Genome-Wide Association Study Identifies Candidate Loci Underlying Agronomic Traits in a Middle American Diversity Panel of Common Bean

Samira Mafi Moghaddam,* Sujan Mamidi, Juan M. Osorno, Rian Lee, Mark Brick, James Kelly, Phillip Miklas, Carlos Urrea, Qijian Song, Perry Cregan, Jane Grimwood, Jeremy Schmutz, and Phillip E. McClean

ABSTRACT

Common bean (Phaseolus vulgaris L.) breeding programs aim to improve both agronomic and seed characteristics traits. However, the genetic architecture of the many traits that affect common bean production are not completely understood. Genome-wide association studies (GWAS) provide an experimental approach to identify genomic regions where important candidate genes are located. A panel of 280 modern bean genotypes from race Mesoamerica, referred to as the Middle American Diversity Panel (MDP), were grown in four US locations, and a GWAS using >150,000 single-nucleotide polymorphisms (SNPs) (minor allele frequency [MAF] ≥ 5%) was conducted for six agronomic traits. The degree of inter- and intrachromosomal linkage disequilibrium (LD) was estimated after accounting for population structure and relatedness. The LD varied between chromosomes for the entire MDP and among race Mesoamerica and Durango–Jalisco genotypes within the panel. The LD patterns reflected the breeding history of common bean. Genome-wide association studies led to the discovery of new and known genomic regions affecting the agronomic traits at the entire population, race, and location levels. We observed strong colocalized signals in a narrow genomic interval for three interrelated traits: growth habit, lodging, and canopy height. Overall, this study detected ~30 candidate genes based on a priori and candidate gene search strategies centered on the 100-kb region surrounding a significant SNP. These results provide a framework from which further research can begin to understand the actual genes controlling important agronomic production traits in common bean.

Published in Plant Genome

© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA
This is an open access article distributed under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Common bean is the most important grain legume for direct human consumption (Broughton et al., 2003). It has a relatively small diploid genome (2n = 22; 521.1 Mb; Schmutz et al., 2014) and consists of two 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 genetic diversity (Mamidi et al., 2013; Schmutz et al., 2014). Selection under domestication within each gene pool has generated distinct ecogeographical races and market classes (Singh et al., 1991; Beebe et al., 2000; Diaz and Blair, 2006; Mamidi et al., 2011b). Members of each race have specific morphological, agronomic, physiological, and molecular traits in common and can differ in the allele frequencies of the genes responsible for a trait (Singh et al., 1991). The

S.M. Moghaddam, S. Mamidi, P.E. McClean, Genomics and Bioinformatics Program and Dep. of Plant Sciences, North Dakota State Univ., Fargo, ND 58105-6050; Present address S. Mamidi, HudsonAlpha Institute of Biotechnology, Huntsville, AL 35806; J.M. Osorno and R. Lee, Dep. of Plant Sciences, North Dakota State Univ., Fargo, ND 58105-6050; M. Brick, Dep. of Soil and Crop Sciences, Colorado State Univ., Fort Collins, CO 80523-1170; J. Kelly, Dep. of Plant, Soil and Microbial Sciences, Michigan State Univ., East Lansing, MI 48824; P. Miklas, Vegetable and Forage Crops Research Unit, USDA–ARS, Prosser, WA 99350; C. Urrea, Dep. of Agronomy and Horticulture, Univ. of Nebraska, NE 69361-4939; Q. Song and P. Cregan, Soybean Genomics and Improvement Lab, USDA–ARS, Beltsville, MD 20705; J. Grimwood and J. Schmutz, HudsonAlpha Institute of Biotechnology, Huntsville, AL 35806; Received 8 Feb. 2016. Accepted 8 July 2016. *Corresponding author (samira.mm@gmail.com; samira.mafi@ndsu.edu).

Abbreviations: DJ, Durango–Jalisco; GWAS, genome-wide association studies; LD, Linkage disequilibrium; MA, Mesoamerican; MAF, minor allele frequency; MDP, Middle American diversity panel; MLMM, multi-locus mixed model; PCA, principle component analysis; QTL, quantitative trait loci; SNP, single nucleotide polymorphism.
Middle American gene pool consists of races Durango–Jalisco (DJ) and Mesoamerica (MA). Pinto, great northern, medium red, and pink bean are major US market classes within race DJ, while navy and black bean are major race Mesoamerican market classes in the United States. (Mensack et al., 2010; Vandemark et al., 2014). Other studies investigated the population structure in the common bean gene pools and confirmed these classifications (Blair et al., 2009; Kwak and Gepts, 2009).

Cultivated forms of common bean show considerable morphological variation (Hedrick et al., 1931) for growth habit, seed size, and color (Leakey, 1988; Singh, 1989). Domestication and breeding selected for agronomic traits important for production and altered the genetic architecture underlying those traits. Breeding for higher yield in bean is affected by many interdependent traits including growth habit, seed size, and maturity (Kelly et al., 1998; Kornegay et al., 1992). Genetic factors controlling important agronomic traits have been mapped to common bean linkage groups using biparental populations and significant quantitative trait loci (QTL) mapped to wide intervals (Beaute et al., 2003; Blair et al., 2006, 2012; Checa and Blair, 2012; Pérez-Vega et al., 2010; Tar’an et al., 2002; Wright and Kelly, 2011). Nevertheless, our knowledge of genes controlling agronomic traits is limited. Quantitative trait loci analysis usually has low resolution because of the limited number of recombination events in a biparental population (Balasubramanian et al., 2009). Quantitative trait loci intervals can span a few centiMorgans (cM), which can indeed be megabases (Mb) long in physical distance and contain hundreds of candidate genes.

