Manhattan plots. Abbreviations: All (across all the locations), CO (Colorado), MI (Michigan), ND (North Dakota), NE (Nebraska), DJ (Durango/Jalisco), MA (Mesoamerican).


Days to Flower
(A)
SEU
SWIC
KNU
BAF60
LWD1
SPY
All (7PC-EMMA)
CO (7PC-EMMA)
MI (7PC-EMMA)
ND (7PC-EMMA)
NE (4PCA)
All-MA (EMMA)
All-DJ (EMMA)
-log10(P)
Chromosome
Observed −log10(p)
Expected −log10(p)
Frequency
All
CO
MI
ND
NE
All−MA
All−DJ

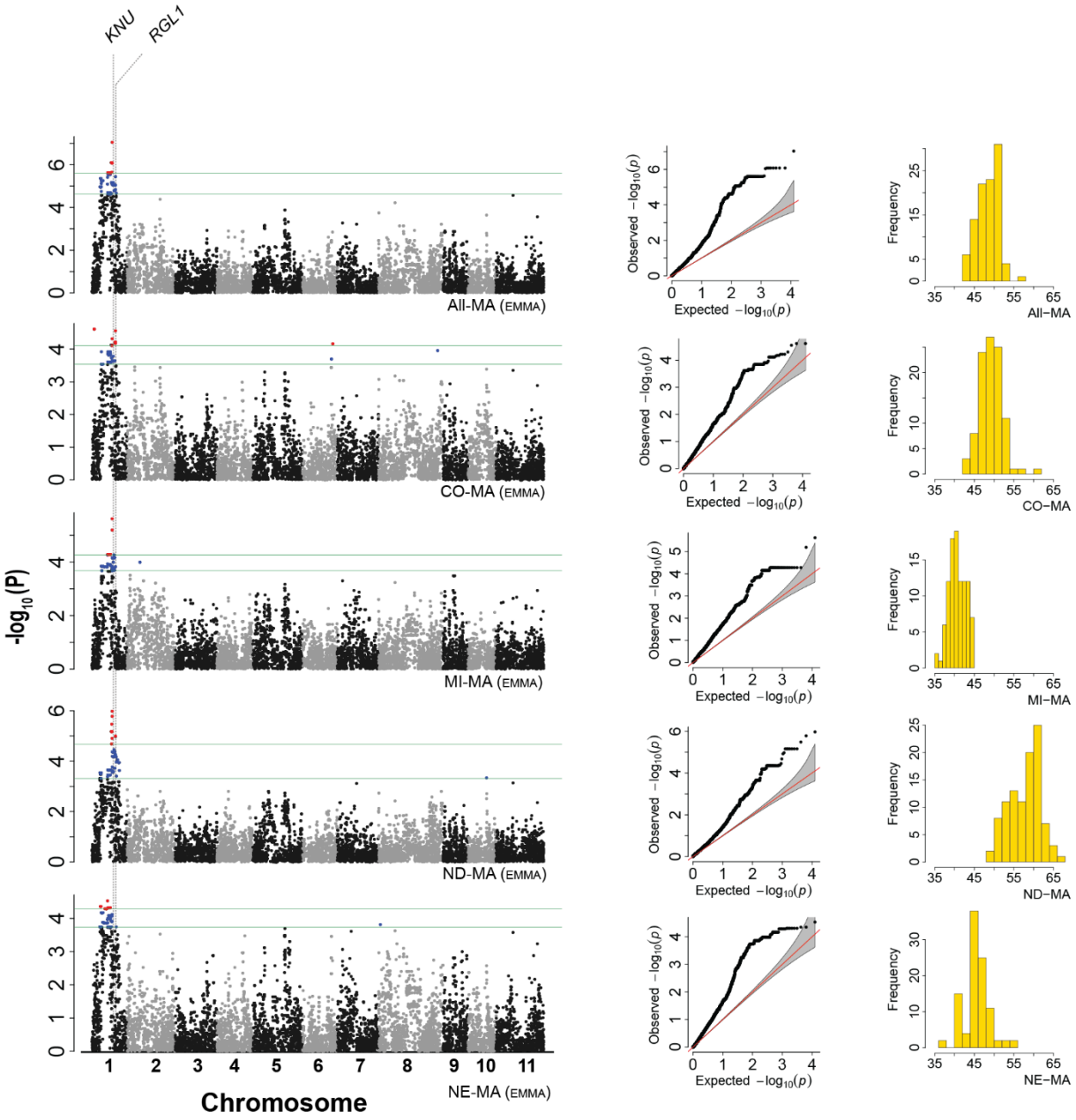
Days to Flower
Durango/Jalisco

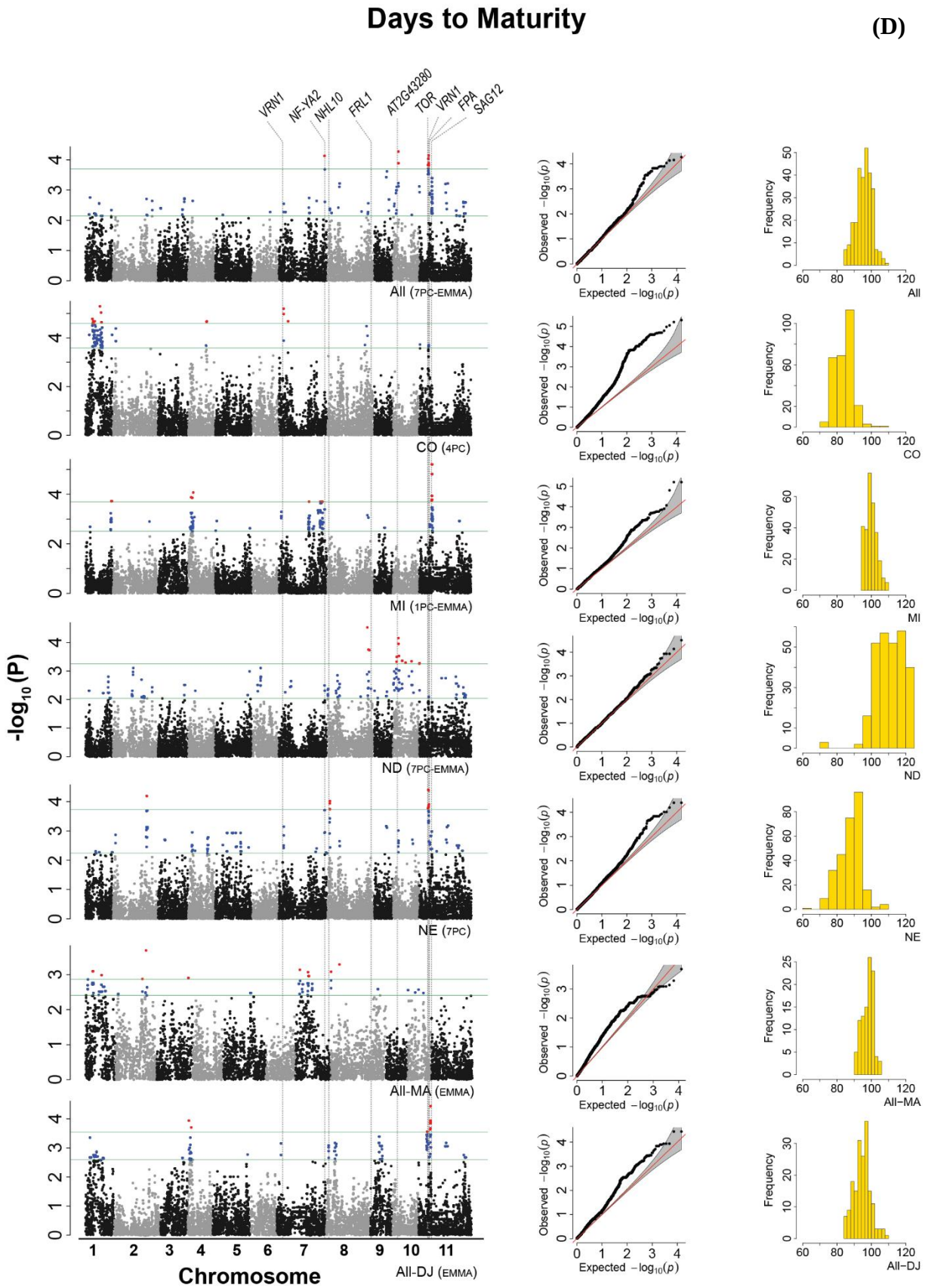
(B)



