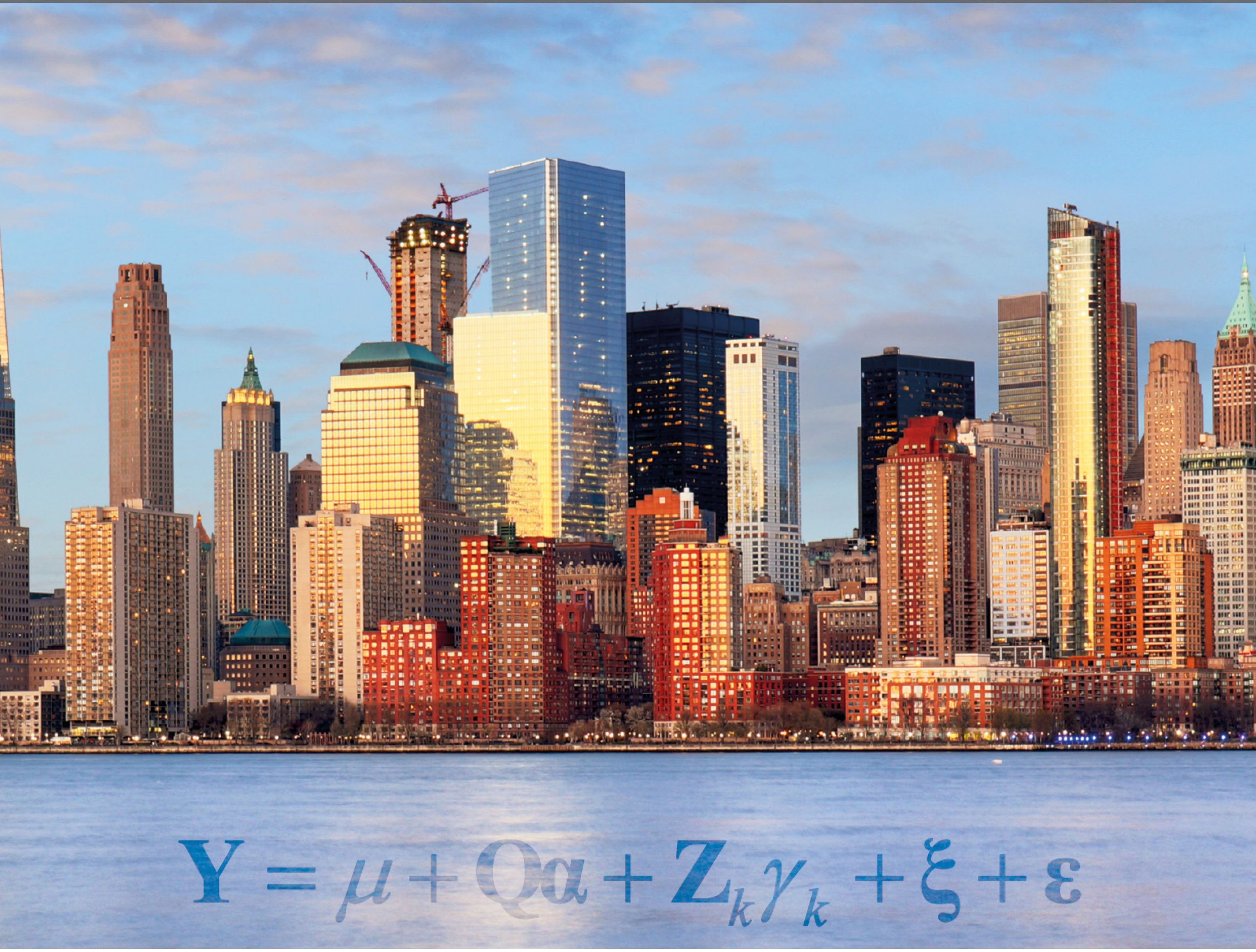
THE APPLICATIONS OF NEW MULTI-LOCUS GWAS METHODOLOGIES IN THE GENETIC DISSECTION OF COMPLEX TRAITS

EDITED BY: Yuan-Ming Zhang, Zhenyu Jia and Jim M. Dunwell
PUBLISHED IN: Frontiers in Plant Science


$\mathbf{Y} = \mu + \mathbf{Q}\alpha + \mathbf{Z}_k\gamma_k + \xi + \varepsilon$

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-834-9
DOI 10.3389/978-2-88945-834-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

THE APPLICATIONS OF NEW MULTI-LOCUS GWAS METHODOLOGIES IN THE GENETIC DISSECTION OF COMPLEX TRAITS

Topic Editors:
Yuan-Ming Zhang, Huazhong Agricultural University, China
Zhenyu Jia, University of California, Riverside, United States
Jim M. Dunwell, University of Reading, United Kingdom

$\mathbf{Y} = \mu + \mathbf{Q}\alpha + \mathbf{Z}_k\gamma_k + \xi + \varepsilon$

Image: TTstudio/Shutterstock.com

Genome-Wide Association Studies (GWAS) are widely used in the genetic dissection of complex traits. Most existing methods are based on single-marker association in genome-wide scans with population structure and polygenic background controls. To control the false positive rate, the Bonferroni correction for multiple tests is frequently adopted. This stringent correction results in the exclusion of important loci, especially for GWAS in crop genetics. To address this issue, multi-locus GWAS methodologies have been recommended, i.e., FASTmrEMMA, ISIS EM-BLASSO, mrMLM, FASTmrMLM, pLARmEB, pKWmEB and FarmCPU.

In this Research Topic, our purpose is to clarify some important issues in the application of multi-locus GWAS methods. Here we discuss the following subjects:

First, we discuss the advantages of new multi-locus GWAS methods over the widely-used single-locus GWAS methods in the genetic dissection of complex traits, metabolites and gene expression levels.

Secondly, large experiment error in the field measurement of phenotypic values for complex traits in crop genetics results in relatively large P-values in GWAS, indicating the existence of small number of significantly associated SNPs. To solve this issue, a less stringent P-value critical value is often adopted, i.e., 0.001, 0.0001 and 1/*m* (*m* is the number of markers). Although lowering the stringency with which an association is made could identify more hits, confidence in these hits would significantly drop. In this Research Topic we propose a new threshold of significant QTN (LOD=3.0 or P-value=2.0e-4) in multi-locus GWAS to balance high power and low false positive rate.

Thirdly, heritability missing in GWAS is a common phenomenon, and a series of scientists have explained the reasons why the heritability is missing. In this Research Topic, we also add one additional reason and propose the joint use of several GWAS methodologies to capture more QTNs. Thus, overall estimated heritability would be increased.

Finally, we discuss how to select and use these multi-locus GWAS methods.

Citation: Zhang, Y.-M., Jia, Z., Dunwell, J. M., eds. (2019). The Applications of New Multi-Locus GWAS Methodologies in the Genetic Dissection of Complex Traits. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-834-9

Table of Contents

frontiers in Plant Science

EDITORIAL
published: 11 February 2019
doi: 10.3389/fpls.2019.00100

Editorial: The Applications of New Multi-Locus GWAS Methodologies in the Genetic Dissection of Complex Traits

Yuan-Ming Zhang[1]*, *Zhenyu Jia*[2] *and Jim M. Dunwell*[3]

[1] *Crop Information Center, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China,* [2] *Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA, United States,* [3] *School of Agriculture, Policy and Development, University of Reading, Reading, United Kingdom*

Keywords: genome-wide association study, mixed linear model, multi-locus model, mrMLM, omics big dataset

Editorial on the Research Topic

The Applications of New Multi-Locus GWAS Methodologies in the Genetic Dissection of Complex Traits

OPEN ACCESS

Edited and reviewed by:
Luigi Cattivelli,
Council for Agricultural and Economics Research, Italy

****Correspondence:***
Yuan-Ming Zhang
soyzhang@mail.hzau.edu.cn

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *21 December 2018*
Accepted: *22 January 2019*
Published: *11 February 2019*

Citation:
Zhang Y-M, Jia Z and Dunwell JM (2019) Editorial: The Applications of New Multi-Locus GWAS Methodologies in the Genetic Dissection of Complex Traits. Front. Plant Sci. 10:100. doi: 10.3389/fpls.2019.00100

Since the establishment of the mixed linear model (MLM) method for genome-wide association studies (GWAS) by Zhang et al. (2005) and Yu et al. (2006), a series of new MLM-based methods have been proposed (Feng et al., 2016). These methods have been widely used in genetic dissection of complex and omics-related traits (**Figure 1**), especially in conjunction with the development of advanced genomic sequencing technologies. However, most existing methods are based on single marker association in genome-wide scans with population structure and polygenic background controls. To control false positive rate, Bonferroni correction for multiple tests is frequently adopted. This stringent correction results in the exclusion of important loci, especially for large experimental error inherent in field experiments of crop genetics. To address this issue, multi-locus GWAS methodologies have been recommended, i.e., mrMLM (Wang et al., 2016), ISIS EM-BLASSO (Tamba et al., 2017), pLARmEB (Zhang et al., 2017), FASTmrEMMA (Wen et al., 2018a), pKWmEB (Ren et al., 2018), and FASTmrMLM (Zhang and Tamba, 2018). Here we summarize their advantages and potential limitations for using these methods (**Table 1**).

MULTI-LOCUS GENOME-WIDE ASSOCIATION STUDIES FOR COMPLEX TRAITS

Comparison of GWAS Methodologies

Our methodological papers have showed their advantages in terms of quantitative trait nucleotide (QTN) detection power and QTN effect estimation accuracy over existing methods (Wang et al., 2016; Tamba et al., 2017; Zhang et al., 2017; Ren et al., 2018; Wen et al., 2018a). This conclusion has been echoed in a number of other applied studies in this Research Topic. For example, Ma et al. and Zhang et al. indicated that mrMLM, FASTmrEMMA, pLARmEB, and ISIS EM-BLASSO outperform the R package GAPIT, with ISIS EM-BLASSO being the most powerful multi-locus approach. Xu et al. compared one single-locus method (GEMMA) and three multi-locus methods (FASTmrEMMA, FarmCPU, and LASSO) in the genetic dissection of starch pasting properties in maize. As a result, FASTmrEMMA detected the most QTNs (29), followed by FarmCPU (19) and LASSO (12), and GEMMA detected the least QTNs (7). In the genetic dissection of salt tolerance traits in rice, Cui et al. compared all the six multi-locus approaches and identified the most