In contrast, GWAS considers more recombination events by using an association panel of individuals each with unique recombination histories. This provides higher mapping resolution, as a result of shorter LD blocks, than a biparental 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; Song et al., 2015). Resequencing GWAS populations and mapping the reads on to the reference genome sequence of common bean (Schmutz et al., 2014) lead to higher mapping resolution because of the greater depth of coverage. Multiple GWAS studies have led to the 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). Therefore, GWAS has recently been implemented in common bean (Cichy et al., 2015; Kamfwa et al., 2015a,b).

In this study, a GWAS was performed followed by candidate gene identification for six important agronomic traits that affect common bean production: days to flower, days to maturity, growth habit, canopy height, lodging, and seed weight. Over 240,000 SNPs were obtained from a collection of 280 diverse genotypes from a population of Middle American genotypes.

Materials and Methods

Middle American Diversity Panel and Single-Nucleotide Polymorphism Data Set

The MDP consists of 280 modern Middle American dry bean cultivars from among the 507 genotypes of the entire BeanCAP diversity panel (http://www.beancap.org/_pdf/DRY_BEAN_GENOTYPES_BEANCAP.pdf). This subpopulation was chosen to reduce the effect of population structure during the GWAS. The MDP itself consists of 100 race MA and 180 race DJ genotypes.

The SNP data set was obtained by genotyping the MDP with two Illumina iSelect 6K Gene Chip (BARCBEAN6K_1 and BARCBEAN6K_2) sets (Song et al., 2015) and low-pass sequencing. The Gene Chips are described in Song et al. (2015). A total of 8526 SNPs from the two 6K Gene Chips were mapped to the common bean reference genome (Schmutz et al., 2014). Single-enzyme low-pass sequencing was performed by Institute for Genomic Diversity at Cornell University according to Elshire et al. (2011) using the enzyme ApeKI for digestion and yielded 17,611 mapped SNPs. The two-enzyme, low-pass sequencing SNP set was created according to Schröder et al. (2016) using TaqI and MseI enzymes. The depth of high-quality mapped reads ranged from 0.1× to 3.5× among the genotypes with an average of 0.7×. A total of 217,486 mapped SNPs were identified. The SNPs with <50% missing data were imputed. All SNP data sets were merged and imputed in fastPHASE (Scheet and Stephens, 2006). The chromosome distribution of SNPs that resulted from the three different platforms is summarized in Table 1. The unimputed HapMap data for each of the three platforms and the combined imputed data in numeric format can be found at http://www.beancap.org/Research.cfm.

Phenotypic Analysis

Data for days to flower, days to maturity, growth habit, canopy height, lodging, and seed weight were collected from field trials grown in Colorado, Michigan, Nebraska, and North Dakota. Days to flower were measured as the number of days from planting to when ~50% of the plants in a plot had at least one open flower. Days to maturity were measured as the number of days from planting until complete physiological development and senescence. Growth habit was recorded during flowering and verified at senescence as type I (determinate erect or upright), type II (indeterminate erect), or type III (indeterminate prostrate). Canopy height was measured at harvest and was recorded in centimeters from the base of the plant at the soil surface to the top of the canopy. Lodging was scored at harvest on a 1-to-5 scale, where 1 indicates 100% plants standing erect and 5 indicates 100% plants flat on the ground. Seed weight was measured as the weight of 100 randomly selected undamaged seeds recorded in grams at 16% moisture. Data for growth habit in Nebraska and growth habit and lodging in North Dakota were not available. The experimental design was an α-lattice design with two replicates. 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). Trait correlations were calculated and plotted in R 3.0 (R Development Core Team, 2013) using cor.matrix() and corrplot() from the corrplot package (Wei and Wei, 2013). Histograms were created in R 3.0 (R Development Core Team, 2013) using the hist() function. Adjusted phenotypic values were calculated using the lsmeans statement in SAS 9.3 (SAS Institute, 2011) PROC MIXED per location and across all locations. A total of 27 phenotypic datasets were analyzed. The lsmeans are available at http://www.beancap.org/_data/Adjusted-means-for-Agronomic-traits-with-race-and-market-calss-info.txt.