Days to Flower
Mesoamerican

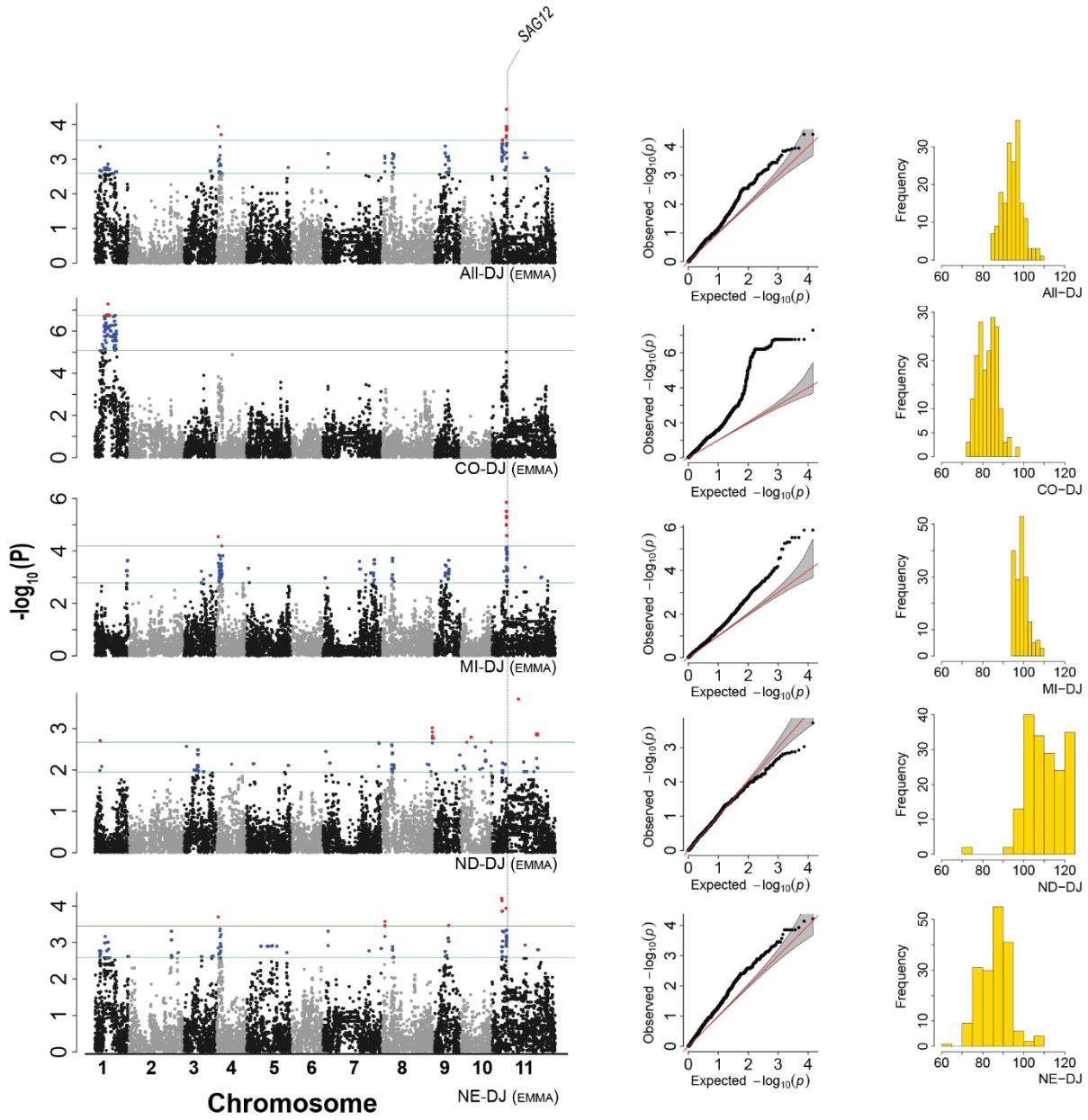
(C)




Days to Maturity
(D)
VRN1
NF-YA2
NHL10
FRL1
AT2G43280
TOR
VRN1
FPA
SAG12
All (7PC-EMMA)
CO (4PC)
MI (1PC-EMMA)
ND (7PC-EMMA)
NE (7PC)
All-MA (EMMA)
All-DJ (EMMA)
-log10 (P)
Chromosome
Observed −log10(p)
Expected −log10(p)
Frequency
All
CO
MI
ND
NE
All−MA
All−DJ

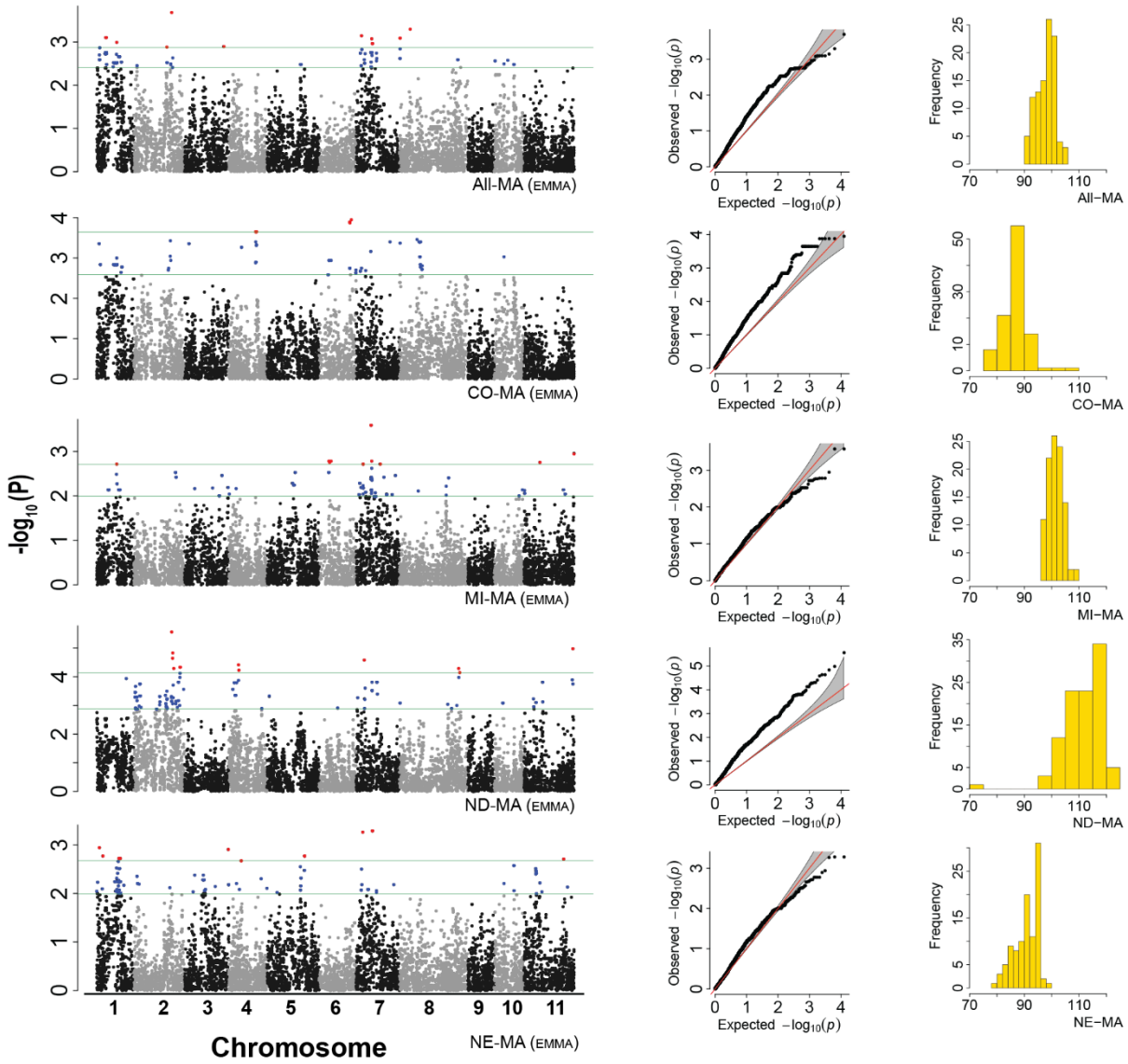
Days to Maturity
Durango/Jalisco

(E)



Days to Maturity
Mesoamerican

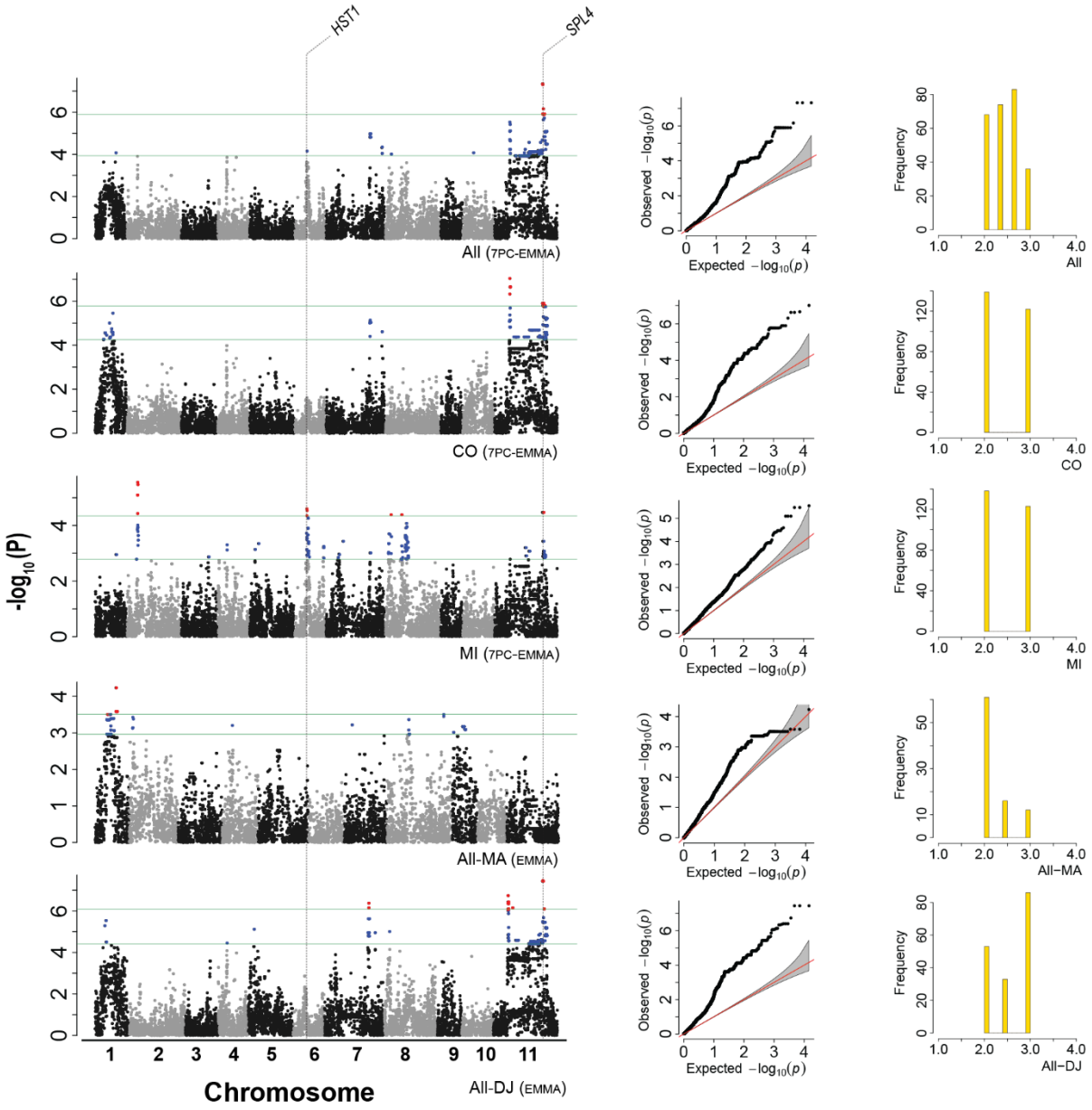
(F)