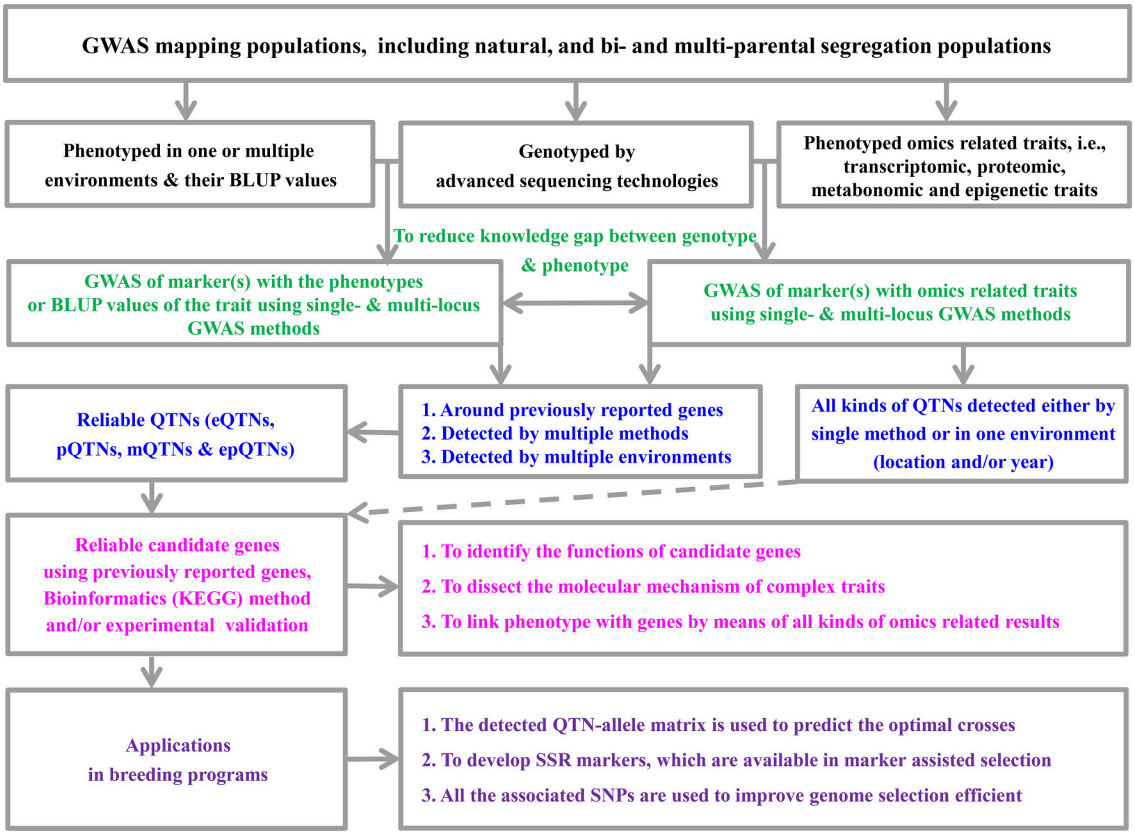

GWAS mapping populations, including natural, and bi- and multi-parental segregation populations
Phenotyped in one or multiple environments & their BLUP values
Genotyped by advanced sequencing technologies
Phenotyped omics related traits, i.e., transcriptomic, proteomic, metabonomic and epigenetic traits
To reduce knowledge gap between genotype & phenotype
GWAS of marker(s) with the phenotypes or BLUP values of the trait using single- & multi-locus GWAS methods
GWAS of marker(s) with omics related traits using single- & multi-locus GWAS methods
Reliable QTNs (eQTNs, pQTNs, mQTNs & epQTNs)
1. Around previously reported genes
2. Detected by multiple methods
3. Detected by multiple environments
All kinds of QTNs detected either by single method or in one environment (location and/or year)
Reliable candidate genes using previously reported genes, Bioinformatics (KEGG) method and/or experimental validation
1. To identify the functions of candidate genes
2. To dissect the molecular mechanism of complex traits
3. To link phenotype with genes by means of all kinds of omics related results
Applications in breeding programs
1. The detected QTN-allele matrix is used to predict the optimal crosses
2. To develop SSR markers, which are available in marker assisted selection
3. All the associated SNPs are used to improve genome selection efficient

FIGURE 1 | The pipeline framework of genome-wide association studies and their application.

co-detected QTNs from ISIS EM-BLASSO. Peng et al. used our six multi-locus GWAS methods to analyze 20 free amino acid levels in kernels of bread wheat (*Triticum aestivum* L.) and found the reliability and complementarity of these methods. In the detection of small-effect QTNs for fiber-quality related traits in the early-maturity varieties of upland cotton, Su et al. claimed that the multi-locus GWAS methods are more powerful and robust than the MLM method in TASSEL v5.0. Hou et al. demonstrated that 20 QTNs were associated with drought stress response using mrMLM, while three QTNs were associated with resistance to Verticillium wilt using EMMAX. Although the above studies have shown the advantages of multi-locus GWAS methods over single-locus GWAS methods, Chang et al., He et al., Li et al., and Xu et al. recommended the combination of single-locus methods and/or multi-locus methods to improve the detection power and robustness of GWAS, and Cui et al. recommended adding a bin analysis to the models or developing a hybrid method that merges the results from different methods. Our previous results in the analysis of real and simulated dataset support the above recommendations.

In addition, Liu et al. adopted four multi-locus GWAS algorithms (mrMLM, FASTmrEMMA, ISIS EM-BLASSO, and pLARmEB) to dissect the genetic foundation for fiber quality and yield component traits in RILs. As a result, a significant number of QTNs were found to coincide with the physical regions of the confidence intervals of reported QTLs, demonstrating the effectiveness and feasibility of multi-locus GWAS methods in RILs.

The Critical *P*-Value or LOD Score for Significant QTN

In single-locus GWAS, one key concern is the high false positive rate (FPR). To reduce FPR, Bonferroni correction is frequently applied in the single-locus methods, including EMMAX (Kang et al., 2010), GEMMA (Zhou and Stephens, 2012), ECMLM (Li et al., 2014), and MLM (Yu et al., 2006). In human genetics, the genome-wide significance *P*-value threshold of 5×10^{-8} has become a standard for common-variant GWAS (Barsh et al., 2012; Fadista et al., 2016; Chang et al., 2018). However, this correction or the critical *P*-value in human genetics is too stringent to detect certain associated loci for complex traits in crop genetics. To address this issue, a modified Bonferroni correction has been proposed; in other words, the number of markers (m) in the correction formulas is replaced by the effective number of markers (m_e) (Wang et al., 2016; Guan et al.). In real data analysis in crop genetics, some subjective and less stringent *P*-value thresholds for significant level are frequently applied owing to large experimental error, i.e., $1/m$ (m is the number

TABLE 1 | Comparison of single- and multi-locus GWAS methodologies.

	Single-locus GWAS	**Multi-locus GWAS***
QTN detection power	Low	High
P-value threshold of significant QTN	5×10^{-8} (human genetics for common variants) $0.05/m \sim 1/m$ (crop genetics; *m* is no. of markers)	2×10^{-4} (or LOD = 3.0)
False positive rate	Low (with Bonferroni correction)	Low (with LOD = 3.0 or $P = 2 \times 10^{-4}$)
Multiple test correction	Yes	No
Polygenic background control	Yes	Yes (First step); No (Second step; all the potential genes have been included)
Population structure control	Yes	Yes
SNP effect	Fixed	Random
No. of variance components	Two (polygenic background and residual variances)	Three (QTN, polygenic background and residual variances; First step)
Multi-locus genetic model	No	Yes (second step)
How to reduce no. of variances	a) To fix the polygenic-to-residual variance ratio b) To estimate residual variance along with fixed effects	a) To fix the polygenic-to-residual variance ratio (1∼5) b) To estimate residual variance along with fixed effects (1∼4) c) Let the number of non-zero eigenvalues of $X_CX_C^T$ be one (3∼5) d) To whiten the covariance matrix of polygenic K and noise (3∼5)
Running time	Fast (GEMMA & EMMAX), slow (EMMA)	Fast (2, 6), slow (5), moderate (others)
Software	GEMMA: http://www.xzlab.org/software.html EMMAX: http://genetics.cs.ucla.edu/emmax	mrMLM: https://cran.r-project.org/web/packages/mrMLM/index.html mrMLM.GUI: https://cran.r-project.org/web/packages/mrMLM.GUI/index.html Parallel calculation with multiple CPU; quickly read big datasets; graphical user interface (GUI); To continuously run the programs for multiple traits

**mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, pKWmEB, and ISIS EM-BLASSO are marked by 1, 2, 3, 4, 5, and 6 respectively.*

of markers) (Li et al.; Xu et al.), 10^{-5} (Misra et al.), and 10^{-4} (Chang et al.). To balance high QTN detection power and low false positive rate, Xu et al. replaced Bonferroni correction by a less stringent criterion ($1/m$) for GEMMA, and a satisfactory result was achieved in their Monte Carlo simulation studies.

Theoretically, correction for multiple tests is unnecessary in multi-locus GWAS because all the potential genes or loci for complex traits are fitted to a single linear model and their effects are estimated and tested simultaneously. For example, 0.05 was chosen as the *P*-value threshold in QTN detection of Khan et al. (2018). Although relaxing the stringency of significance level in multi-locus GWAS can identify more hits, confidence in these hits will drop significantly. Thus, Segura et al. (2012) and Liu et al. (2016) imposed Bonferroni correction on QTN detection in their multi-locus GWAS methods. Our results indicated that Bonferroni correction in multi-locus GWAS of (Segura et al., 2012) and Liu et al. (2016) may be too stringent, while the cutoff of 0.05 in multi-locus GWAS of Khan et al. (2018) may be too relaxed due to the fact that a significance level of 0.05 can result in a high false positive rate. Lü et al. simply used LOD score ≥ 5 as a threshold for QTN detection in their multi-locus GWAS. Based on our studies, we proposed using LOD = 3.0 (or $P = 0.0002$) as a cutoff in multi-locus GWAS to balance the high power and low false positive rate for QTN detection.

Heritability Missing in GWAS

Heritability missing is a common issue in GWAS (Maher, 2008). Human geneticists ascribe heritability missing to a few reasons, including rare alleles, gene-by-gene and gene-by-environment interactions, and miniature genetic effects of DNA variants that can hardly reach the level of genome-wide significance (Eichler et al., 2010). In our opinion, the stringent threshold in genome-wide detection is also a factor, because certain QTNs cannot meet the significant level if such *P*-value cutoff is applied. This viewpoint is supported by the simulation results of Xu et al.

In most GWAS methodologies, the genotypes of a SNP, for example, QQ, Qq, and qq, are conventionally coded as 2, 1, and 0, respectively. Thus, the estimated QTN effect is actually the average effect of allelic substitution, being $a + (q - p)d$. Let $a + (q - p)d = 0$, then $d = a/(p - q)$. Where p takes different values, such as $p = 0.1, 0.3, 0.5, 0.7$, and 0.9, so $d = -1.25\,a$, $-2.5\,a$, ∞, $2.5\,a$, and $1.25\,a$, respectively, indicating the difficulty in the detection of QTNs with over-dominance. This may be another reason for the heritability missing.

New methodologies have been proposed to handle heritability missing, for example, GCTA (Yang et al., 2011) and GREML-LDMS (Yang et al., 2015). In this Research Topic, we suggest that part of the missing heritability may be regained by using multi-locus GWAS methods, since more QTNs can be detected and overall estimated heritability will be increased.

HOW TO DETERMINE RELIABLE QTNS AND MINE RELIABLE CANDIDATE GENES?

How to Determine Reliable QTNs?

Firstly, when several multi-locus methods are used to analyze a same dataset, the QTNs identified by multiple approaches are usually reliable. For example, all the 31 genomic regions associated with four photosynthesis related traits were detected by at least three multi-locus methods in Lü et al., five QTNs associated with forage quality-related traits were detected by at least two methods in Li et al., and all the common QTNs either between single-locus methods and multi-locus methods, or across several multi-locus methods were declared in Misra et al. Secondly, the QTNs near previously reported trait-associated genes should be reliable. For example, the QTNs around genes *GRMZM2G163761*, *GRMZM2G412611,* and *GRMZM2G066749* likely contribute to the callus regenerative capacity (Ma et al.), the QTNs around genes *GRMZM2G032628* (*ae1*) and *GRMZM2G392988* may be associated with starch biosynthesis (Xu et al.), and the QTNs around genes *Gh_D102255* and *Gh_A13G0187* perhaps participate in cellular activities for fiber elongation (Liu et al.). Finally, the QTNs identified across various environments (locations and/or years) are also reliable, i.e., Liu et al. identified 57 QTNs that were associated with cotton fiber quality and yield components in at least two environments; Hu et al. repeatedly detected 39 QTNs clusters to be associated with 14 agronomic traits in 122 barley doubled haploid lines in multiple environments; Zhang et al. repeatedly detected 22 common QTNs to be associated with protein content in 144 soybean four-way recombinant inbred lines in 20 environments.