### Population Structure and Kinship

Population structure was calculated with STRUCTURE 2.3 (Pritchard et al., 2000) using SNPs whose pairwise $r^2$
was <0.5 for all pairwise comparisons. STRUCTURE uses a model-based method to assign the subpopulation membership to each individual. The STRUCTURE parameters used were 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 subpopulations with five replications. The Wilcoxon signed rank test implemented in SAS 9.3 (SAS Institute, 2011) was used to select the optimum number of subpopulations (Rosenberg et al., 2001). The optimum number of subpopulations was the smallest $K$ in the first nonsignificant Wilcoxon test. Distruct 1.1 (Rosenberg, 2003) was used for graphical display of the STRUCTURE output. Population structure for GWAS was determined using principal component analysis (PCA) using the prcomp() R 3.0 function (R Development Core Team, 2013; Price et al., 2006) with the complete SNP data set for loci with MAF ≥ 5%. To account for individual relatedness, an identity-by-state kinship matrix was generated by the EMMA algorithm (Kang et al., 2008) embedded in GAPIT (Lipka et al., 2012) using the complete SNP data set with MAF ≥ 0.05.

**Linkage Disequilibrium**

Pairwise LD between markers in the null model was calculated as the squared allele frequency correlation in PLINK (Purcell et al., 2007) using SNPs with MAF ≥ 5%. To calculate genome-wide LD, a set of SNPs spaced every 50 Kb across the genome was used. The R package LDcorSV (Mangin et al., 2012) was used to calculate the pairwise LD when accounting for population structure, relatedness, or both. The LD heatmaps were generated for each chromosome in R 3.0 (R Development Core Team, 2013) using the LDheatmap package (Shin et al., 2006). The positions of the pericentromeric borders were obtained from Supplementary Table 6 in Schmutz et al. (2013) using the LDheatmap package (Shin et al., 2006). The positions of the pericentromeric borders were obtained from Supplementary Table 6 in Schmutz et al. (2013) using the LDheatmap package (Shin et al., 2006).

**Genome-Wide Association Study**

From a total of 243,623 SNPs, 159,884 with a MAF ≥ 5% were used for the MDP GWAS. The DJ subpopulation ($n = 180$) GWAS analyses used 151,239 SNPs (MAF ≥ 5%), while the MA subpopulation GWAS analyses used 134,924 SNPs (MAF ≥ 5%). Analyses were performed using pooled data across all locations and for each location separately. More than 200 GWAS analyses were performed using GAPIT, an R function developed by Lipka et al. (2012). Multiple models were tested per trait in MDP: (i) null general linear model, (ii) general linear model with fixed effects to control for population structure, (iii) univariate unified mixed linear model (Yu et al., 2005) using the population parameters previously determined (Zhang et al., 2010) protocol to control for individual relatedness (random effect), and (iv) a model controlling for both individual relatedness and population structure. We used the first two principal components (controls for ~25% of variation) to account for population structure. The best model was selected based on the mean squared difference as described by Mamidi et al. (2011a). The final Manhattan plots were created using the mplot() function from R package gap (Zhao, 2007). The phenotypic variation explained by significant markers in the best model was calculated based on the likelihood-ratio-based $R^2$ ($R^2_{L R}$) (Sun et al., 2010) using the genABLE package in R (Aulchenko et al., 2007).

**Multilocus Mixed Model**

Multilocus mixed model (MLMM) (Segura et al., 2012) was used to evaluate the results for single-marker tests. The 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. The best model was selected considering both multiple Bonferroni and extended Bayesian information criteria.

**Significant Single-Nucleotide Polymorphisms and Candidate Gene Identification**

The significance cutoff threshold affects candidate gene identification in GWAS. Although controlling for false positives is very crucial, the effect of false negatives should not be ignored. False negatives can certainly occur if the cutoff value is too stringent. Different approaches to setting significance cutoffs have included selecting the top 50 SNPs (Stanton-Geddes et al., 2013), considering two false discovery rate significance levels (Lipka et al., 2013), or adhering to the stringent Bonferroni cutoff. We defined significant markers by two cutoff levels. For the first, SNPs falling in the lower 0.01 percentile tail of the empirical distribution of $P$-values after 1000 bootstraps (Mamidi et al., 2014), which is a stringent cutoff but does not imply that if a SNP does not pass this criterion it does not affect the trait of interest. This was demonstrated by Atwell et al. (2010), where a SNP close to or even within a candidate gene previously validated by transgenic complementation experiments did not meet specific significance criteria. Therefore, the second cutoff used was more relaxed. SNPs falling in the lower 0.1 percentile tail of the empirical distribution of $P$-values after 1000 bootstraps were considered. Two approaches were used to define candidate genes: (i) a priori candidates based on the published literature and (ii) evaluation of genes within a 100-Kb window centered on a significant SNP. The primary focus was on SNPs falling in the more stringent cutoff, but SNPs falling in the less stringent cutoff were considered if there was a priori candidate support.

**Results**

**Population Structure**

With $K = 2$ subpopulations, the MDP was split into two subpopulations corresponding to races MA and Durango (Fig. 1a). When the subpopulation number was set to $K = 3$ to $K = 7$, genotypes clustered according to the market classes and admixture was observed between the market classes. Figure 1b shows that the first three principal components from PCA correspond to the STRUCTURE results. The first component separates race MA and
Durango. Pink and small red bean genotypes are distributed across both races, a result consistent with the admixture pattern in these market classes observed with the STRUCTURE analysis (Fig. 1a). The PC2 clusters some of the pinto and great northern bean genotypes closer to pink and small red bean market classes; the PC3 separates the race MA navy and black bean market classes and shows that small red and some pink bean are more closely related to the black bean market class.