Growth Habit

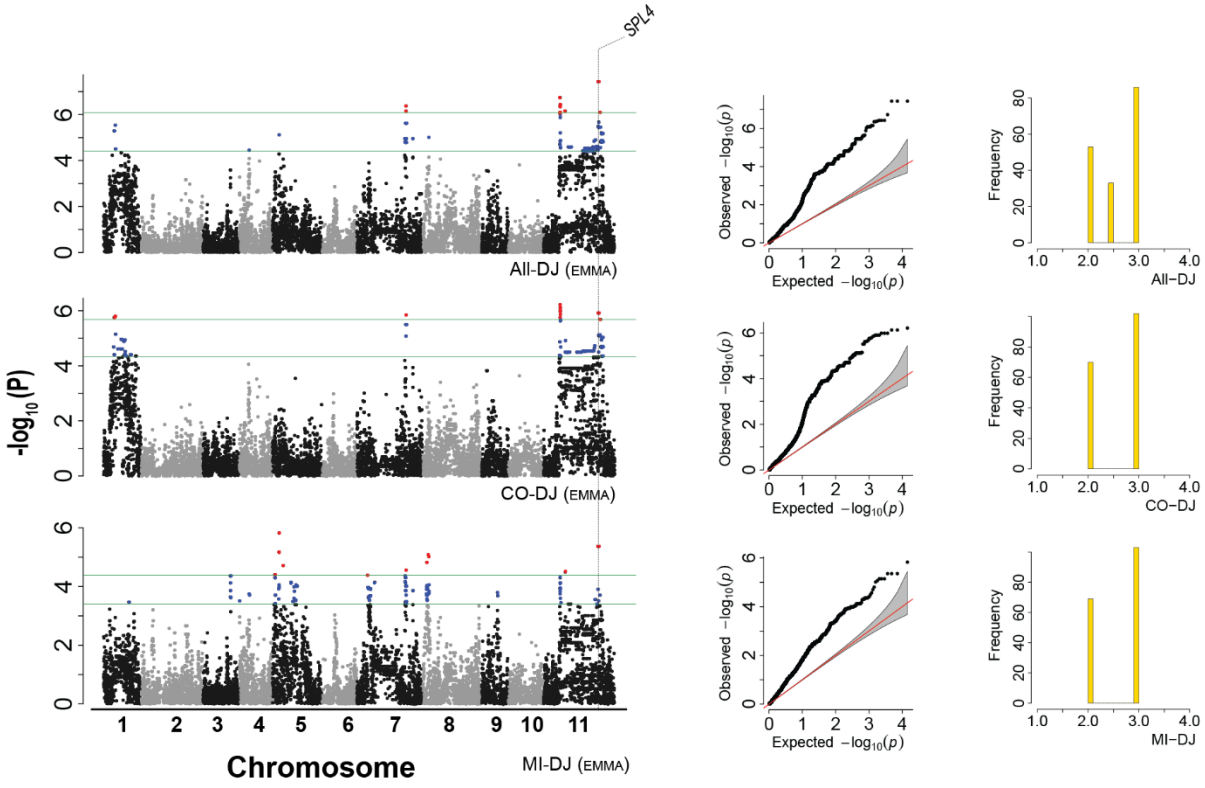
No Determinates

(G



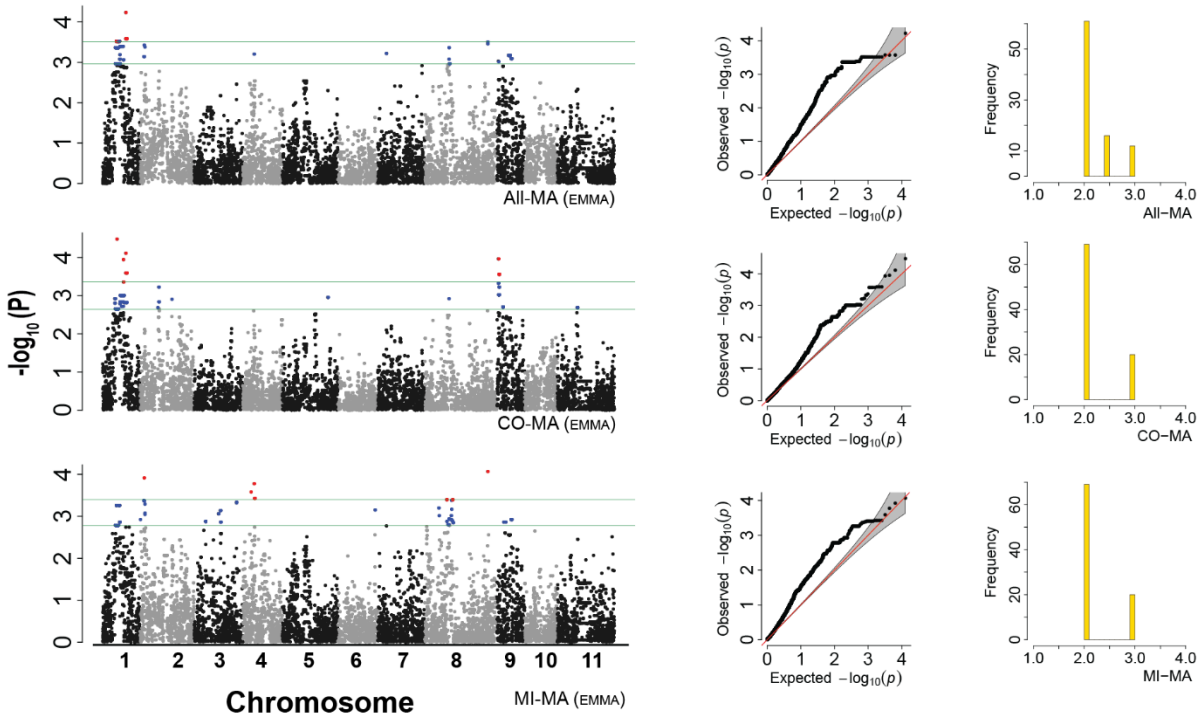
Growth Habit
No Determinate
Durango/Jalosco

(H)



Growth Habit
No Determinates
Mesoamerican

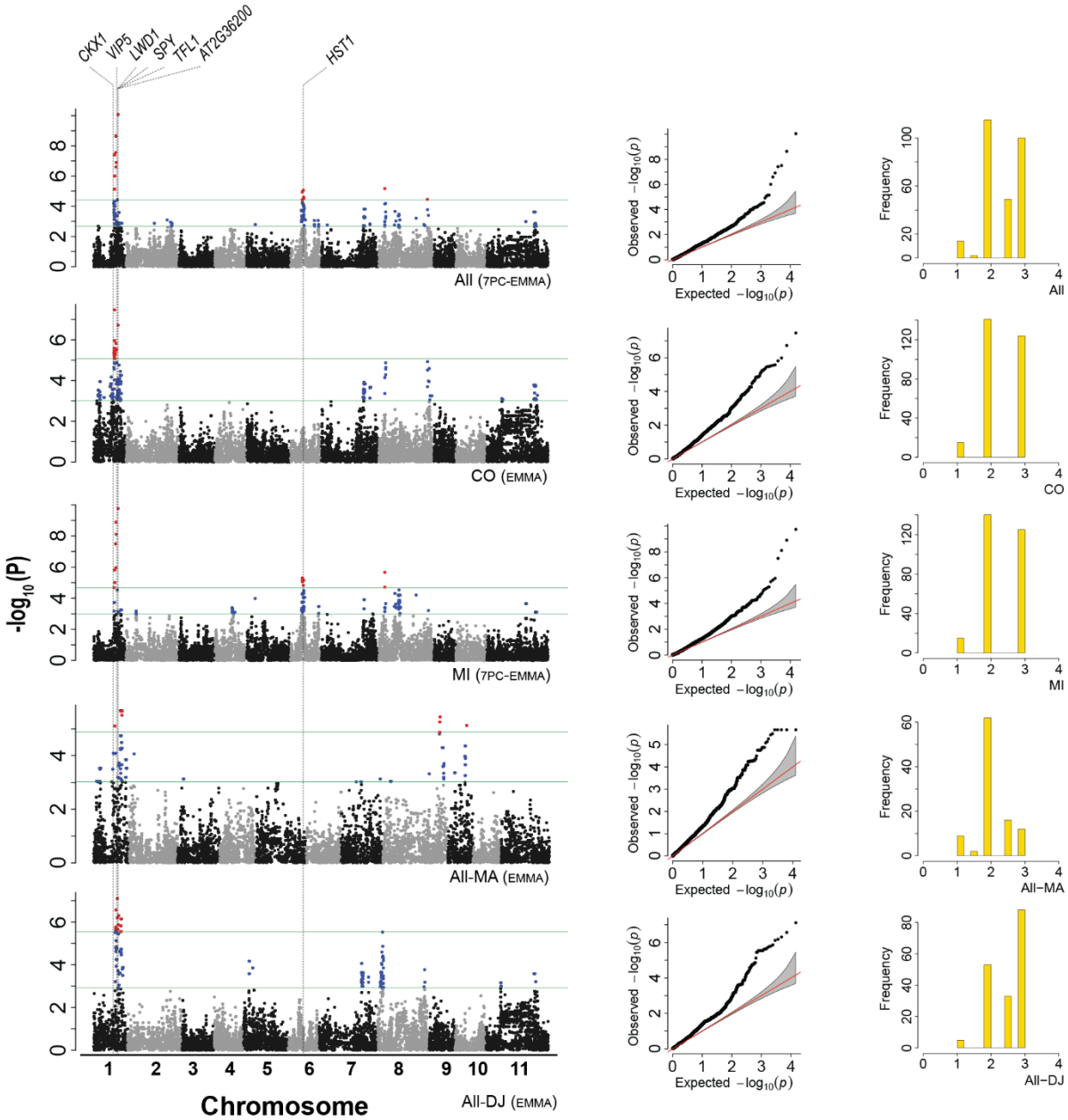
(I)