How to Mine Reliable Candidate Genes?

All known genes in the regions around reliable QTNs potentially contribute to the traits of interest. However, only a subset of them may be reliable candidate genes which are worthy of further investigation. We can use homolog (previously reported genes) in other species, e.g., *Arabidopsis thaliana*, to mine reliable candidate genes in these regions. For example, *WOX2* in *Arabidopsis* has been reported to increase the rate of resistant seedlings from transformed immature embryos in maize and, therefore, the homologous gene *GRMZM2G108933* might play an important role in controlling maize callus regeneration (Ma et al.). Bioinformatics approaches, such as the KEGG pathway analytic tool, may be used for mining reliable candidate genes and relevant gene networks. For example, two genes (*LOC_Os01g45760* and *LOC_Os10g04860*) are found to be involved in auxin biosynthesis in rice using KEGG (Cui et al.). Experimental validations are often needed to confirm the associations between these candidate genes and the traits of interest. For instance, RNA-seq analysis and qRT-PCR experiments verified that four genes (*RD2*, *HAT22*, *PIP2*, and *PP2C*) are associated with drought tolerance in cotton (Hou et al.); genomic DNA sequencing showed that two candidate genes *BnaA08g08280D* and *BnaC03g60080D* are different between the high- and low-oleic acid lines (Guan et al.). The combined use of GWAS and experimental validation has great potential for detection of new genes and their biological functions. For example, a new gene *GRMZM2G065083* was found by Xu et al. to play a critical role in starch biosynthesis in maize by being involved in the gluconeogenesis process, hexose biosynthetic and metabolic process, and glucose-6-phosphate isomerase activity, providing insights into the molecular mechanism underlying the pasting properties of maize starch.

Important genes may be missed if we only select consensus QTNs identified by more than one methodology or in more than one experiment/environment. In practice, we found that some QTNs detected by only one multi-locus method or one environment may lead to important discoveries. These QTNs may be used to mine candidate genes through network analysis using bioinformatics analysis and/or experimental validation.

How to Make Use of the GWAS Results?

The main product of GWAS includes the detected QTNs and the candidate genes nearby. Three approaches are available for applying these results to breeding programs. Firstly, one can organize the detected QTN-allele matrix as the population genetic constitution to facilitate the selection of optimal crosses. For example, the top 10 optimal crosses were predicted according to their 95th percentile weighted average values (Khan et al., 2018). Secondly, we can develop SSR markers around the reliable QTNs and utilize them in marker assisted selection of crops (Li et al., 2018). Thirdly, all the SNPs that are significantly associated with the trait of interest can be used for improving genome selection (He et al.; He et al., 2019).

Figure 1 summarizes how to design a GWAS to identify QTNs and mine candidate genes, of which the biological functions may be further investigated or validated at a molecular level.

FUTURE PERSPECTIVES

It is becoming common to use multiple statistical methods to detect major quantitative trait loci (QTLs) in the linkage analyses of complex traits. Thus, we recommend using a few GWAS methods, especially several multi-locus GWAS methods which do not need correction for multiple comparisons, to investigate complex traits. However, not all QTNs can be identified by all these methods, posing difficulties for using these GWAS results. This may be ascribed to the fact that the various GWAS models are based upon different genetic or statistical assumptions. Possible solutions have been provided in this editorial to compare the results from various GWAS models and screen for candidate QTNs or genes, facilitating the subsequent validation or application.

Interaction at different omics levels, including QTL-by-environment and QTL-by-QTL interactions, can be detected with various software programs in linkage analysis. Nevertheless, methods and software programs with comparable function are quite limited in GWAS, especially for the studies of quantitative traits in natural populations where large numbers of genomic markers are analyzed. The number of variables in GWAS models will increase sharply if interactions are considered, challenging both computational efficiency and detection power. Multicollinearity among highly saturated and linked markers is another issue in GWAS, which impairs the efficiency and accuracy of the current statistical methods. Innovative strategies are needed to distill many thousands of variables by removing the redundant genomic markers such that the computational burden and impact from multicollinearity can be reduced and the studies of interactions made more feasible.

Zhang et al. (2018) showed that the explained heritability increases with sample size in GWAS, and also estimated that the required sample size may range from a few hundred thousand to multiple millions to account for most of the heritability. The samples used in crop genetics, however, is often small, therefore, increasing sample size in crop GWAS has a great potential in future research.

With the rapid advances in various technologies, other types of omic data, including transcriptomic, proteomic, metabolomic and epigenetic data, have been recently exploited in crop research (Peng et al.; Wen et al., 2018b). These multi-omic variables may be treated as additional traits in GWAS, which promises to reduce knowledge gap between genotype and phenotype and will eventually benefit selective breeding. For example, omic-traits (at various layers) that are mapped to the same genomic locations with agronomic traits will provide multi-dimensional insights of genetic architectures and the underlying biological pathways. We believe multi-locus GWAS methodologies will become useful and popular tools for analysis of omics big datasets and help understand the mysterious world of genetics.

AUTHOR CONTRIBUTIONS

Y-MZ, ZJ, and JMD contributed to manuscript writing, provided important interpretations, and revised the work. All the authors checked and confirmed the final version of the manuscript.

FUNDING

The work was supported by the National Natural Science Foundation of China (31571268, 31871242, and U1602261), and Huazhong Agricultural University Scientific and Technological Self-innovation Foundation (2014RC020).

ACKNOWLEDGMENTS

We would like to thank Dr. Chengsong Zhu at Southwestern Medical Center, University of Texas and Dr. Jia Wen at Department of Bioinformatics and Genomics, University of North Carolina at Charlotte for their comments and suggestions.

REFERENCES

Barsh, G. S., Copenhaver, G. P., Gibson, G., and Williams, S. M. (2012). Guidelines for genome-wide association studies. *PLoS Genet.* 8:e1002812. doi: 10.1371/journal.pgen.1002812

Chang, M., He, L., and Cai, L. (2018). An overview of genome-wide association studies. *Methods Mol. Biol.* 1754, 97–108. doi: 10.1007/978-1-4939-7717-8_6

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., et al. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* 11, 446–450. doi: 10.1038/nrg2809

Fadista, J., Manning, A. K., Florez, J. C., and Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur J. Hum. Genet.* 24, 1202–1205. doi: 10.1038/ejhg.2015. 269

Feng, J. Y., Wen, Y. J., Zhang, J., and Zhang, Y. M. (2016). Advances on methodologies for genome-wide association studies in plants. *Acta Agron. Sin.* 42, 945–956. doi: 10.3724/SP.J.1006.2016.00945

He, L., Xiao, J., Rashid, K. Y., Jia, G., Li, P., Yao, Z., et al. (2019). Evaluation of genomic prediction for pasmo resistance in flax. *Int. J. Mol. Sci.* 20:E359. doi: 10.3390/ijms20020359

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Khan, M. A., Tong, F., Wang, W., He, J., Zhao, T., and Gai, J. (2018). Analysis of QTL-allele system conferring drought tolerance at seedling stage in a nested association mapping population of soybean [*Glycine max* (L.) Merr.] using a novel GWAS procedure. *Planta* 248, 947–962. doi: 10.1007/s00425-018- 2952-4

Li, C. X., Xu, W. G., Guo, R., Zhang, J. Z., Qi, X. L., Hu, L., et al. (2018). Molecular marker assisted breeding and genome composition analysis of Zhengmai 7698, an elite winter wheat cultivar. *Sci. Rep.* 8:322. doi: 10.1038/s41598-017-18726-8

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y. M., Todhunter, R. J., et al. (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol.* 12:73. doi: 10.1186/ s12915-014-0073-5

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12:e1005767. doi: 10.1371/journal.pgen.1005767

Maher, B. (2008). Personal genomes: the case of the missing heritability. *Nature* 456, 18–21. doi: 10.1038/456018a

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2018). pKWmEB: integration of Kruskal-Wallis test with empirical bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018a). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Wen, Y. J., Zhang, Y. W., Zhang, J., Feng, J. Y., Dunwell, J. M., and Zhang, Y. M. (2018b). An efficient multi-locus mixed model framework for the detection of small and linked QTLs in F_2. *Brief. Bioinform.* doi: 10.1093/bib/bby058. [Epub ahead of print].

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Lee, S. H., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47, 1114–1120. doi: 10.1038/ng.3390

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical bayes for multi-locus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y., Qi, G., Park, J. H., and Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* 50, 1318–1326. doi: 10.1038/s41588-018-0193-x

Zhang, Y. M., Mao, Y., Xie, C., Smith, H., Luo, L., and Xu, S. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics* 169, 2267–2275. doi: 10.1534/genetics.104.033217

Zhang, Y. M., and Tamba, C. L. (2018). A fast mrMLM algorithm for multi-locus genome-wide association studies. *bioRxiv [Preprint]*. doi: 10.1101/341784

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 26 April 2018
doi: 10.3389/fpls.2018.00561

Genetic Dissection of Maize Embryonic Callus Regenerative Capacity Using Multi-Locus Genome-Wide Association Studies

Langlang Ma[1†], *Min Liu*[1†], *Yuanyuan Yan*[1], *Chunyan Qing*[1], *Xiaoling Zhang*[1], *Yanling Zhang*[1], *Yun Long*[1], *Lei Wang*[1], *Lang Pan*[1], *Chaoying Zou*[1], *Zhaoling Li*[1], *Yanli Wang*[1], *Huanwei Peng*[2], *Guangtang Pan*[1], *Zhou Jiang*[1] and *Yaou Shen*[1]*

[1] Key Laboratory of Biology and Genetic Improvement of Maize in Southwest Region, Maize Research Institute, Sichuan Agricultural University, Chengdu, China, [2] Institute of Animal Nutrition, Sichuan Agricultural University, Chengdu, China

OPEN ACCESS

Edited by:
*Yuan-Ming Zhang,
Huazhong Agricultural University,
China*

Reviewed by:
*Wenqin Wang,
Shanghai Jiao Tong University, China
Zhiyong Zhang,
Rutgers University, The State
University of New Jersey,
United States
Haochuan Li,
Henan Agricultural University, China*

***Correspondence:**
*Yaou Shen
shenyaou@sicau.edu.cn*

[†] These authors have contributed equally to this work.

Specialty section:
*This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science*

Received: *24 February 2018*
Accepted: *10 April 2018*
Published: *26 April 2018*

Citation:
Ma L, Liu M, Yan Y, Qing C, Zhang X, Zhang Y, Long Y, Wang L, Pan L, Zou C, Li Z, Wang Y, Peng H, Pan G, Jiang Z and Shen Y (2018) Genetic Dissection of Maize Embryonic Callus Regenerative Capacity Using Multi-Locus Genome-Wide Association Studies. Front. Plant Sci. 9:561. doi: 10.3389/fpls.2018.00561

The regenerative capacity of the embryonic callus, a complex quantitative trait, is one of the main limiting factors for maize transformation. This trait was decomposed into five traits, namely, green callus rate (GCR), callus differentiating rate (CDR), callus plantlet number (CPN), callus rooting rate (CRR), and callus browning rate (CBR). To dissect the genetic foundation of maize transformation, in this study multi-locus genome-wide association studies (GWAS) for the five traits were performed in a population of 144 inbred lines genotyped with 43,427 SNPs. Using the phenotypic values in three environments and best linear unbiased prediction (BLUP) values, as a result, a total of 127, 56, 160, and 130 significant quantitative trait nucleotides (QTNs) were identified by mrMLM, FASTmrEMMA, ISIS EM-BLASSO, and pLARmEB, respectively. Of these QTNs, 63 QTNs were commonly detected, including 15 across multiple environments and 58 across multiple methods. Allele distribution analysis showed that the proportion of superior alleles for 36 QTNs was <50% in 31 elite inbred lines. Meanwhile, these superior alleles had obviously additive effect on the regenerative capacity. This indicates that the regenerative capacity-related traits can be improved by proper integration of the superior alleles using marker-assisted selection. Moreover, a total of 40 candidate genes were found based on these common QTNs. Some annotated genes were previously reported to relate with auxin transport, cell fate, seed germination, or embryo development, especially, GRMZM2G108933 (*WOX2*) was found to promote maize transgenic embryonic callus regeneration. These identified candidate genes will contribute to a further understanding of the genetic foundation of maize embryonic callus regeneration.