**Linkage Disequilibrium**

The LD heatmaps were created for each chromosome for the MDP and the MA and DJ race subpopulations to illustrate the extent of LD at different levels of population structure. The LD pattern not only varied between subpopulations but also varied among chromosomes within a subpopulation. Pv06, Pv07, and Pv11 showed dramatically different LD patterns among the DJ and MA race subpopulations (Fig. 2a, 2b).
The majority of interchromosomal LD ($r^2 \geq 0.6$ after controlling for population structure and relatedness in MDP) occurred among pericentromeric regions. The largest interchromosomal LD is between Pv07 and other chromosomes. A ~30-Mb region of Pv07 that encompasses the complete pericentromeric region (27 Mb) and a small amount of neighboring euchromatic DNA is in LD with either a single SNP or narrow regions on other chromosomes. Only Pv05 and Pv09 do not exhibit any LD with Pv07. Although the mixed model significantly reduced the overall interchromosomal LD, some long-range LD blocks persisted in the DJ subpopulation, but...
very little interchromosomal LD was detected in the MA subpopulation (Fig. 3a, 3b).

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 d in the MA and DJ subpopulations, respectively. Days to maturity ranged from 60–125 d. The plants matured earlier in Michigan (96–110 d) than in North Dakota (75–125 d). Seed weight showed a bimodal distribution with values ranging from 14.1 to 39.7 g 100 seeds⁻¹ for race MA genotypes and 21.1 to 53.7 g 100 seeds⁻¹ for race DJ genotypes. Figure 4 shows the distribution of each trait across
Fig. 3. Genome-wide linkage disequilibrium heatmap 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 individual relatedness. A subset of single-nucleotide polymorphisms spaced every 50 kb was used, and only pairwise $r^2 > 0.6$ are shown. The gray rectangular areas show the pericentromeric regions. Gray dashed lines define the chromosome boundaries.
Fig. 4. Density plot of phenotypic distribution of six agronomic traits. 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.
different locations and combined across locations. Trait values were highly correlated among locations except for 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, while canopy height was negatively correlated with growth habit and lodging (Fig. 5). Table 2 shows that heritability was moderate to high for each trait in races DJ and MA.

### Genome-Wide Association Study

Multiple genomic regions were found to be associated with the six agronomic traits. The growth habit data was analyzed with and without type I determinate genotypes because the strong Pv01 signal associated with the determinacy locus reduced the number of loci at other positions associated with the type II and III phenotypes that passed the cutoff value. The $P$-values for the best model varied among traits, locations, and subpopulations. This suggested that the number and effect size varied among the loci that affected the different traits. Therefore, a single $P$-value cutoff would not be a suitable approach to identify relevant SNPs among all traits. Rather, the $P$-values were bootstrapped 1000 times for each trait, and SNPs falling in the top 0.01 percentile of the empirical distribution were considered significant (Mamidi et al., 2014). This led to a different cutoff threshold for each trait and sometimes the cutoff was even more stringent than the Bonferroni value. Supplemental Table S1 shows the significant SNPs and the candidate gene models within 50 Kb upstream and downstream of the significant marker. For each trait, the amount of variation explained by a SNP was only calculated for those SNPs for which a candidate gene was identified.
For days to flower, collectively, all candidate SNPs explained 14% of the phenotypic variation for the entire MDP and across all the locations analyzed. The strongest signal for days to flower was located on Pv01 and includes two major peaks in a block that spans from 8 to 20 Mb. The most significant GWAS peak on Pv01 differed among the races. The most significant GWAS peak for race DJ occurred at Pv01 (8 Mb), while for race MA, the peak SNP was located at Pv01 (15 Mb) (Fig. 6). 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 is in LD ($r^2 \approx 0.4$) with the 15-Mb region after controlling for both population structure and relatedness. To further investigate this region, we used the MLMM (Segura et al., 2012) for the MDP population and the DJ and MA subpopulations. The most significant SNP controlled all the variation on Pv01 when used as a cofactor in the MLMM analysis. The interaction effect between the candidate SNPs at these two regions was not significant.

For days to maturity, GWAS peaks were noted across the genome and with different peaks at each location and for each subpopulation. Collectively, all candidate SNPs explained 21% of the variation. The most noticeable peaks are on Pv04 and Pv11. The peak at the end of Pv11 is from North Dakota, and the peak at the beginning of Pv11 is from Nebraska and Michigan. The peak on Pv04 originates from Michigan and has mainly a DJ origin. There is a Nebraska-specific peak on Pv01 and a Colorado-specific peak on Pv07 (Supplemental Fig. S1d–f). Most of the candidate genes found for this trait are associated with flowering time, senescence, or nutrient remobilization.

For the entire MDP population, a major GWAS signal was noted on Pv01 for growth habit where the flowering time candidate genes $LWD1$, SPY, and $TFL-1$ are located. All candidate SNPs on Pv01 collectively explained 42% of the phenotypic variation. When the determinate type I genotypes were excluded from the growth habit analysis, the Pv01 peak disappeared and major peaks appeared at Pv04 (3 Mb), Pv06 (30 Mb, Pv07 (46 Mb), and Pv11 (10 and 43 Mb) (Fig. 7). The candidate SNPs across the genome and those on Pv07...
explained 33 and 12% of the phenotypic variation, respectively.