Growth Habit

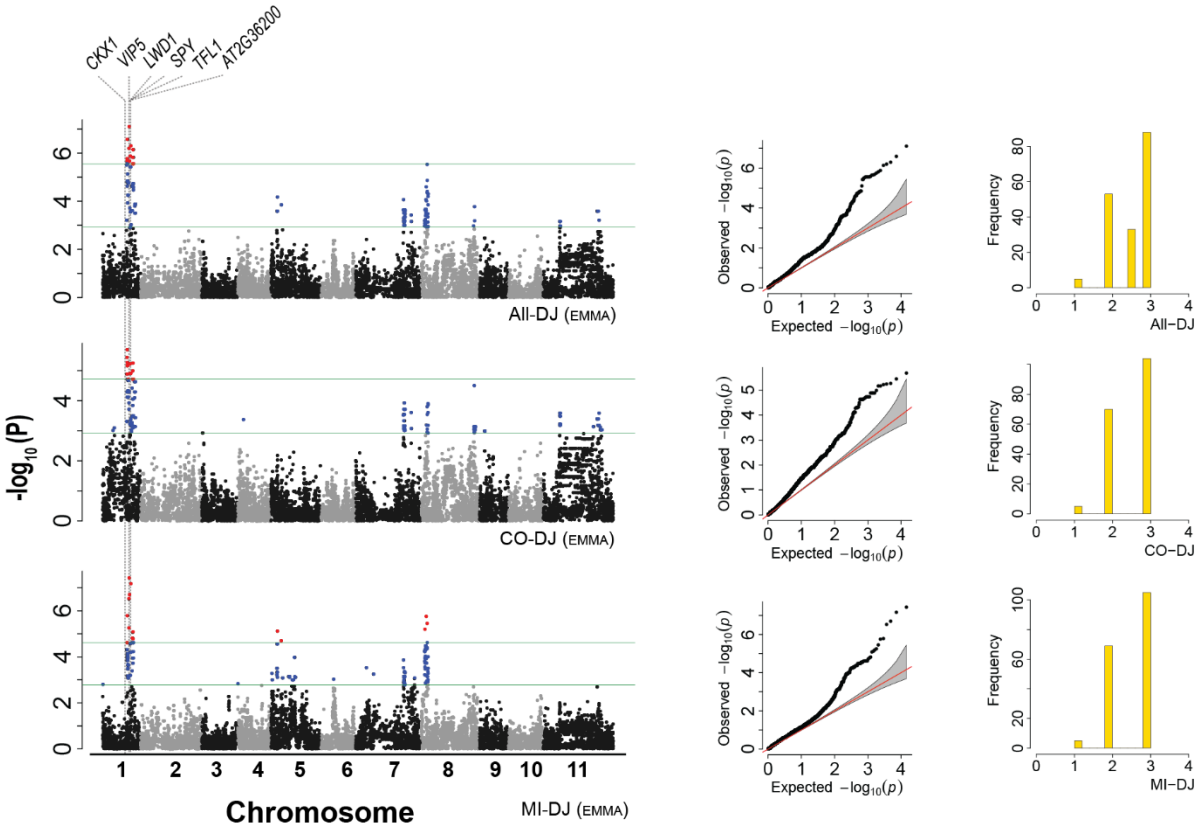
With Determinates

(J)



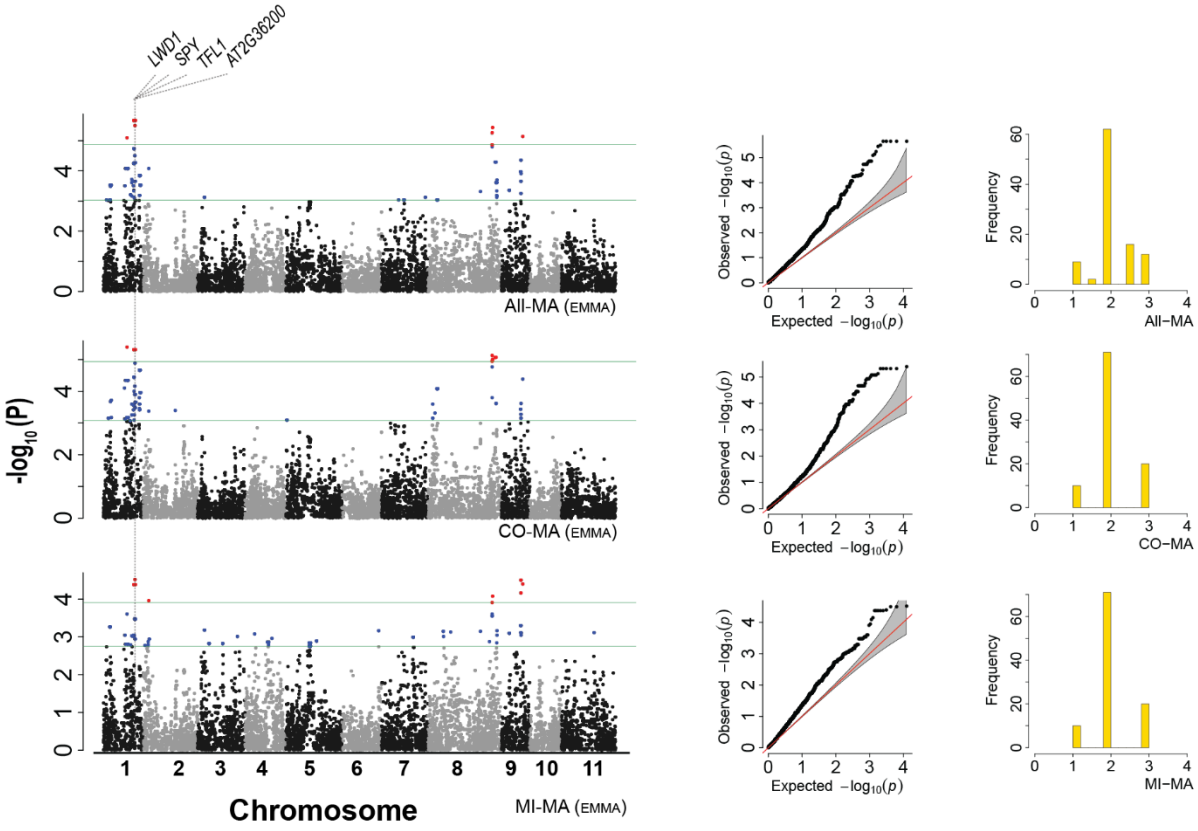
Growth Habit
with Determinate
Durango/Jalosco

(K)



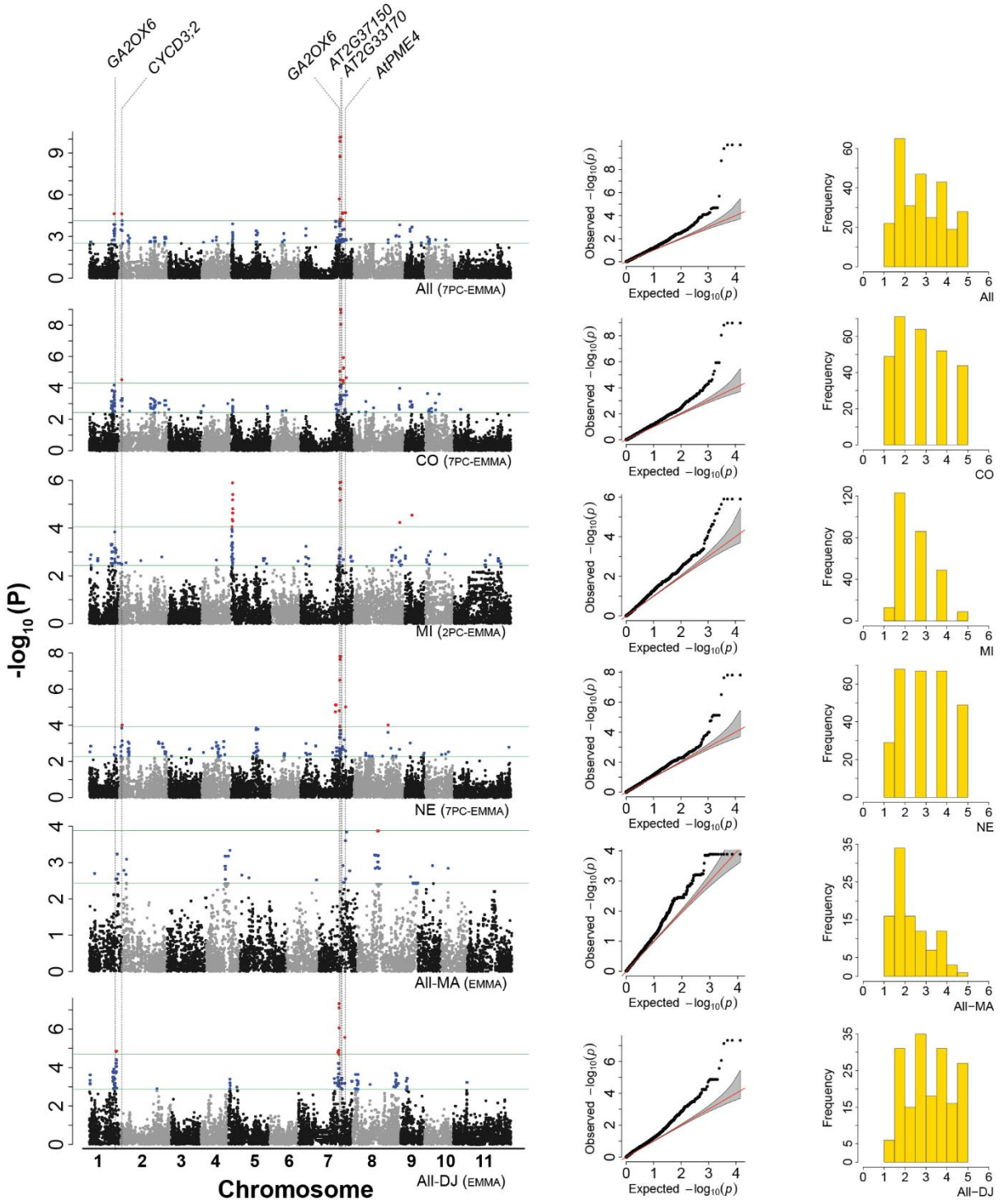
Growth Habit
With Determinates
Mesoamerican