Keywords: maize, embryonic callus, regenerative capacity, multi-locus GWAS, candidate gene

INTRODUCTION

As one of the main crops for animals and humans, maize (*Zea mays* L.) is an important target for genetic manipulation (Zhang et al., 2014; Li et al., 2016). However, during maize transformation, difficulty in embryonic callus induction and regeneration, which occurs in most elite lines, presents a major bottleneck (Shen et al., 2012, 2013; Ge et al., 2016). Previous studies have suggested that both genotypes and exogenous hormones affect embryonic callus induction from maize immature embryos, such as abscisic acid (ABA), indole acetic acid (IAA), and gibberellic acid (GA3) widely considered to play important roles in callus formation (Jiménez and Bangerth, 2001; Ge et al., 2016). Genetic research has suggested that embryonic callus induction is controlled by nuclear genes in maize (Schlappi and Hohn, 1992). Furthermore, eight quantitative trait loci (QTL) and three epistatic interactions were found to control type I callus formation in a maize recombinant inbred line (RIL) population (Krakowsky et al., 2006). In previous studies, some transcription factors and microRNAs in hormone signal transduction pathways were found to regulate the process of embryonic callus induction (Shen et al., 2013; Ge et al., 2016). To date, research exploring callus regenerative capacity has mainly focused on *Arabidops*is, rice, wheat, maize, and other plants. In *Arabidopsis*, *PLT* genes (PLETHORA) were proved to modulate the regenerative capacity by a two-step mechanism (Kareem et al., 2015). First, *PLT3, PLT5,* and *PLT7* activated the expression of root stem cell regulators *PLT1* and *PLT2* to establish pluripotency and form shoot progenitors. Then, *PLT3*, *PLT5,* and *PLT7* up-regulated the expression of shoot-promoting factor Cup-shaped cotyledon1 (*CUC1*) and Cup-shaped cotyledon2 (*CUC2*) to complete the shoot regeneration process. Inhibitor of cyclin-dependent kinase (*ICK*), a cyclin-dependent kinase (*CDK*) inhibitor, has been shown to enhance the regenerative capacity of *Arabidopsis* embryonic callus (Cheng et al., 2015). Moreover, WUSCHEL-related homeobox 5 (*WOX5*) expression in the quiescent center (QC) is considered as a marker of the root stem cell niche in *Arabidopsis* (Sarkar et al., 2007). In addition, as an AP2/ERF transcription factor, wound induced dedifferentiation1 (WIND1) promoted the *Arabidopsis* shoot regeneration by up-regulating the expression of enhancer of shoot regeenration1 *(ESR1)* gene which encoded another AP2/ERF transcription factor (Iwase et al., 2015, 2017). For wheat, genes controlling green shoot re-differentiation were mapped to chromosomal sites 3A, 5B, 2D, and 1B (Szakács et al., 1988). Additionally, QTL mapping showed that two QTLs on chromosomes 1 and 9 control green shoot re-differentiation in rice, with the latter considered to be a major locus (Ping et al., 1998). Nishimura et al. (2005) observed a main QTL encoding ferredoxin-nitrite reductase (NiR) which is responsible for regenerate ability in rice. Recently, WUSCHEL-related homeobox 2 (*WOX2*) and Baby Boom (*BBM*) genes were introduced into maize by genetic transformation, which resulted in the increased rate of resistant seedlings from transformed immature embryos (Lowe et al., 2016). So far, the genetic basis of plant regeneration has not been well understood especially for maize, in which few functional genes have been revealed to directly control embryonic callus regeneration. Therefore, more systematic studies are required to reveal the genetic basis of maize embryonic callus regenerative capacity.

Genome-wide association analysis (GWAS) is a useful tool in the dissection of complex traits (Abdel-Ghani et al., 2013; Pace et al., 2015). Using mixed linear model (MLM) and general linear model (GLM), 4 and 263 significant SNPs were found to be associated with root architecture traits at maize seedling stage, respectively. More specifically, GRMZM2G153722, which is located on chromosome 4, was found to contain nine significant SNPs that are likely expressed in the roots and shoots (Pace et al., 2015). When using GWAS, several genes that modulate maize leaf architecture were identified in a nested association mapping (NAM) population (Tian et al., 2011). GWAS also aided in the identification of 74 candidate genes associated with maize oil biosynthesis (Li et al., 2013). Furthermore, another study identified a total of 51 SNPs significantly associated with maize leaf blight by adopting a NAM population, with most of the candidate genes reported in previous studies as relating to plant disease resistance (Kump et al., 2011). To our knowledge, there is no study that has utilized GWAS when detecting the embryonic callus regenerative capacity until now.

In this study, four multi-locus GWAS approaches were used to dissect the genetic foundations for the five regenerative capacity-related traits in a natural population containing rich genetic information across multiple environments. Our objectives were: (i) to understand the significance of genotype, environment, and genotype × environment on traits relating to regenerative capacity; (ii) to identify significant quantitative trait nucleotides (QTNs) and candidate genes that modulate the five traits and resolve the genetic basis of maize embryonic callus regenerative capacity; and (iii) to analyze and compare the detection powers of different methods and identify the optimal multi-locus GWAS approach. To our knowledge, this is the first comprehensive study aimed at understanding the genetic basis of maize embryonic callus regenerative capacity using multi-locus GWAS approaches.

MATERIALS AND METHODS

Plant Materials and Phenotypic Data Analysis

In a previous study, we examined the embryonic callus induction rate in immature embryos from a natural maize population of 362 inbred lines, with 144 of the lines exhibiting efficient induction (**Table S1**) and thus they were used to detect regenerative capacity. The details of planting and culturing processes were described by Zhang et al. (2017b). Herein, five regeneration ability-related traits, namely, embryonic green callus rate (GCR), callus differentiating rate (CDR), callus plantlet number (CPN), callus rooting rate (CRR), and callus browning rate (CBR), were examined (the features of the five traits were shown in **Figure 1**). The data were transformed as previously described with the GCR, CDR, CRR, and CBR values calculated by $\sin^{-1}\sqrt{p}$ and the CPN value calculated by $\sqrt{p+1}$, with p being the initial value (Zhang et al., 2017b). The analysis of variance (ANOVA), phenotypic correlation, BLUP values and broad-sense heritability (H_B^2) were all completed in our previous study (Zhang et al., 2017b).

Genotypic Data Analysis

Genomic DNA was extracted from mixed leaf tissues from eight plants per line using the CTAB method (Zhang et al., 2016). All of the accessions were genotyped using the Illumina MaizeSNP50 BeadChip containing 56,110 SNPs (http://support.illumina.com/array/array_kits/maizesnp50_dna_analysis_kit/downloads.html). A total of 43,427 SNPs across 10 chromosomes remained after quality filtering (**Figure S1**), with SNPs having a missing rate >20%, heterozygosity >20%, and minor allele frequency (MAF) <0.05 deleted. These 43,427 SNPs were subsequently used for calculating the population structure and kinship and to perform GWAS.

Population Structure, Linkage Disequilibrium, and Multi-Locus Association Studies

STRUCTURE 2.3.4 was used to estimate subgroup numbers within the population structure (Q matrix) (Evanno et al., 2005). Among the 43,427 SNPs, 5,000 high quality SNPs with a rare allele frequency (RAF) >20% were randomly selected for the estimating panel. Based on the subgrouping results, the obtained evaluated data were used for further analysis.

TASSEL 4.0 was utilized to analyze linkage disequilibrium (LD) (Bradbury et al., 2007), with the LD decay calculated by plotting r^2 onto the genetic distance in base pairs with a cutoff of $r^2 = 0.2$. The LD decay was calculated using only markers that remained after quality filtering. Additionally, the Loiselle kinship coefficients between inbred lines in a panel (K matrix) were calculated using SpAGeDi software (Hardy and Vekemans, 2002).

In this study, four multi-locus GWAS approaches were used to detect significant QTNs for five embryonic callus regenerative capacity-related traits (mrMLM v2.1, https://cran.r-project.org/web/packages/mrMLM/index.html), including mrMLM (Wang et al., 2016), FASTmrEMMA (Wen et al., 2017), ISIS EM-BLASSO (Tamba et al., 2017), and pLARmEB (Zhang et al., 2017a). Owing to the fact that these multi-locus methods were more powerful and accurate than the single-locus MLM methods in their simulation experiments, thus we adopted these multi-locus methods in this study. Moreover, Q- and K-matrices were applied to correct the population structure and Loiselle kinship coefficients that were calculated between inbred lines. The setting parameters for these methods were as follows: (i) mrMLM, critical *P*-value of 0.01 in rMLM and critical LOD score of 3.0 in mrMLM (Wang et al., 2016); (ii) FASTmrEMMA, critical *P*-value of 0.005 in first step of FASTmrEMMA and critical LOD score of 3.0 in the last step of FASTmrEMMA (Wen et al., 2017); (iii) ISIS EM-BLASSO, critical *P*-value of 0.0002 in ISIS EM-BLASSO (Tamba et al., 2017); and (iv) pLARmEB, critical LOD score of 3.0 in pLARmEB and the number of potentially associated variables for each chromosome: 143 ("144–1") (Zhang et al., 2017a).

Superior Allele Analysis and Annotation of Candidate Genes

For QTNs (RefGen_v2) that were detected consistently in multiple environments or methods, a superior genotype was determined based on the effect value of each significant QTN. For each QTN, the superior allele percentage in these elite inbred lines was equal to number of lines containing superior alleles divided by the total line number. For each line, the proportion of superior alleles in these QTNs was calculated as superior allele number divided by total QTN number. A heat map visualizing the percentage of superior alleles was obtained in the R (heatmap package) program (Mellbye and Schuster, 2014).

Herein, the QTNs which locate in gene regions were used to identify the candidate genes. Furthermore, the corresponding candidate genes of the consistent QTNs that were stably expressed in multi-environment or multi-method were annotated by performing a GENE search on the NCBI website (RefGen_v2) (https://www.ncbi.nlm.nih.gov/).

Real-Time PCR for Candidate Genes

Four candidate genes GRMZM2G108933 (*WOX*2), GRMZM2G066749, GRMZM2G163761, and GRMZM2G371033 were randomly selected for identification of expression levels at different regeneration stages (0 d, 3 d, 6 d, and 9 d) by quantitative real-time PCR analysis (qPCR, ABI 7500 real-time PCR System, Torrance, CA, USA). Firstly, RNA samples were extracted using TRIZOL reagent (Invitrogen, Beijing, China) and RNase-free DNase (Takara, Beijing, China). Then, cDNA was obtained by PrimeScript RT Reagent Kit With gDNA Eraser (TaKaRa, Beijing, China). Moreover, the primers were designed using the software Primer Premier 5.0. The detailed PCR amplification programmes were described as Shen et al. (2012), and the $2^{-\Delta\Delta Ct}$ method was used for calculating the expression levels (Schefe et al., 2006). Here, *Actin* 1 (GRMZM2G126010) was used as the reference gene.