For lodging, a very strong and consistent Pv07 (46 Mb) peak was observed at all locations as well as across all locations. The peak is 68.6 Kb wide and consists of multiple SNPs, with pairwise \( r^2 \) LD values >0.5. There are two smaller signals at Pv07 (45 Mb) and Pv07 (48 Mb) that are not in LD with the Pv07 (46 Mb) region (Fig. 8). There is another peak at 35.42 Mb, which falls in the 0.1 percentile tail of the empirical distribution of the \( P \)-values. All of the Pv07 variation is accounted for when SNPs at both Pv07 (46.11 Mb) and Pv07 (35.42 Mb) are included as cofactors in a MLMM analysis. Using SNPs at 48 or 45 Mb as cofactors did not control for all the Pv07 variation. The candidate SNPs on Pv07 collectively explained 21% of the phenotypic variation. A significant SNP at the proximal end of Pv05 was observed in Michigan, but no candidate genes were identified (Supplemental Fig. S1m–o).

The same lodging Pv07 (46 Mb) region was significant for canopy height across all locations (Fig. 7) and in each location except for North Dakota (Supplemental Fig. S1p–r). We also found a major peak on Pv04 for this trait that was not present for the lodging analysis. However, this peak did not remain significant when the SNP at Pv07 (46 Mb) was used as a cofactor in MLMM.

Major peaks for seed weight were observed on Pv10 across all locations and individually in Colorado and Michigan (Supplemental Fig. S1s–u). This signal is mainly
from the DJ subpopulation, while no significant peak in the MA subpopulation was observed. The MLMM analysis did not select any signals for seed weight. This might imply the presence of other environmental or confounding variables that affect the causal loci nonlinearly and therefore make signal detection difficult.

Fig. 8. Manhattan plot and linkage disequilibrium (LD) heatmap over 3.7-Mb region (45.03 – 48.37 Mb) on Pv07 for lodging. This region includes the three significant peaks (Pv07 [45.18 Mb], Pv07 [46.11 Mb], and Pv07 [48.63 Mb]) and their associated candidate gene homologs in Arabidopsis. The marker at 46.11 Mb, which resides in Phvul.007G221700, is colored in blue in the Manhattan plot and the rest of the markers in this region are color coded based on their pairwise $r^2$ value with this marker. Single-nucleotide polymorphisms (SNPs) with $r^2 < 0.25$ are white colored, those with $0.25 \leq r^2 \leq 0.5$ are orange, and SNPs with $r^2 > 0.5$ are colored in red. The $r^2$ values in Manhattan plot and LD heatmap are corrected for population structure and relatedness based on the best model using all the SNPs with minor allele frequency $\geq 5\%$. 

MOGHADDAM ET AL.: GWAS ON AGRONOMIC TRAITS IN COMMON BEAN 13 OF 21
The STRUCTURE analysis, when clustered according to race and market classes, revealed patterns of admixture. Market classes in common bean are distinguished based on seed size, shape, and color. Therefore, admixed lines within a market class would occur based on only a few traits. Admixture observed among the genotypes within market classes at K = 7 reflects the breeding strategies used in bean improvement. During the first half of the 21st century, bean production, and therefore breeding in United States, was regional, and crosses were made within market classes or between closely related market classes in that region. At that time, pinto and great northern bean were grown in the Central Great Plains, and in the far West, pink, great northern, red Mexican, and pinto bean were the main bean market classes (McClean et al., 1993). Even today, breeders commonly make crosses among black, small red, and navy bean in race Mesoamerica and between pinto and great northern bean in race Durango. Black bean has tropical origin and are more recent to North America. In Central America, black bean has been used to improve red bean by introgressing yield and pest-resistance traits. Thus, for some red bean lines, the majority of their pedigree consists of black bean parents (Beebe et al., 1995). Moreover, breeding for new red and pink bean varieties in recent decades involved the introduction of plant architectural traits from the Mesoamerican black and navy bean market classes (Kelly, 2001; Vandemark et al., 2014). The pink, red, and pinto bean were landraces in the early 1900s that were crossed among each other primarily to introgress disease resistance traits. Even though some lines are admixed, general clustering occurs by market class (Miklas, 2000).