(L)



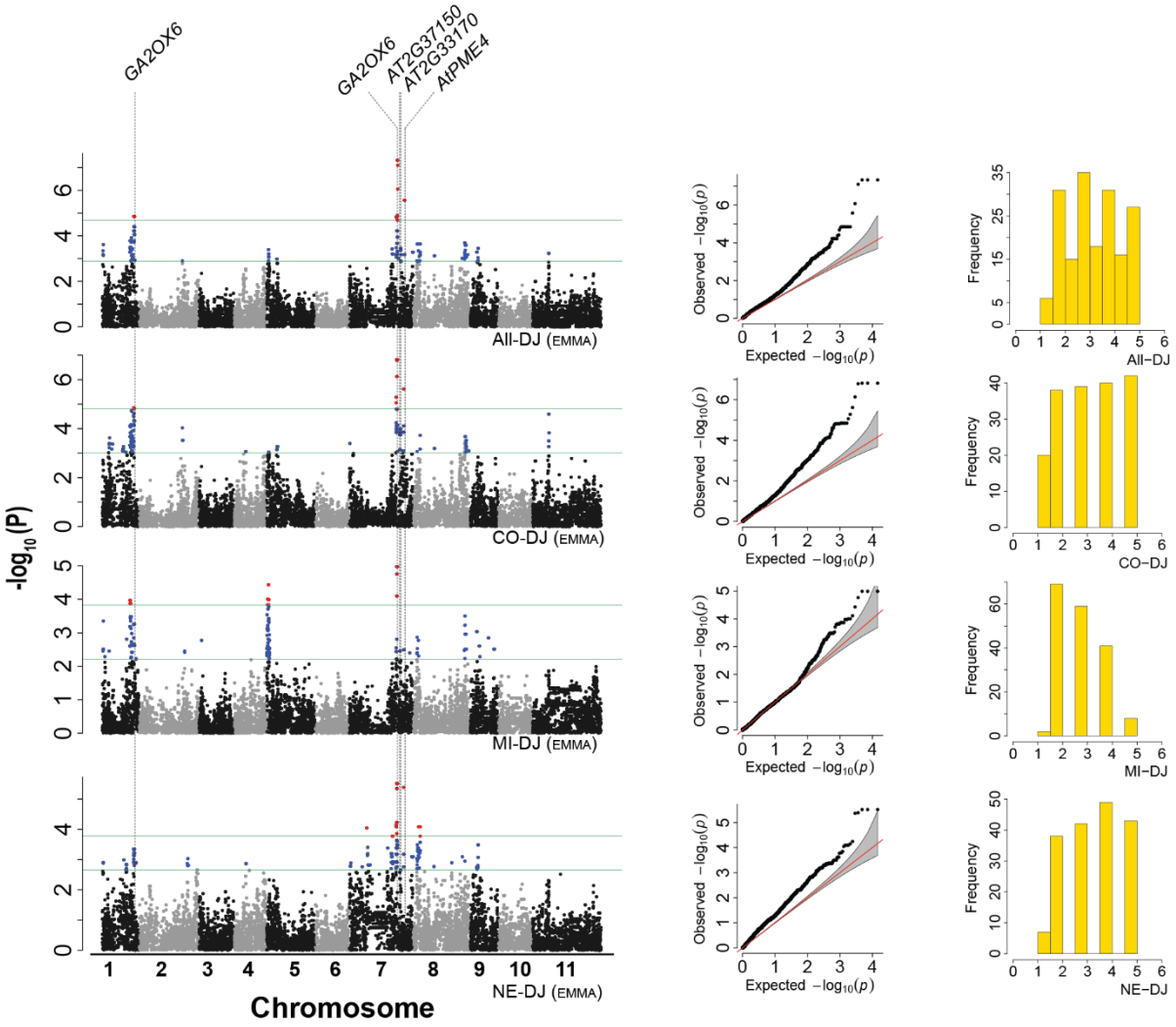
Lodging

(M)


GA2OX6
CYCD3;2
GA2OX6
AT2G37150
AT2G33170
AtPME4
All (7PC-EMMA)
CO (7PC-EMMA)
MI (2PC-EMMA)
NE (7PC-EMMA)
All-MA (EMMA)
All-DJ (EMMA)
-log10(P)
Chromosome
Observed −log10(p)
Expected −log10(p)
Frequency
All
CO
MI
NE
All−MA
All−DJ