RESULTS

Phenotype for Regenerative Capacity-Related Traits

The phenotypes for CBR, CDR, CPN, CRR, and GCR have been described by Zhang et al. (2017b), readers are encouraged to refer to the original study (Zhang et al., 2017b). The results were briefly described here. The average values for the above five traits across three environments were 37.70, 17.30, 1.28, 11.50, and 43.16 with the standard deviations 26.99, 17.52, 0.51, 14.25 and 24.94, respectively. Additionally, the heritability (h_B^2) of the five traits ranged from 47.09 to 78.91%, suggesting that genetic effects play an important role in the formation of these traits. A significantly positive correlation was observed between CDR and CPN, while a significantly negative correlation was found between CBR and GCR ($P = 0.01$). The high correlation coefficient between the BLUP value and the phenotypic value in a single environment (>0.9) indicated the reliability of the phenotypic values for most of the traits (**Figure S2**; Zhang et al., 2017b).

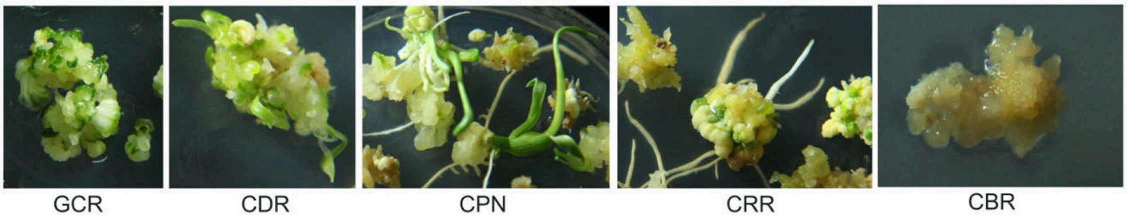


FIGURE 1 | Features of the five traits. The traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate).

Linkage Disequilibrium Decay in the Population

To obtain the average distance of LD decay, 43,427 SNPs were adopted. As shown in **Figure S3**, r^2 decreased gradually with increased distance. However, the r^2-value reached a plateau when it decreased to a certain level. The corresponding distance was considered as the average distance of LD decay in this population. Herein, the average LD decay distance was 220 kb ($r^2 = 0.2$), which is consistent with a previous study (Zhang et al., 2016). Moreover, the distance was greater than the average distance between markers of 48 kb, thus indicating sufficient coverage.

Population Structure

A subset of 5,000 high quality SNPs were randomly chosen to define the subpopulations within the panel of 144 lines. Delta K (ΔK) was calculated using STRUCTURE 2.3.4 (**Figure 2A**; $K =$ 2–9), with two subpopulations (selected $K = 2$) presented based on ΔK-values (**Figure 2B**). These two subgroups contained 109 (75.69%) and 35 (24.31%) lines (**Table S1**), respectively. The larger subpopulation included tropical, temperate, and mixed germplasms, while the other was composed of mostly temperate lines (**Table S1**).

QTNs Detected by Multi-Locus GWAS Methods

Four multi-locus GWAS approaches were utilized in this study. A total of 127, 56, 160, and 130 significant QTNs were identified in mrMLM, FASTmrEMMA, ISIS EM-BLASSO, and pLARmEB, respectively, for five traits across three environments and the BLUP model (**Figures 3**, **4**, **Figure S4**, **Tables S2–S5**). Among them, 26, 29, 27, 16, and 29 QTNs were identified for CBR, CDR, CPN, CRR, and GCR, respectively, in multi-location and BLUP model by mrMLM method (**Figure 3A**; **Table S2**). When using FASTmrEMMA, the number of QTNs detected for the five traits were 14, 13, 7, 11, and 11, respectively (**Figure 3B**; **Table S3**). The ISIS EM-BLASSO method also identified 29, 37, 26, 31, and 37 QTNs for the above five traits (**Figure 3C**; **Table S4**). Moreover, 29, 28, 25, 27, and 21 QTNs were identified for the above five traits, respectively, using the pLARmEB approach (**Figure 3D**; **Table S5**).

We further analyzed the common QTNs that were co-identified in at least two of the environments (or environments and BLUP model) using a certain multi-locus GWAS approach. A total of 15 common QTNs were identified by combination of these four methods (**Table 1**). Among them, six, two, eight, and three environment-stable QTNs were identified using mrMLM, FASTmrEMMA, ISIS EM-BLASSO, and pLARmEB method, respectively (**Table 1**). These common QTNs were separately located on chromosomes 1, 2, 3, 4, 5, 7, 8, and 9, with LOD values ranging from 3.06 to 9.40 (**Table 1**). The proportion of phenotypic variance explained (PVE) by each QTN ranged from 1.83 to 19.26% (**Table 1**). Furthermore, three, four, two, six, and four common QTNs were found significantly associated with CBR, CDR, CPN, CRR, and GCR, respectively (**Table 1**).

When comparing the results across different methods, 58 QTNs were consistently identified by two or more methods (**Table 2**), and they were associated with the CBR (15), CDR (13), CPN (11), CRR (9), and GCR (16) traits (**Figure 4**; **Table 2**). Especially, three QTNs (SYNGENTA15901, SYN39155, and SYN32084) were found to be significantly associated with CBR, CDR, and CDR, respectively, in all the multi-locus methods (**Table 2**). Meanwhile, the average LOD-values and PVE ranges of the three QTNs for the CBR (5.69; 4.77–13.19%), CDR (4.68; 3.10–6.45%), and CDR (7.61; 2.65–7.25%) traits were also generated (**Table 2**).

Remarkably, 10 QTNs were co-detected not only in multi-environment (including environment and the BLUP model) but also by different methods (**Table 3**). Among these QTNs, the three QTNs (SYNGENTA15901, SYN39155, and SYN32084) were detected by all the methods as well as in BLUP model and CZ (**Table 3**). Furthermore, two other QTNs (SYN8267 and PZE-107005556) that are associated with CBR and CRR were identified by three methods and in two environments. The remaining six QTNs were associated with CPN, CRR, and GCR, and they were identified by two methods and found in two locations (**Table 3**).

Distribution of Superior Alleles in Elite Inbred Lines

The 63 common QTNs, detected in multiple environments or using multiple methods, were considered as important QTNs associated with regenerative capacity-related traits. Since 31 elite inbred lines were included in the constructed panel, this enabled us to evaluate the utilization of superior alleles during maize breeding. Herein, the allele associated with a higher phenotypic value was defined as the superior allele for each of the traits, except for CBR and CRR, because callus browning and callus rooting are both disadvantageous phenotypes for regeneration.

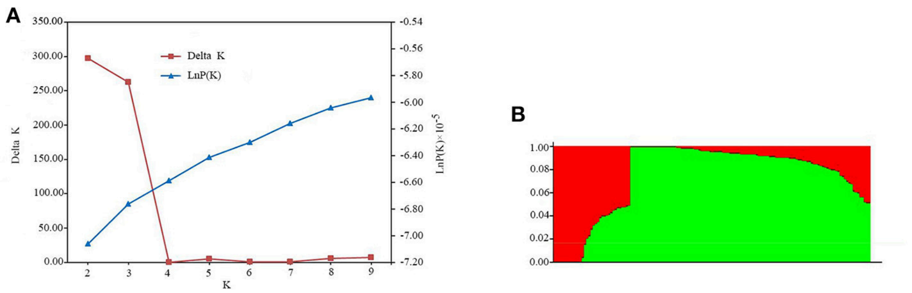


FIGURE 2 | Population structure estimates based on 43,427 SNPs distributed across 10 chromosomes. **(A)** Plot of ln*P*(*D*), with Δ*K* calculated for *K* = 2–9. **(B)** Population structure estimates (*K* = 2), the areas of the two colors (green and red) illustrate the proportion of each subgroup.

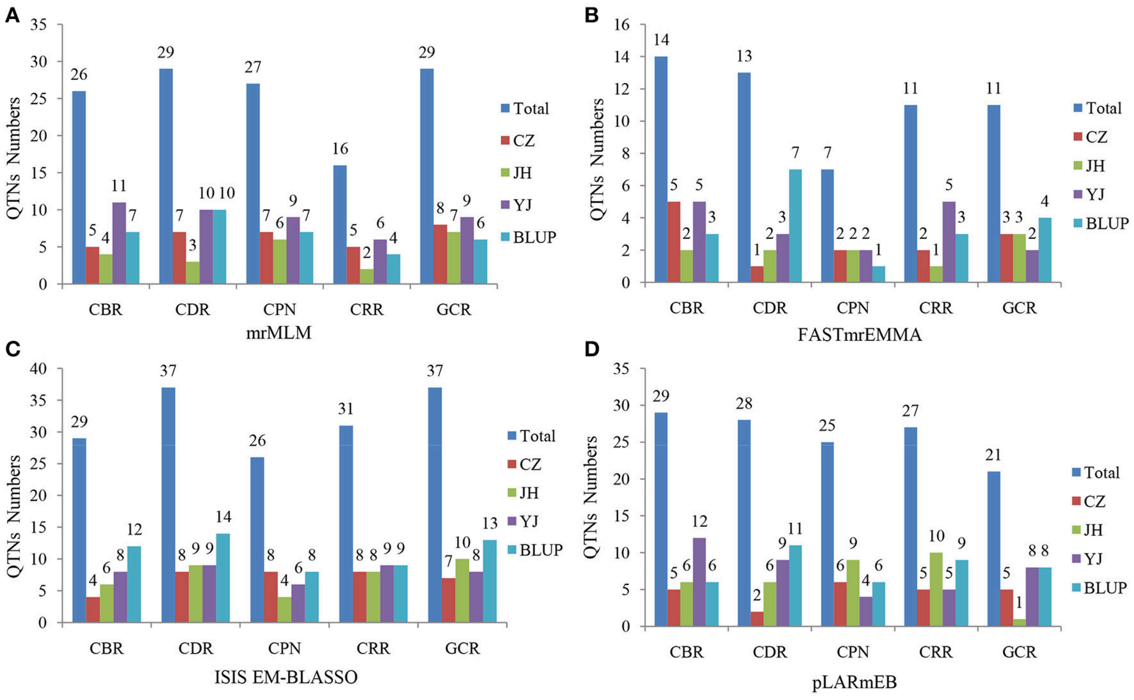


FIGURE 3 | Number of detected QTNs for the five traits across three environments and BLUP model in four methods. The traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate). CZ, JH, and YJ denote the population planted in Chongzhou (2015), Jinghong (2014), and Yuanjiang (2015), respectively. The approaches utilized included **(A)** mrMLM, **(B)** FASTmrEMMA, **(C)** ISIS EM-BLASSO, and **(D)** pLARmEB.

As described in **Table 4**, the superior allele percentages for the QTNs ranged from 0.00 to 96.67% in the elite lines, with 27 of the QTNs containing ≥50% superior alleles while the remaining 36 QTNs contained <50% (**Figure 5**; **Table 4**). Three QTNs (PZE-101213720, PZE-103108199, and PZE-108021239) had superior allele percentages >80%, while eight (PZE-104066682, PZE-103049772, PZE-101220149, PZE-107024505, PZE-102109640, PZE-109067144, PZE-109121058, and PZE-109066380) had percentages <10% (**Figure 5**; **Table 4**).