Patterns of Linkage Disequilibrium

Linkage disequilibrium patterns vary by specific chromosome regions and between populations (Tailon-Miller et al., 2000; Huttley et al., 1999; Laan and Pääbo, 1997; Huang et al., 2010). Evidence shows that smaller populations show higher level of LD (Laan and Pääbo, 1997; Flint-Garcia et al., 2003). Within the MDP, race MA is a smaller subpopulation than DJ, which could have resulted in higher levels of LD as well as lower polymorphism rate within that race. Many markers with long-range interchromosomal LD in the MDP and DJ subpopulation were monomorphic in MA subpopulation or had a MAF < 5%. On the other hand, the differences in LD between specific regions of the genome might be stronger that the effect of population on LD (Zavattari et al., 2000). For example, both subpopulations exhibited higher LD in the pericentromeric than the ends of the chromosomes. In contrast to populations in linkage equilibrium, where LD only depends on the relative rate of mutation and recombination, the patterns of LD within the MA and DJ subpopulations depend on (i) the evolutionary steps that led to the divergence within the wild Mesoamerican gene pool, (ii) domestication events within that gene pool, (iii) the development of the landraces, and (iv) the effects of artificial selection for beneficial alleles specific to each race during variety development (Gepts and Debouck, 1991; Kwak and Gepts, 2009; Hamblin et al., 2011; Bitocchi et al., 2012; Mamidi et al., 2013; Schmutz et al., 2014). All these factors result in changes in allele frequencies as a result of genetic drift or selection. Some allelic combinations can be lost, leading to higher LD between the remaining alleles. Strong selection during breeding in conjunction with local adaptation can create LD between unlinked loci, which does not decay with a longer genetic distance (Long et al., 2013). The SNP allele frequencies and LD patterns within the two subpopulations in the MDP clearly varied. Any conclusion on how the demographic factors shaped the LD pattern in this population needs in-depth population genetic studies using data from higher-pass sequencing of genotypes.

Linkage Disequilibrium affects the Interpretation of Genome-Wide Association Study Results

The GWAS for days to flower within the MDP identified two SNP peaks that mapped near each other (~8 and ~15 Mb on Pv01). These two peaks are located in the low-recombination pericentromeric region of the chromosome, and the LD value was r² = 0.8 with the null model, a value that dropped to r² = 0.4 when LD was corrected for population structure and individual relatedness. An important question is whether these Pv01 peaks were significant because of LD or whether they both might contain genes associated with the trait. To evaluate the possibilities, we compared the GWAS results with those obtained with MLMM. At the 0.01% cutoff, a group of linked SNPs were significant at ~15 Mb in the MA subpopulation, while within the DJ subpopulation, a group of linked SNPs were significant at ~8 Mb. The best MLMM forward and backward stepwise regression model for the MA subpopulation only included the 15-Mb peak SNPs, while the best MLMM model for the DJ subpopulation only included the 8-Mb peak SNPs. Therefore, the result observed with the full MDP was actually the union of the effects of the two subpopulations and is not a spurious result resulting from LD.

Interrelated Traits show Colocalized Genome-Wide Association Study Peaks

Multiple traits exhibited colocalized GWAS signals. Most prominent among these was a Pv07 peak observed in nearly all test locations for lodging, canopy height, and growth habit (when determinate genotypes are excluded). This region spans the 45- to 48-Mb region and is the main peak for lodging and canopy height. Stepwise regression defined the 46-Mb region on Pv07 as the most significant region for these three interrelated traits. This could indicate the existence of a gene with a pleiotropic effect for the three traits or that these are interrelated subtraits associated with plant architecture. The Pv07 (46
Candidate Genes

Days to Flower
Flowering is controlled by a complex network of genes that collectively integrate the photoperiod, vernalization, gibberellin, autonomous, and flowering time clock pathways. These pathways merge into two floral integrator genes, FLOWERING LOCUS T (FT) and SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 (SOC1) that promote flowering in Arabidopsis (Fornara et al., 2010; Chen et al., 2015). Although neither of these genes was associated with a days-to-flowering GWAS peak, multiple genes involved in the flowering pathways were identified as candidate genes. One candidate, *Phvul.001G064600* (Pv01 [8 Mb]), is an ortholog of MYB56, a gene that negatively regulates FT (Chen et al., 2015). The peak marker for the df1.1 days-to-flowering QTL (Blair et al., 2006) also maps ~1 Mb from the MYB56 candidate. Another candidate, *Phvul.001G192200*, is located near the Pv01 (45 Mb) GWAS peak and is an ortholog of LIGHT-REGULATED WDI (LWD1). Alternate alleles at this gene change the expression periodicity of genes in the circadian rhythm pathway leading to earlier flowering (Wu et al., 2008). This acceleration of flowering is mediated through CONSTANS (CO), which activates the integrator FT. A second days-to-flowering candidate gene at the same Pv01 (45 Mb) GWAS peak, *Phvul.001G192300*, is a SPINDLY (SPY) ortholog and is located immediately adjacent to LWD1. SPINDLY delays flowering by suppressing the gibberellic acid pathway and directly interacts with GI1 (GI), a gene within the photoperiod pathway (Tseng et al., 2004). *Phvul.001G094300*, an ortholog of Arabidopsis BRG-1 ASSOCIATED FACTOR 60 (AtBAF60), maps near the Pv01 (20 Mb) GWAS peak. AtBAF60 is a subunit of the SWI/SNF polycomb complex, a complex that regulates the photoperiod and vernalization pathway by directly interacting with and suppressing FLC (Jégu et al., 2014). FLC is a flowering-time suppressor that functions upstream of FT and SOCI. *Phvul.007G213900*, a homolog of Arabidopsis SWINGER (SWN), maps near the Pv07 (45 Mb) peak. SWINGER is a component of the polycomb repressive complex 2 (PRC2) that promotes flowering via FT through the suppression of the flowering suppressor FLC (Wood et al., 2006; Jiang et al., 2008). Polycomb repressive complex 2 not only affects flowering time, but also affects floral meristem development. The release of PRC2 from a repressive complex that includes KNUCKLES (KNU) leads to determinate flower development. The bean ortholog of KNU, *Phvul.001G087900*, maps near the Pv01 (15 Mb) days-to-flowering GWAS peak. When transitioning to a determinate floral meristem, KNU is activated and functions in a feed-back loop that promotes determinate development. This results in flower set over a defined time period (Lenhart et al., 2001; Payne et al., 2004; Sun et al., 2009, 2014; Liu et al., 2011; Wu et al., 2012; Mozgova et al., 2015). Figure 9 maps the common bean days-to-flowering candidate genes relative to the Arabidopsis flowering pathway.