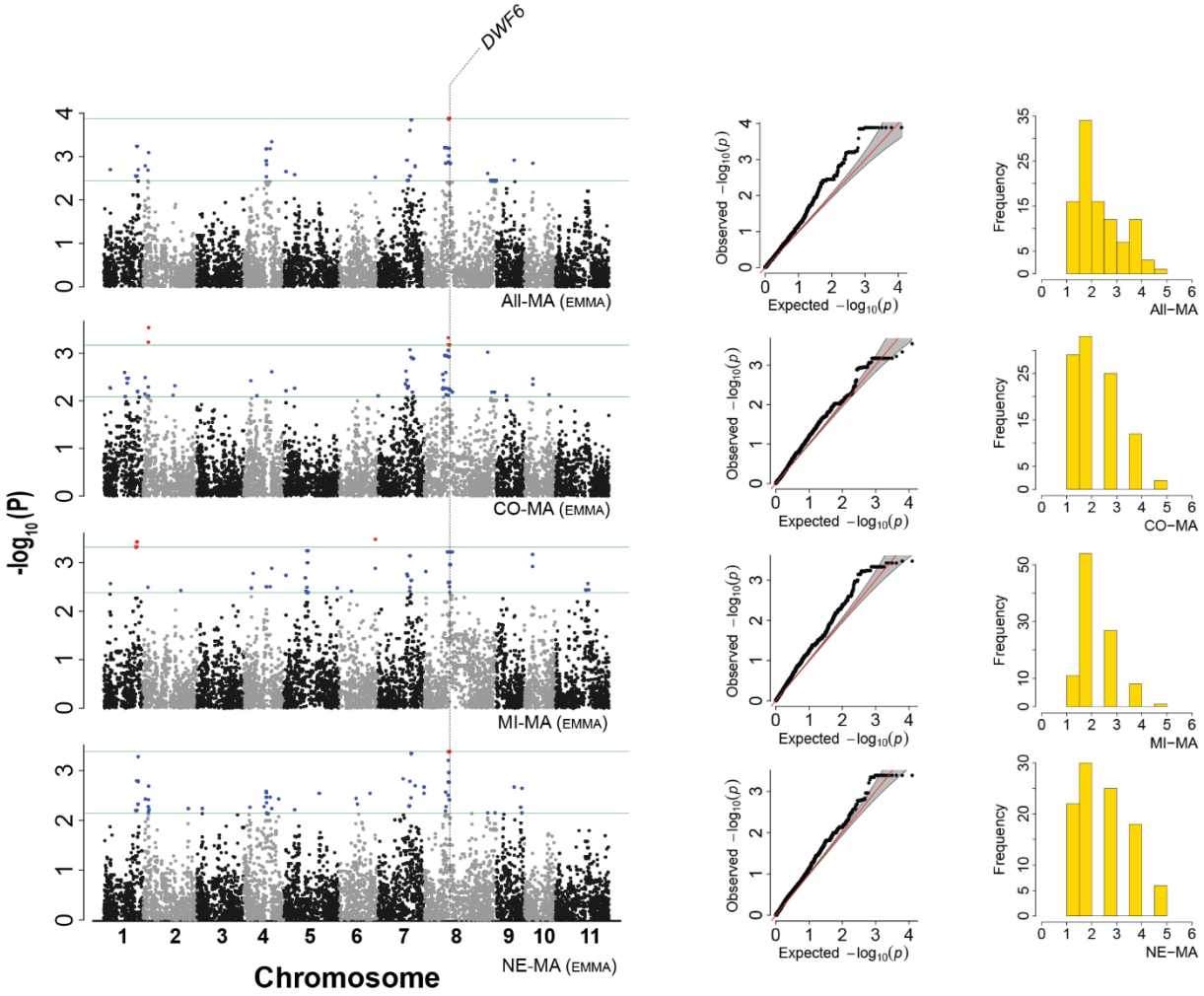
Lodging
Durango/Jalisco

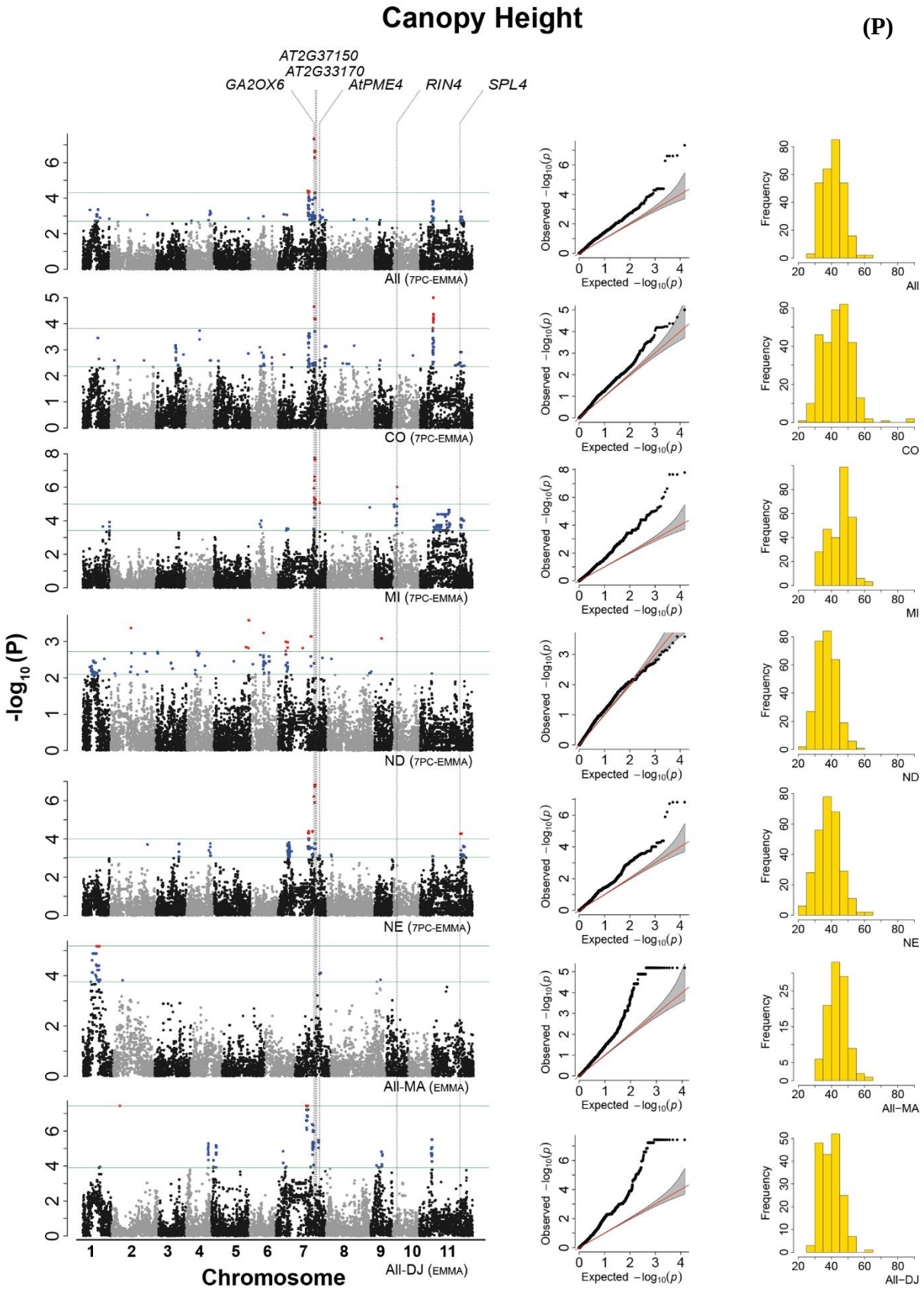
(N)



Lodging
Mesoamerican

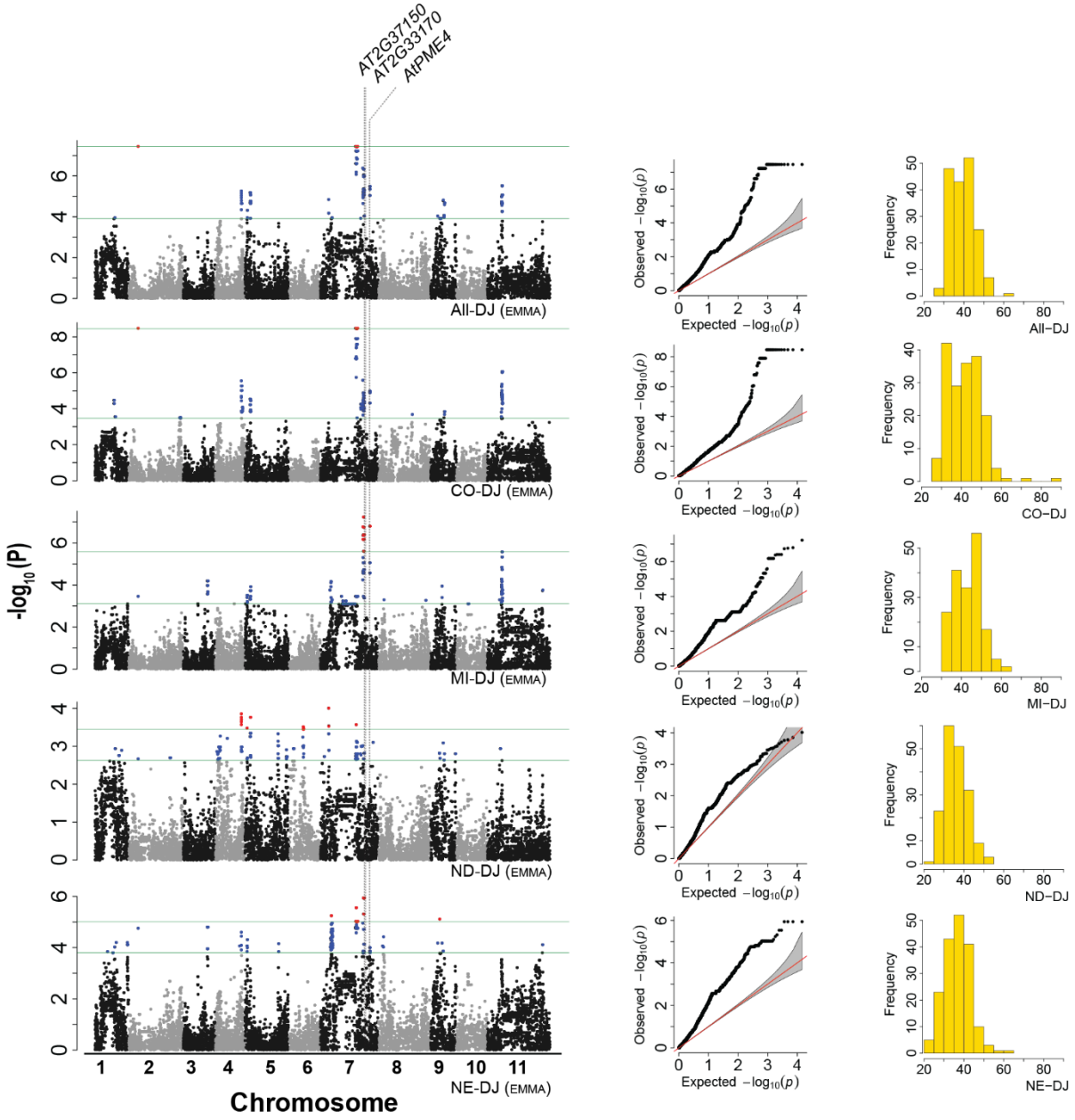
(O)




Canopy Height
(P)
AT2G37150
AT2G33170
GA2OX6
AtPME4
RIN4
SPL4
All (7PC-EMMA)
CO (7PC-EMMA)
MI (7PC-EMMA)
ND (7PC-EMMA)
NE (7PC-EMMA)
All-MA (EMMA)
All-DJ (EMMA)
-log10(P)
Chromosome
Observed −log10(p)
Expected −log10(p)
Frequency
All
CO
MI
ND
NE
All−MA
All−DJ