Moreover, 18 of the elite lines that contained 26–40 superior alleles showed higher phenotypic values, with increased percentages of 109.81% (CDR), 32.91% (CPN), and 75.63% (GCR), relative to the other 13 elite lines that contained 10–25 superior alleles (**Table 5** and **Table S6**). However, for CBR and CRR, the 18 elite lines that contained between 26 and 40 superior alleles had the averaged phenotypic values of 34.17 and 9.68, respectively, which were 26.08 and 28.55% lower than the other 13 elite lines that contained 10–25 superior alleles (**Table 5** and **Table S6**). These findings suggest that the superior alleles have obviously additive effects on regenerative capacity. Therefore, the maize callus regenerative capacity can be improved by increasing the numbers of superior alleles in the lines with low regenerative capacities by marker assisted selection (MAS). Among them, CDR and GCR are the most attractive traits for MAS modification due to them having the most significant enhancement effect. In addition, we found some lines with high regenerative capacity shared common superior alleles, such as lines 178, 18-599, and DH40 which all contained the superior alleles of SYN15872, PZE-104024889, PZE-103108199, PZE108057325, PZE-101096007, PZE-101223466, PZE-108021239, PUT-163a-31909945-2005, PZE-106043314, PZE-102109721, SYN11739, SYN8144, SYN18315, PZE-102186765, SYN31996, PZE-104025174, and PZE-110088629 (**Figure 5**). This suggested these superior alleles may play an important role in callus regeneration process. All these findings will be more useful in the application of superior alleles in maize breeding.

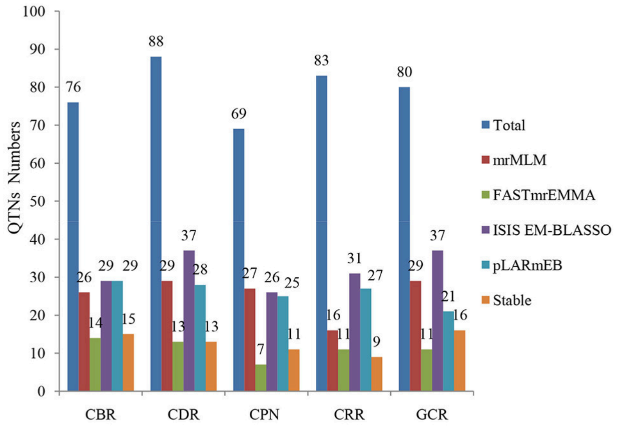


FIGURE 4 | Comparison of the number of detected QTNs from the four methods. The four methods are mrMLM, FASTmrEMMA, ISIS EM-BLASSO, and pLARmEB. The traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate). Total: denotes the total QTN number for each trait and Stable: denotes the number of stably expressed QTNs across multiple methods for each trait.

Candidate Genes Determined Based on Common QTNs

According to the 63 common QTNs, we further focused on the associated candidates. The results showed that a total of 40 candidate genes were obtained based on the B73 genome (RefGen_v2, **Table 6**). Among them, 8, 11, 7, 8, and 13 candidate genes were associated with CBR, CDR, CPN, CRR, and GCR, respectively (**Table 6**). Moreover, one QTN correlated with the CDR trait was associated with GRMZM2G589579 and had the largest LOD-value of 16.25 (**Tables 2**, **6**). Based on the functional annotations, these genes were mainly classified as transcription factors and kinases (**Table 6**). Specifically, seven genes were located on chromosomes 1, 2, 3, and 6, with each associated with two of the regeneration capacity-related traits (**Table 6**). In detail, gene models GRMZM2G108933, GRMZM2G072264,

TABLE 1 | Stably expressed QTNs for the five traits in each method across three environments and BLUP model.

Method	Trait	Environment and BLUP	Marker	Chromosome	Marker position (bp)	QTN effect	LOD score	r^2 (%)[a]
mrMLM	CBR	YJ, BLUP	PZE-109066380	9	109,317,272	−16.82, −8.67	5.61, 3.23	13.92, 15.08
	CDR	CZ, BLUP	SYN15872	8	161,523,427	−6.45, −2.16	5.18, 4.75	8.44, 5.45
	CPN	CZ, BLUP	PZE-101220149	1	271,749,865	−0.28, −0.11	8.26, 4.42	14.86, 8.27
		CZ, BLUP	SYN39155	3	2,446,145	0.12, 0.06	3.85, 3.06	5.86, 5.14
	CRR	CZ, BLUP	PZE-101160089	1	202,300,686	−14.92, −3.37	8.18, 8.28	16.38, 19.26
	GCR	YJ, BLUP	SYNGENTA13688	2	5,681,488	−7.20, −6.40	4.26, 4.08	7.55, 9.82
FASTmrEMMA	CBR	CZ, BLUP	SYNGENTA15901	7	5,038,808	21.95, 5.60	7.63, 3.74	13.19, 4.77
	GCR	CZ, BLUP	PZE-104024889	4	28,985,737	−14.27, −7.42	3.31, 3.66	6.48, 5.99
ISIS EM-BLASSO	CBR	CZ, BLUP	SYN8267	4	169,213,008	−10.11, −1.80	5.36, 3.14	10.05, 1.83
	CDR	CZ, BLUP	SYN32084	1	256,515,262	5.76, 2.51	6.35, 8.87	7.25, 6.09
		CZ, BLUP	SYN39155	3	2,446,145	4.47, 1.69	4.76, 4.59	4.68, 3.10
		CZ, JH	PZE-107024505	7	26,451,809	−5.72, −4.80	3.54, 4.09	5.14, 4.39
	CRR	CZ, BLUP	PZE-101160089	1	202,300,686	−8.84, −1.97	3.26, 5.38	6.23, 6.74
		CZ, BLUP	SYN35026	5	1,946,471	−3.60, −0.83	3.28, 4.72	3.07, 4.09
		YJ, BLUP	PZE-107005556	7	3,824,391	−4.20, −0.75	6.29, 3.90	7.74, 3.76
	GCR	CZ, JH	PZE-102138070	2	186,820,524	8.49, 8.27	6.33, 9.02	10.12, 12.43
pLARmEB	CRR	CZ, BLUP	SYN35026	5	1,946,471	−8.14, −1.35	7.37, 9.40	17.27, 6.53
		JH, BLUP	PZE-107005556	7	3,824,391	−4.03, −0.70	4.47, 3.43	4.80, 2.00
	GCR	YJ, BLUP	PZE-104068814	4	136,958,099	−4.17, −3.01	3.18, 3.82	2.69, 4.74

Traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate).
JH, CZ, and YJ denote the population planted in Jinghong (2014), Chongzhou (2015), and Yuanjiang (2015), respectively.
[a] *r^2 (%), phenotypic variation of traits explained by each QTN.*

TABLE 2 | Stably expressed QTNs for the five traits among different multi-methods.

Trait	Method (1, 2, 3, 4)[a]	Marker	Chromosome	Maker position (bp)	LOD score	r^2 (%)[b]
CBR	1, 3, 4	PZE-102138070	2	186,820,524	5.68, 3.92, 6.73	9.48, 5.00, 11.56
	1, 4	PZE-108009888	8	10,298,512	5.78, 3.17	8.01, 4.60
	1, 3	PZE-110088629	10	138,832,523	3.85, 3.35	5.43, 2.65
	1, 3	PZE-103123331	3	181,066,730	3.29, 4.53	6.40, 3.58
	1, 2, 3	SYN6514	3	196,351,287	3.56, 4.05, 7.36	3.63, 6.32, 8.30
	1, 3, 4	SYN7221	2	6,200,684	3.75, 11.83, 5.94	7.76, 10.78, 6.22
	1, 2	PZE-109067144	9	110,459,334	3.27, 3.14	2.59, 4.92
	1, 2, 3, 4	SYNGENTA15901	7	5,038,808	3.69, 3.74 (7.63),7.03, 5.84	10.75,4.77 (13.19) 12.35, 11.39
	1, 3	PZE-102109721	2	141,363,433	4.18, 5.46	3.64, 4.91
	2, 3	PZE-102151093	2	197,600,202	5.56, 6.37	10.24, 12.66
	2, 3, 4	PZE-101213720	1	264,163,677	4.65, 6.84, 4.20	6.88, 11.24, 6.73
	2, 3, 4	SYN8267	4	169,213,008	5.19, 3.14 (5.35),7.29	8.43, 1.83 (10.05), 16.35
	2, 4	PZE-101152052	1	19,548,4495	4.16, 6.94	6.30, 4.76
	2, 4	PZE-108020924	8	19,855,121	3.75, 13.53	4.89, 9.52
	3, 4	PZE-104067972	4	134,998,323	4.90, 9.53	4.89, 7.06
CDR	1, 4	PZE-106032634	6	75,630,749	16.15, 4.43	18.87, 5.68
	1, 2, 3, 4	SYN39155	3	2,446,145	4.73, 5.07, 4.59 (4.76), 4.05	4.85, 6.45, 3.10 (4.68), 4.21
	1, 2, 3, 4	SYN32084	1	256,515,262	4.04, 3.05, 8.87 (6.35), 8.74	5.89, 2.65, 6.09 (7.25), 5.19
	1, 2, 3	PZE-101216827	1	267,908,158	3.48, 4.33, 6.19	4.77, 4.23, 3.25
	1, 2	SYN11739	9	9,965,031	4.92, 3.66	14.28, 8.27
	1, 3	SYN31996	6	163,506,361	3.13, 3.86	9.16, 3.58
	1, 3	PZE-108002411	8	2,512,300	16.25, 4.97	14.64, 6.03
	1, 4	PZE-101096007	1	94,367,481	4.86, 4.45	3.72, 5.66
	1, 3	PZE-102109640	2	141,173,773	7.67, 5.68	4.86, 5.53
	3, 4	PZE-104025174	4	29,335,471	9.52, 3.01	7.33, 0.89
	3, 4	PZE-106036875	6	84,672,851	4.31, 6.72	1.97, 4.40
	3, 4	PUT-163a-31909945-2005	6	110,706,817	4.29, 3.58	2.04, 4.30
	3, 4	SYN8144	10	142,358,869	4.21, 4.09	3.31, 5.42
CPN	1, 3, 4	PZE-106032634	6	75,630,749	7.88, 4.58, 4.56	14.66, 7.12, 0.62
	1, 3	PZE-101220149	1	271,749,865	4.42 (8.26), 4.96	8.27 (14.86), 6.35
	1, 3	SYN39155	3	2,446,145	3.06 (3.85), 4.20	5.14 (5.86), 3.04
	1, 3, 4	PZE-108105282	8	159,954,599	3.88, 6.55, 5.42	6.91, 5.57, 1.57
	1, 4	PZE-108057325	8	102,454,042	3.07, 6.36	5.32, 0.47
	1, 3	PZE-104066682	4	131,771,972	7.49, 3.31	17.57, 8.73
	1, 3	PZE-106043314	6	93,212,668	6.64, 3.46	9.61, 4.20
	1, 2, 3	PZE-102186765	2	230,884,488	9.11, 7.26, 6.50	12.45, 0.30, 9.41
	2, 3	PZE-109062403	9	104,884,301	6.36, 13.00	16.33, 14.58
	3, 4	PZE-103049772	3	54,177,469	6.93, 12.72	14.03, 12.98
	1, 3	SYN29447	5	213,294,101	3.29, 6.18	5.79, 6.17
CRR	2, 3	SYN18315	1	252,377,691	7.13, 7.27	12.37, 9.45
	2, 3	PZE-106008760	6	25,291,385	6.76, 4.44	13.33, 6.04
	2, 4	PZE-101085779	1	75,627,286	4.81, 7.65	10.52, 6.47
	2, 3	SYN28088	5	68,653,887	4.18, 6.13	6.74, 5.74
	2, 3	PZE-106000504	6	1,234,387	3.04, 4.17	4.13, 2.57

(Continued)