Days to Maturity
Since the biological pathway that controls progress toward maturity is not fully elucidated, we searched for candidate genes known (or hypothesized) to be involved in senescence, nutrient remobilization, and seed development. Flowering time candidates were also considered because of the positive correlation between days to flowering and days to maturity (Córdoba et al., 2010; Blair et al., 2012). The GWAS peak at Pv11 (4.3 Mb) maps within the bean ortholog of the Arabidopsis TARGET of RAPAMYCIN (TOR), *Phvul.011G053000*. TOR is a member of a signaling pathway that perceives the nutrient and energy status of the plant (Wullschleger et al., 2006; Kim and Guan, 2011) and controls senescence and nutrient recycling by regulating SERINE/THREONINE

Moghaddam et al.: GWAS on Agronomic Traits in Common Bean 15 of 21
PROTEIN PHOSPHATASE 2A (PP2A) activity (MacKintosh, 1992). Enhanced TOR expression increases organ and cell size and seed production, two biological processes required for the plants migration toward maturity (Deprost et al., 2007). Phvul.004G011400, an ortholog of ATMYB118, is a candidate gene at the Pv04 (20 Mb) GWAS peak. ATMYB118 is an Arabidopsis transcription factor that regulates seed maturation at the spatial level by suppressing maturation related genes within the endosperm of seeds and maintains the embryo as the major nutrient sink (Barthole et al., 2014).

Three days-to-maturity candidate genes function in flowering pathways. Phvul.011G053300, maps near the Pv11 (4 Mb) peak and is an ortholog of the Arabidopsis FPA gene. FPA is a member of the autonomous flowering pathway (Koornneef et al., 1991), and genotypes overexpressing FPA flower earlier in short days and are insensitive to photoperiod (Schomburg, 2001). Its role in maturation is unclear, but early flowering is often correlated with early maturity. The bean ortholog of REDUCED VERNALIZATION RESPONSE 1 (VRN1), Phvul.011G050800, maps to the same Pv11 (4 Mb) peak as FPA. Levy et al. (2002) reported that constitutive overexpression of VRN1 led to early flowering without vernalization in Arabidopsis. Another days-to-maturity candidate gene related to flowering is Phvul.011G158300, which maps at the Pv11 (41 Mb) peak. This gene is homologous to one of the two Arabidopsis SHL genes (At4g22140), which encode a transcriptional regulator and chromatin remodeling protein. Genotypes overexpressing SHL lead to early flowering and senescence (Müssig and Altmann, 2003).

**Growth Habit, Lodging, and Canopy Height**

Plant architecture is a complex trait that involves many hormone and shoot branching pathways that affect main stem length. A major signal in the Pv11 (43 Mb) peak contains four significant SNPs located inside Phvul.011G164800, an ortholog of Arabidopsis SPL4. SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL) is one regulator of Arabidopsis BRC1 transcription factor that integrates the hormone and branching pathways. Three SNPs are in complete LD with each other, and two of these cause nonsynonymous substitutions (Cys145Gly and Ser146Asn). In Arabidopsis, SPL3, SPL4, and SPL5 act redundantly to affect the timing of the juvenile to adult phase transition (Wu and Poethig, 2006; Schwarz et al., 2008; Wang et al., 2009; Poethig, 2009; Yamaguchi et al., 2009). SPL genes in rice (Oryza sativa L.) and alfalfa (Medicago sativa L.) have been reported to cause architectural changes, such as lodging, internode length, branching, and biomass (Jiao et al., 2010; Aung, 2014), and have a pleiotropic effect on shoot branching, plant height, and floral induction. These pleiotropic effects of the candidate gene might explain why early efforts to create early flowering type II architecture in common bean were not successful (Kelly, 2001). The Arabidopsis and bean homologs both contain a miR156 cleavage site in their 3' untranslated region, and branching may be regulated by the miR156-SPL pathway.