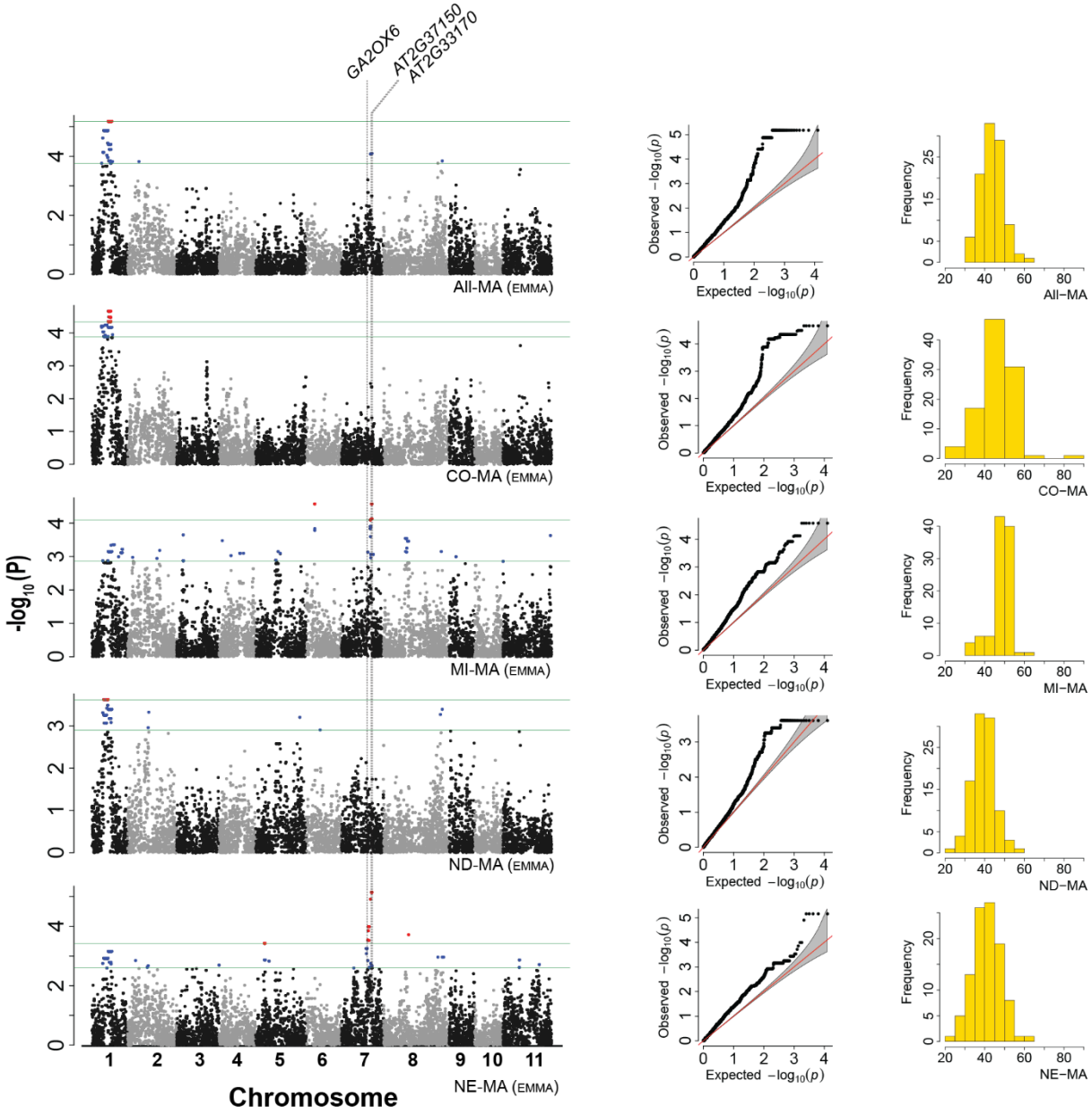
Canopy Height
Durango/Jalisco

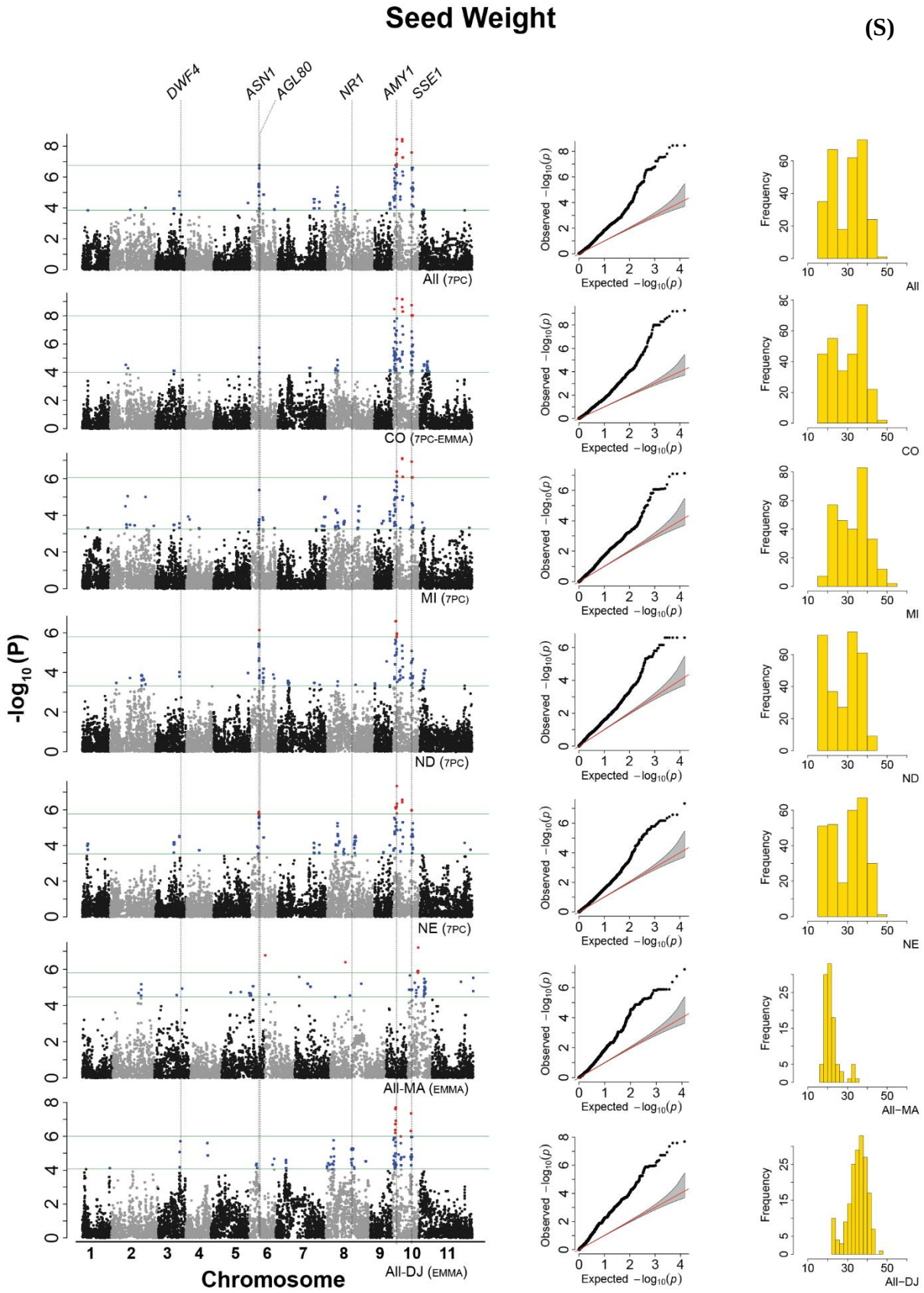
(Q)



Canopy Height
Mesoamerican

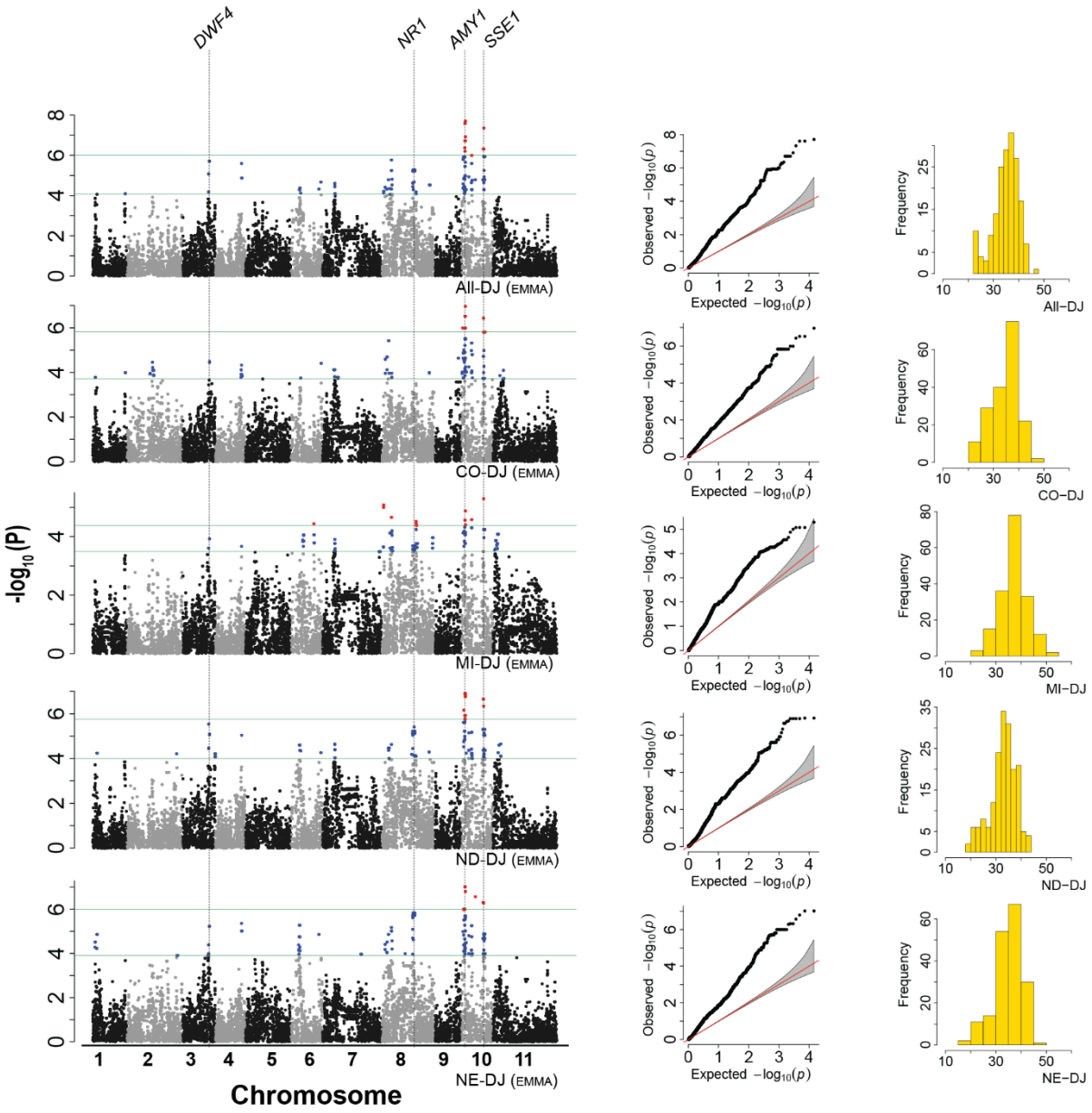
(R)




Seed Weight
(S)
DWF4
ASN1
AGL80
NR1
AMY1
SSE1
All (7PC)
CO (7PC-EMMA)
MI (7PC)
ND (7PC)
NE (7PC)
All-MA (EMMA)
All-DJ (EMMA)
-log10(P)
Chromosome
Observed −log10(p)
Expected −log10(p)
Frequency
All
CO
MI
ND
NE
All−MA
All−DJ