TABLE 2 | Continued

Trait	Method (1, 2, 3, 4)[a]	Marker	Chromosome	Maker position (bp)	LOD score	r^2 (%)[b]
	2, 3, 4	PZE-107005556	7	3,824,391	4.13, 3.90 (6.29) 3.43(4.47)	9.14, 3.76 (7.74) 2.00(4.80)
	3, 4	SYN35026	5	1,946,471	4.72 (3.28) 9.40 (7.37)	4.09 (3.07) 6.53 (17.27)
	3, 4	PZE-105122012	5	179,270,149	5.07, 7.25	4.17, 4.70
	3, 4	SYN18708	1	21,466,619	5.18, 4.33	4.15, 3.11
GCR	1, 3	SYN32084	1	256,515,262	3.99, 4.08	6.10, 4.35
	1, 3	SYNGETA13688	2	5,681,488	4.08 (4.26), 4.96	9.82 (7.55), 5.25
	1, 3, 4	PZE-109121058	9	154,807,596	8.07, 7.71, 5.76	13.73, 11.26, 8.20
	1, 3, 4	PZE-103123331	3	181,066,730	3.42, 4.03, 7.25	6.67, 7.75, 5.46
	1, 3	PZE-108010908	8	11,504,308	4.61, 6.26	6.88, 6.58
	1, 4	SYN37974	2	10,782,867	3.25, 6.41	7.78, 10.77
	2, 3	PZE-103108199	3	169,053,554	3.28, 5.74	4.47, 6.96
	2, 3	PZE-104024889	4	28,985,737	3.66 (3.31), 4.93	5.99 (6.48), 5.90
	2, 4	PZE-104069507	4	138,153,696	5.14, 4.45	10.42, 14.05
	2, 4	PZE-101106628	1	110,914,630	4.86, 5.53	15.03, 9.59
	3, 4	SYN7221	2	6,200,684	4.70, 3.42	3.57, 3.82
	3, 4	PZE-109081358	9	129,514,761	5.86, 6.96	3.98, 9.62
	3, 4	PZE-101223466	1	274,722,612	9.02, 7.28	12.32, 8.94
	3, 4	PZE-102138070	2	186,820,524	6.33 (9.02), 6.42	10.12 (12.43), 9.73
	3, 4	SYN21743	9	1,347,687	7.10, 5.60	9.66, 7.20
	3, 4	PZE-108021239	8	20,231,393	4.19, 4.09	4.87, 5.42

Traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate).
[a] *Method numbers correspond to (1) mrMLM, (2) FASTmrEMMA, (3) ISIS EM-BLASSO, and (4) pLARmEB.*
[b] *r^2 (%), phenotypic variation of traits explained by each QTN.*
The values in parentheses denote the means for QTNs in different environments.

TABLE 3 | Stably expressed QTNs in both multi-environment (including BLUP model) and multi-method.

Trait	Marker	Method (1, 2, 3, 4)[a]	Environment and BLUP	LOD score
CBR	SYNGENTA15901	1, 2, 3, 4	BLUP and CZ (2)	3.74 and 7.63
	SYN8267	2, 3, 4	BLUP and CZ (3)	3.14 and 5.35
CDR	SYN39155	1, 2, 3, 4	BLUP and CZ (3)	4.59 and 4.76
	SYN32084	1, 2, 3, 4	BLUP and CZ (3)	8.87 and 6.35
CPN	PZE-101220149	1, 3	BLUP and CZ (1)	4.42 and 8.26
	SYN39155	1, 3	BLUP and CZ (1)	3.06 and 3.85
CRR	PZE-107005556	2, 3, 4	BLUP and YJ (3, 4)	3.90 and 6.29, 3.43 and 4.47
	SYN35026	3, 4	BLUP and YJ (3, 4)	4.72 and 3.28, 9.40 and 7.37
GCR	SYNGETA13688	1, 3	BLUP and YJ (1)	4.08 and 4.26
	PZE-104024889	2, 3	BLUP and CZ (2)	3.66 and 3.31
	PZE-102138070	3, 4	CZ and YJ (3)	6.33 and 9.02

Traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate).
[a] *Method numbers correspond to (1) mrMLM, (2) FASTmrEMMA, (3) ISIS EM-BLASSO, and (4) pLARmEB.*
JH, CZ, and YJ denote the population planted in Jinghong (2014), Chongzhou (2015), and Yuanjiang (2015), respectively.

and GRMZM2G026095 were individually correlated with both CBR and GCR, while GRMZM2G309660 was associated with CBR and CRR (**Table 6**). Moreover, GRMZM2G163761 was correlated with CDR and GCR, while GRMZM2G097959 and GRMZM5G835276 were associated with CDR and CPN (**Table 6**).

TABLE 4 | Distribution of superior alleles in stably expressed QTNs among 31elite inbred lines.

QTN	Superior alleles	Percentage (%)[a]	QTN	Superior alleles	Percentage (%)[a]	QTN	Superior alleles	Percentage (%)[a]
PZE-101213720	GG	96.67	SYN15872	AA	54.84	PZE-108002411	AA	32.26
PZE-103108199	TT	83.87	PUT-163a-31909945-2005	GG	53.33	PZE-102186765	CC	32.26
PZE-108021239	GG	80.65	PZE-106043314	GG	53.33	PZE-106036875	CC	32.26
SYN18315	CC	76.67	SYN18708	TT	51.61	PZE-101216827	CC	26.67
SYN31996	CC	75.86	PZE-108010908	CC	50.00	PZE-108105282	CC	25.81
SYN35026	AA	74.19	PZE-104025174	CC	50.00	PZE-109062403	AA	25.81
SYN32084	AA	73.33	PZE-109081358	CC	48.39	SYN29447	AA	24.14
PZE-102109721	GG	73.33	SYN28088	CC	46.67	SYN8267	AA	24.14
SYN8144	CC	70.00	SYN21743	TT	43.33	PZE-108009888	GG	19.23
PZE-106008760	GG	67.74	SYN37974	CC	42.86	PZE-108020924	CC	16.67
PZE-107005556	AA	67.74	PZE-104067972	CC	40.00	PZE-106032634	GG	13.33
PZE-110088629	CC	65.52	PZE-102151093	GG	40.00	PZE-101160089	TT	12.90
SYN39155	GG	64.52	PZE-101152052	AA	38.71	SYNGENTA13688	AA	12.90
PZE-101096007	GG	63.33	PZE-106000504	AA	38.71	PZE-104066682	GG	9.68
PZE-103123331	AA	63.33	SYNGENTA15901	CC	38.71	PZE-103049772	GG	9.68
SYN11739	GG	61.29	PZE-104068814	AA	38.71	PZE-101220149	GG	9.68
PZE-101223466	GG	60.00	PZE-108057325	TT	35.71	PZE-107024505	TT	7.14
PZE-104024889	AA	60.00	PZE-101085779	CC	35.71	PZE-102109640	AA	6.45
PZE-101106628	TT	58.62	SYN6514	GG	33.33	PZE-109067144	GG	6.45
PZE-102138070	TT	58.62	PZE-105122012	CC	33.33	PZE-109121058	CC	3.23
PZE-104069507	GG	56.67	SYN7221	AA	32.26	PZE-109066380	TT	0.00

[a] *Percentage (%) was calculated as: (superior allele number within the 31 elite inbred lines/total allele number for the 31 elite inbred lines) × 100%.*

Expression Patterns of Candidate Genes

To detect the responses of the candidate genes to callus regeneration, two lines 141 (with high regenerative capacity) and ZYDH381-1 (with low regenerative capacity) were submitted to qRT-PCR analysis for four randomly selected genes at three regenerative stages (3 d, 6 d, and 9 d) and CK (0 d). Among these genes, *WOX2* was up-regulated at all of the stages compared to 0 d in 141 and ZYDH381-1. However, the expression level of *WOX2* in line 141 was higher than that in ZYDH381-1 at each of the stages. Besides, the expression peak occurred at 3 d in 141, which was at 6 d in ZYDH381. These indicate that *WOX*2 was more susceptive in the response of callus regeneration in 141 (**Figure 6A**). GRMZM2G066749 was down-regulated at every of regenerative stage when compared with 0 d in 141, whereas it was slightly up-regulated in ZYDH381-1. Interestingly, the expression level of GRMZM2G066749 in 141 was much higher than that in ZYDH381-1 at all of the stages including 0 d (**Figure 6B**). However, the expression levels of GRMZM2G163761 and GRMZM2G371033 were generally higher in ZYDH381-1 than those in 141 (**Figures 6C,D**). These findings suggested that the difference of expression patterns in different lines could be an important factor which influenced the regenerative capacity of embryonic callus.

DISCUSSION

Population Selection

A population of 144 inbred lines was used for the present study, which is slightly smaller than in other maize GWAS studies (Pace et al., 2014, 2015; Zhang et al., 2016). This is due to the specialty of these maize callus regenerative ability-related traits, which are based on the embryonic callus induction. In our previous study, 362 inbred lines were utilized to identify embryonic callus induction, and only 144 lines had a relatively efficient induction. Therefore, the present study is based on a comparatively small maize population. Moreover, population structure analysis showed that this novel population was divided into two subpopulations. The average LD decay distance was 220 kb ($r^2 = 0.2$), which was relatively consistent with the distance obtained for the initial population of 362 inbred lines (Zhang et al., 2016). This finding indicates that the average LD decay distance is almost stable despite the reduced population size. Additionally, some QTNs for the five traits were co-identified in different methods and multi-environment (in Results section). Of particular interest is the candidate gene *WOX2* (in Candidate Gene Functions in Callus Regenerative Capacity section), which has been proven to promote the formation of resistant seedlings after callus transformation. These findings confirm the reasonability of the population structure used for this study.

Advantages of the New Multi-Locus GWAS Approaches

Previous studies have dissected some complex traits using a GLM or MLM based on a single-locus GWAS (Zhang et al., 2005; Yu et al., 2006; Pace et al., 2015). However, both of these two models have procedural limitations. GLM has a high false positive rate (FPR) because this model does not correct the population

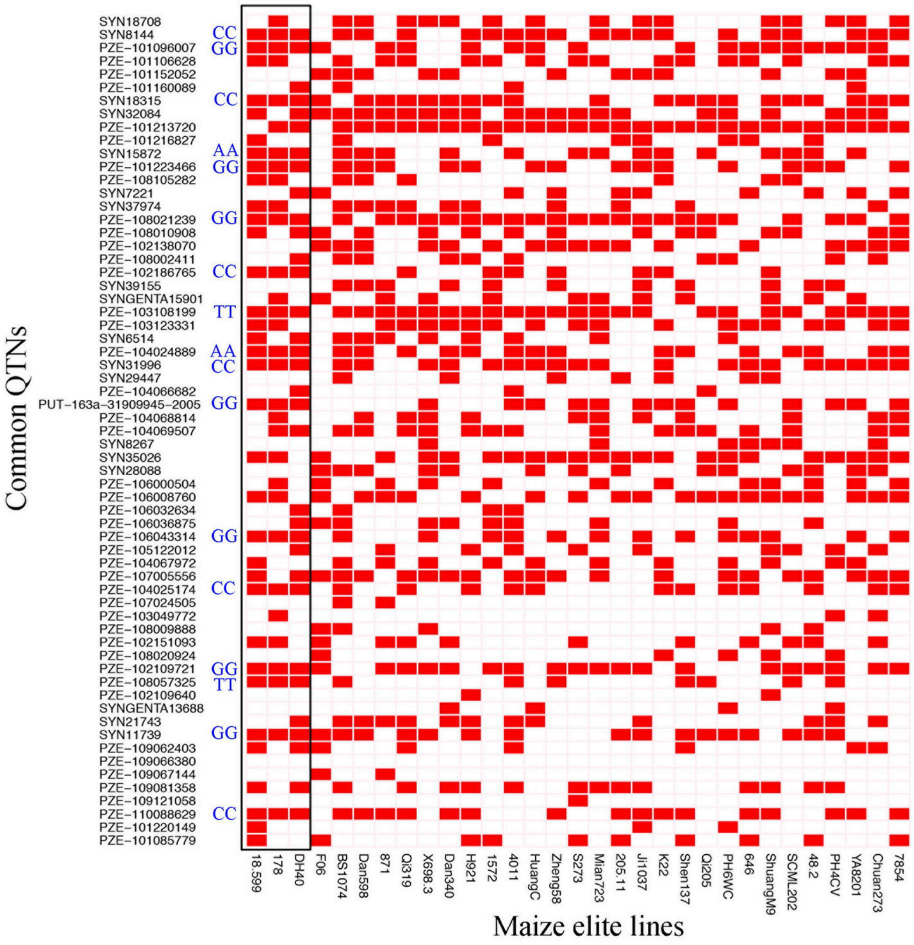


FIGURE 5 | Heat map of the superior alleles distribution for the 63 QTNs in 31 elite inbred lines**.** Red and White colors represent superior and inferior alleles, respectively. Black box means the superior alleles distribution for the 63 QTNs in high-regeneration lines 18-599, 178, and DH40. AA, TT, GG, and CC represent the genotypes for common superior alleles in 18-599, 178, and DH40.