Three other candidate genes act in the auxin pathway upstream of cytokinin and strigolactone to regulate the auxiliary bud outgrowth through TB1/BRC1 (Müller and Leyser, 2011; Rameau et al., 2014). Phvul.006G203400, at the Pv06 (30 Mb) peak, is an ortholog of SUPPRESSOR OF AÇAULIS 51 (SAC51), which encodes a bHLH (basic helix-loop-helix) transcription factor that controls stem elongation in Arabidopsis. SAC51 uORF-mediated translational regulation is affected by thersmostpermine, which limits the auxin signaling that promote xylem differentiation. The GWAS peak at Pv07 (45.8 Mb) includes a candidate gene (Phvul.007G218900) homologous to SUPPRESSOR OF AUXIN RESISTANCE 1 (SARI) and...
appears to control stem thickness. Sar1 is epistatic to AUXIN RESISTANCE 1 (axr1) and together the two genes increase plant height and internode distance in *Arabidopsis* (Cernac et al., 1997; Parry et al., 2006). The GWAS peak at Pv04 (2.9 Mb) is located in the candidate gene *Phvul.004G027800*, a homolog to *ABC\text{B9}* (MDRI) an ABC transporter protein that mediates polar auxin transport in stems and roots (Noh et al., 2001; Kaneda et al., 2011). This gene affects apical dominance (Noh et al., 2001), hypocotyl gravitropism, and inflorescence stem phototropism and curvature, all factors associated with plant architecture (Noh et al., 2003; Kumar et al., 2011).

As noted above, several lodging and canopy-height peaks were colocalated. Another example is the Pv07 (48 Mb) GWAS peak that is located in *Phvul.007G246700*. This gene encodes a homolog of *AtPME41*, an invertase–pectin methylesterase inhibitor. This protein changes the methylesterification level of pectin, a modification that affects cell wall rigidity (Micheli, 2001; Castillejo et al., 2004). This may be a factor that favors an upright, nonlodged architecture. Similarly, a GWAS on soybean agronomic traits reported a pectin lyase-like superfamily protein as a candidate gene for plant height (Zhang et al., 2015).

**Seed Weight**

Seed weight is a complex trait that encompasses many developmental and physiological processes, such as plant architecture, seed development, and metabolic efflux, to enhance sink strength (Van Camp, 2005). There was no clear signal in the MA subpopulation for seed weight, but Pv08 and Pv10 showed strong peaks in the DJ subpopulation. The Pv10 candidate gene, *Phvul.010G017600*, affects metabolic efflux and protein storage in seed. This gene is a homolog of *Arabidopsis\text{ALPHA-AMYLASE LIKE GENE (AMY)}* that is associated with the onset of suspensor and endosperm programmed cell death and early nutrient mobilization to nourish the growing embryo (Mansfield and Briarty, 1994; Sreenivasulu and Wobus, 2013). The significant peak at Pv08 (1 Mb) within the DJ subpopulation contains *Phvul.008G013300*, a subtilase gene family member that is homologous to *SBT1.1* in *Arabidopsis*. In legumes, this gene is specifically expressed during the early stages of endosperm development when the embryo cells are still dividing. It is hypothesized that it provides developmental signals or regulates metabolic flux to control embryo growth by affecting the embryo cell number and therefore might have a role in regulating seed size (Abirached-Darmency et al., 2005; Melkus et al., 2009; D’Erfurth et al., 2012). *Phvul.006G069300*, at the Pv06 (18 Mb) GWAS peak in the Michigan environment, is homologous to *Arabidopsis\text{ASN1}*, a glutamine-dependent asparagine synthase 1 and a domestication gene in the Middle American gene pool (Schmutz et al., 2014). *Arabidopsis* transgenic plants overexpressing *ASN1* show increased N content, higher seed-soluble protein content, and increased seed weight (Lam et al., 2003). Many QTL analyses in common bean have located seed-weight QTL on the same chromosomes as described here (Beattie et al., 2003; Blair et al., 2006, 2012; Pérez-Vega et al., 2010; Wright and Kelly, 2011; Linares-Ramirez, 2013).

**Conclusion**

Genome-wide association studies using a diverse panel and a large set of SNPs was able to dissect the genetic architecture of important economic traits in common bean and define promising candidate genes for six agronomic traits. These candidates are involved in many different molecular biochemical, hormone, developmental, and transcriptional regulation pathways. Adding phenotypic data from multiple years to this data set can improve the accuracy of the results. Moreover, increasing the number of individuals in the MA subpopulation can yield more polymorphic SNPs and hence facilitate the discovery of MA specific signals for some traits. It is noteworthy that the GWAS yielded more promising results for traits with higher heritability than those that are affected more by the environment. Moreover, the extent of LD, especially for signals falling in the heterochromatic region, is wide in this population, which made candidate gene discovery more challenging. The population and genotyping tools developed during this project are now available to study many other phenotypes that are important for the productivity of common bean. These tools also can provide an additional starting point for the discovery, validation, and functional investigation of genes that control the many unique features of common bean that are also shared by other legume species.

**Acknowledgments**

Financial support provided by USDA, National Institute of Food and Agricultural (NIFA), Agriculture and Food Research Initiative (AFRI), Project #2009-01929

**References**


advanced intercross-recombinant inbred lines. PLoS One 4:e4318. doi:10.1371/journal.pone.004318


Sun, B., L.S. Looi, S. Guo, Z. He, E.S. Gan, J. Huang, Y. Xu, W.Y. Wee, and T. Itô. 2014. Timing mechanism dependent on cell division is invoked by Polycamb evocation in plant stem cells. Science 343:1248559. doi:10.1126/science.1248559