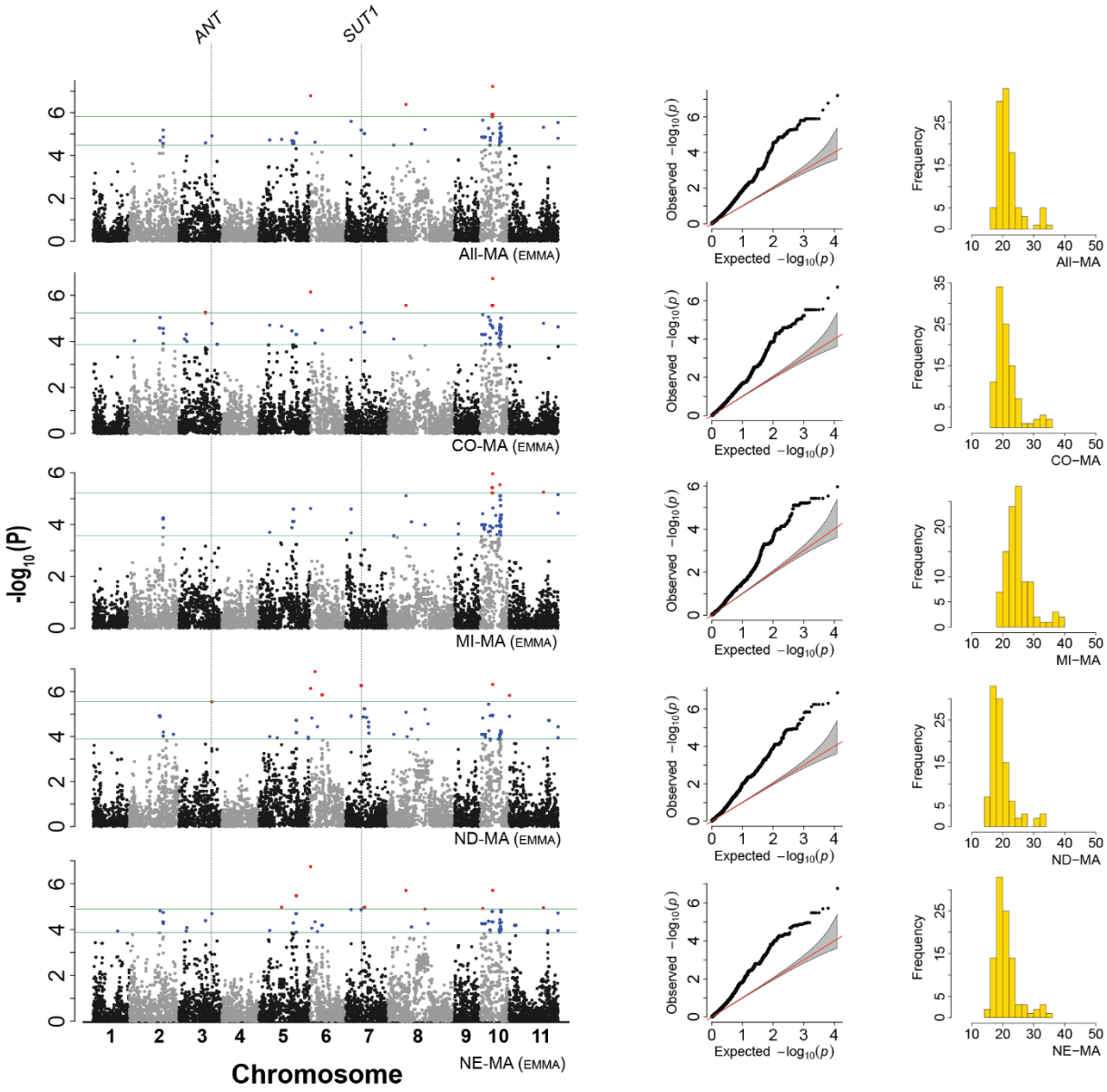
Seed weight
Durango/Jalisco

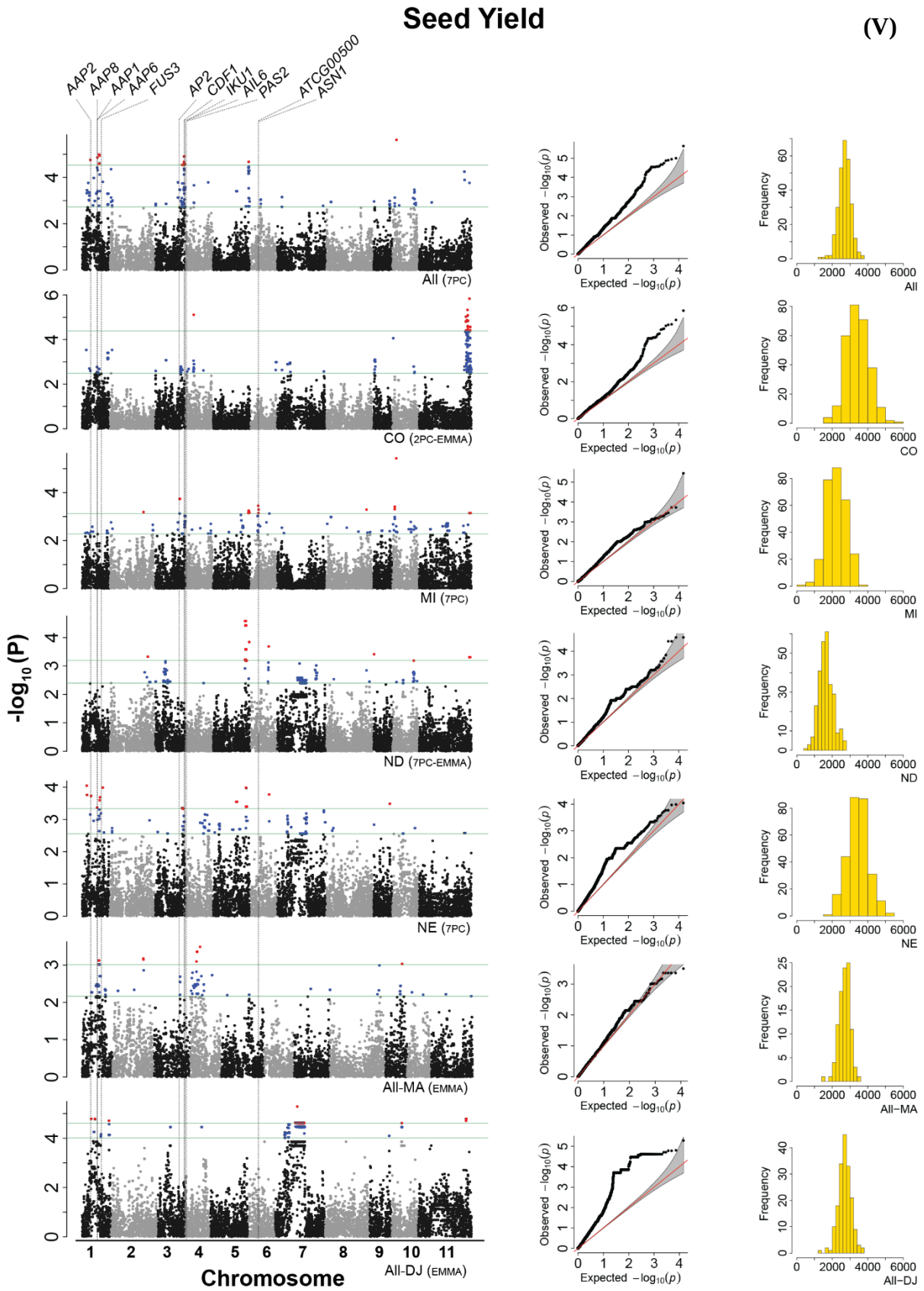
(T)



Seed Weight
Mesoamerican

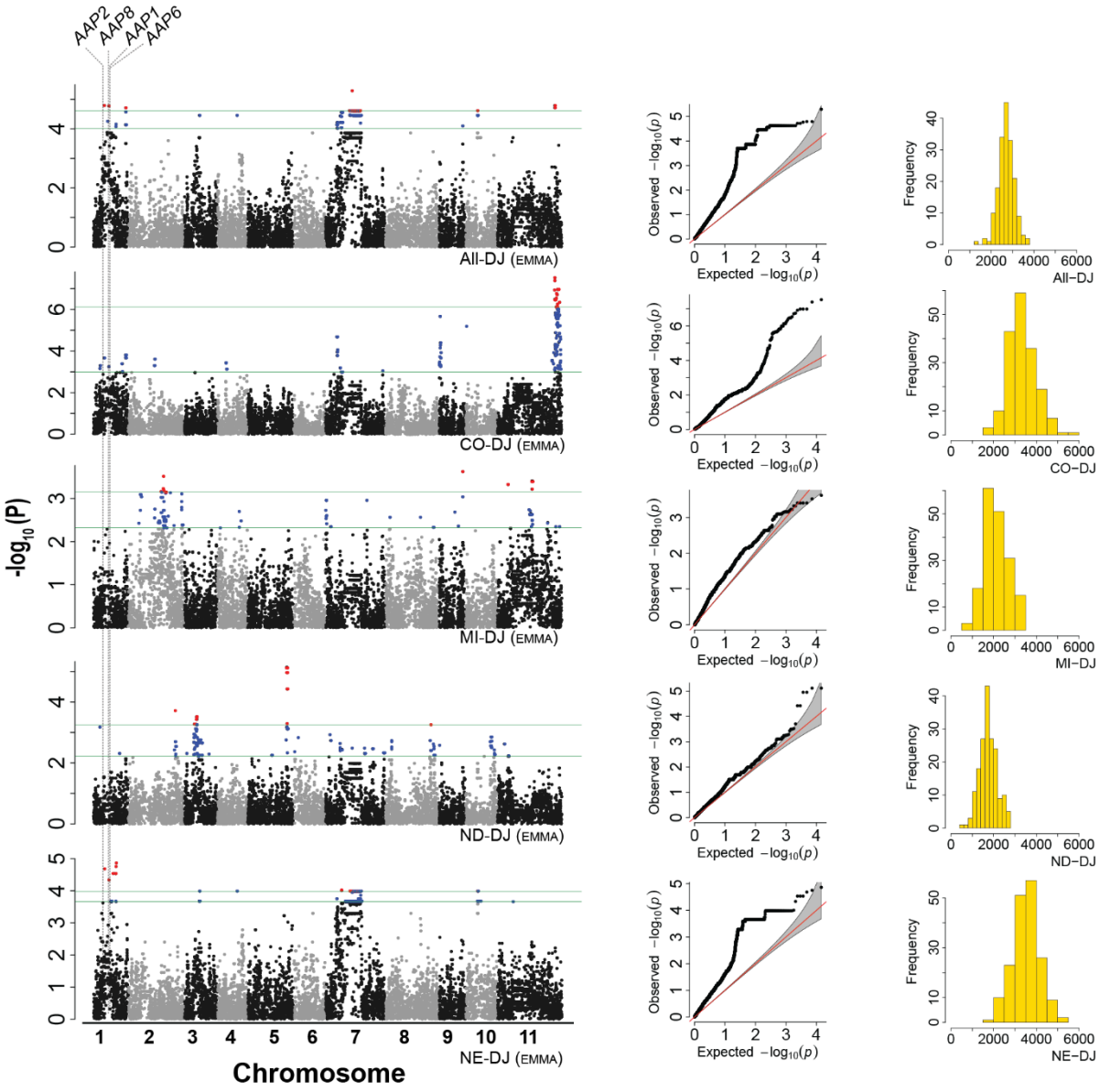
(U)




Seed Yield
(V)
AAP2
AAP8
AAP1
AAP6
FUS3
AP2
CDF1
IKU1
AIL6
PAS2
ATCG00500
ASN1
All (7PC)
CO (2PC-EMMA)
MI (7PC)
ND (7PC-EMMA)
NE (7PC)
All-MA (EMMA)
All-DJ (EMMA)
-log10(P)
Chromosome
Observed −log10(p)
Expected −log10(p)
Frequency
All
CO
MI
ND
NE
All−MA
All−DJ

Seed Yield
Durango/Jalisco

(W)



Seed Yield
Mesoamerican

(X)