structure (Q) or polygenic background (K; Korte and Farlow, 2013). In the MLM, the correction of Q and K is so stringent that many significant loci are missed, especially small-effect loci (Wang et al., 2016). In recent years, researchers developed some multi-locus methodologies to address these limitations, such as mrMLM, FASTmrEMMA, ISIS EM-BLASSO, and pLARmEB (Wang et al., 2016; Tamba et al., 2017; Wen et al., 2017; Zhang et al., 2017a), and they were used in this study. In these new multi-locus methods, the significance level was set to a LOD score = 3, which was equivalent to $P = 0.0002$ (Wang et al., 2016). However, in the single-locus MLM GWAS methods, the significance threshold is generally set to $P = 0.05/m$ (m is the number of markers), thus the multi-locus GWAS methods are less stringent. Furthermore, FPRs for these four multi-locus GWAS approaches were smaller than in the single-locus MLM GWAS methods and other multi-locus GWAS methods (Wang et al., 2016; Tamba et al., 2017; Wen et al., 2017; Zhang et al., 2017a). Therefore, these methods were considered effective alternative approaches (Wang et al., 2016; Tamba et al., 2017; Wen et al., 2017; Zhang et al., 2017a). In this study, 127, 56, 160, and 130 significant QTNs were found for the five traits using mrMLM, FASTmrEMMA, ISIS EM-BLASSO, and pLARmEB, respectively (**Figure 3**; **Tables S2–S5**). However, only one and six significantly associated SNPs were detected when using MLM (R package GAPIT) and FarmCPU (R package FarmCPU; PCA+K, where PCA and K were calculated by GAPIT and SpAGeDi, respectively) models, respectively ($P = 0.05/43427 = 1.15 \times 10^{-6}$; **Table S7**). This suggested that these multi-locus methods were more powerful when used for detecting the QTNs for regeneration-related traits of maize. Furthermore, some stably expressed QTNs were commonly detected in multiple environments (or between environment and the BLUP model) (**Table 1**) and a total of 58 common QTNs were identified by multiple methods (**Table 2**). These evidence verified the reliability of these new multi-locus methods. Comparison of the four methods illustrates that ISIS EM-BLASSO is slightly more powerful than the other three methods (**Figure 3**). Additionally, the running time for these four methods when using the data generated herein are as follows: mrMLM >

TABLE 5 | Phenotypic values of different numbers of superior alleles for the five traits among the common QTNs within the 31 elite lines.

Trait	Mean phenotypic value in three environments	Mean phenotypic value in three environments	Increased percentage (%)
	(contain 10–25 superior alleles)	**(contain 26–40 superior alleles)**	
CBR	46.22	34.17	−26.08
CDR	11.30	23.70	109.81
CPN	1.09	1.45	32.91
CRR	13.54	9.68	−28.55
GCR	27.14	47.66	75.63

Traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate).

FASTmrEMMA > pLARmEB > ISIS EM-BLASSO (**Figure S5**). Notably, the validated functional gene *WOX2* (mentioned above) was commonly detected in multiple methods for both CBR and GCR. These findings suggest that the most robust approach enabling the identification of the most interesting candidate genes is to use a combination of the methods utilized herein.

Application of Superior Alleles in Maize Breeding

When examining the common QTNs within the 31 elite inbred lines, 36 of the 63 QTNs contained <50% superior alleles (**Table 4**), suggested that these alleles were not effectively selected during artificial selection. A possible reason is that maize regenerative ability was not previously a main breeding focus. Instead, breeding efforts have focused on yield-related traits, plant type-related traits, resistance-related traits, and high quality-related traits. In the remaining 27 common QTNs, superior alleles proportions ≥50% was observed, with three of these QTNs (PZE-101213720, PZE-103108199, and PZE-108021239) having proportions >80% (**Table 4**). These findings suggest that in some cases, these superior alleles must be linked with traits of interest for breeders and thus were maintained during artificial selection.

The results presented herein show that the identified superior alleles exhibited additive effects on the regenerative capacity. Furthermore, this study focused on the number of superior alleles in several popular inbred lines (Zheng 58, PH4CV, and PH6WC), whose high yields and high combining abilities were outstanding (Barker, 2005; Ma et al., 2014; Li et al., 2017). The results showed that for each line, the superior allele proportion was <50% in the 63 QTNs (**Figure 5**, **Table S6**). Future studies could focus on these lines acquiring more super alleles and an improved regenerative ability that will contribute to the establishment of callus regeneration and a transformation system. These findings also enable the furthering of gene functional research in these lines.

We further investigated the distribution of these superior alleles in those lines that failed to induce the callus. As a result, the proportions of the superior alleles in different lines ranged from 20.64 to 57.14% and the average value was 40.27%, which was very similar to the averaged proportion (40.96%) of superior alleles in the 144 inducible lines. In addition, the averaged proportion of superior alleles for *WOX2* (PZE-103123331) was 69.29 and 70.14%, respectively, in the uninducible and inducible lines (these data were not provided). These suggested that these QTNs associated with callus regeneration were probably not involved in the induction of embryonic callus.

Candidate Genes Involved in Callus Regenerative Capacity

Based on the identified common QTNs, 40 candidate genes were identified, with several previously reported to be associated with transgenic callus regeneration, auxin transport, cell fate, seed germination, or embryo development (**Table 6**). These gene included GRMZM2G108933, GRMZM2G130442, GRMZM2G315375, GRMZM2G163761, GRMZM2G412611, GRMZM2G066749, and GRMZM2G371033. GRMZM2G108933, which was associated with CBR and CDR, was annotated to *WOX2*, an embryonic transcription factor (Nardmann et al., 2007) (**Table 6**). In *Arabidopsis*, *WOX5* is closely associated with the root stem cell niche (Sarkar et al., 2007). In the recent year, *WOX2* (a homologous gene to GRMZM2G108933) was introduced into maize by genetic transformation, and it increased the rate of resistant seedlings from transformed immature embryos (Lowe et al., 2016). These findings suggest that GRMZM2G108933 could be an important functional gene controlling maize callus regeneration by inhibiting callus browning and promoting callus differentiation. GRMZM2G130442 (associated with GCR) and GRMZM2G315375 (associated with CRR) are thought to regulate plant embryo development, which is consistent with their assigned associations herein (**Table 6**). As a member of the HD-Zip (homeo domain-leucine zipper) family, GRMZM2G130442 was annotated to the *Zea mays* outer cell layer (*ZmOCL*) family (**Table 6**), which has been reported to play roles in defining different regions of the epidermis during embryonic development and it controls the maintenance of cell-layer identity in meristematic regions (Ingram et al., 2000). GRMZM2G315375, known as *br2*, encodes P-glycoproteins (PGPs) (**Table 6**), which has been implicated in auxin transport. Meanwhile, auxin is widely accepted to be one of the most important hormones for embryo dedifferentiation (Pasternak et al., 2002). Moreover, Cassani et al. (2011) proposed that the interaction between *br 2* and *br 3* results in an alteration in embryo development. Regeneration is a process involving callus re-differentiation and it is similar to embryo development, but the opposite of embryo dedifferentiation (Yang et al., 2012). Therefore, these findings suggest that GRMZM2G130442 and GRMZM2G315375 could be modulators of callus regeneration.

Gene model GRMZM2G163761 was annotated to KIP1 (knotted interacting protein1) and was associated with CDR and GCR (**Table 6**). Smith et al. (2002) reported that cell fate in the shoot apical meristem is influenced by the transcriptional regulation from the association of KIP and KN1 (knotted 1), a three amino acid loop extension (TALE) class of

TABLE 6 | Candidate genes based on the stable commonly expressed QTNs.

Trait	Marker	Chromosome	Position (bp)	Candidate genes (v2)	Annotation
CBR GCR	PZE-103123331	3	181,066,730	GRMZM2G108933	(*WOX2*) WUSCHEL related homeobox 2
CBR GCR	SYN7221	2	6,200,684	GRMZM2G072264	RNA-binding (RRM/RBD/RNP motifs) family protein
GCR CBR	PZE-102138070	2	186,820,524	GRMZM2G026095	Tuliposide A-converting enzyme 1, chloroplastic
CBR CRR	SYN6514	3	196,351,287	GRMZM2G309660	Unknown
CDR GCR	SYN32084	1	256,515,262	GRMZM2G163761	kip1 (knotted interacting protein1)
CPN CDR	SYN39155	3	2,446,145	GRMZM2G097959	GTP binding
CDR CPN	PZE-106032634	6	75,630,749	GRMZM5G835276	Alpha-L-fucosidase 2
CBR	SYNGENTA15901	7	5,038,808	GRMZM2G060866	Anther-specific proline-rich protein APG
CBR	PZE-101213720	1	264,163,677	GRMZM2G123204	Adenylosuccinate synthetase, chloroplastic
CBR	PZE-101152052	1	195,484,495	GRMZM2G138425	Hypothetical protein
CBR	SYN8267	4	169,213,008	GRMZM2G383210	jmj21—JUMONJI-transcription factor 21
CDR	SYN15872	8	161,523,427	GRMZM2G371033	sbp18 (SBP-transcription factor 18)
CDR	PZE-101216827	1	267,908,158	GRMZM2G066749	dek35 (defective kernel35)
CDR	SYN31996	6	163,506,361	GRMZM2G136219	Unknown
CDR	PZE-108002411	8	2,512,300	GRMZM2G589579	ago4a (argonaute4a)
CDR	PZE-101096007	1	94,367,481	GRMZM2G088524	mybr32 (MYB-related-transcription factor 32)
CDR	PZE-106036875	6	84,672,851	GRMZM2G088930	Midasin
CDR	PUT-163a-31909945-2005	6	110,706,817	GRMZM2G412611	Alpha-glucan water dikinase 1 chloroplastic
CDR	SYN8144	10	142,358,869	GRMZM2G033724	Trypsin family protein
CPN	PZE-108105282	8	159,954,599	GRMZM2G460576	Unknown
CPN	PZE-104066682	4	131,771,972	GRMZM2G122983	Vacuolar protein sorting-associated protein 20 homolog 2
CPN	PZE-106043314	6	93,212,668	GRMZM2G047969	Protein JASON
CPN	PZE-102186765	2	230,884,488	GRMZM2G082302	Unknown
CPN	SYN2944	5	213,294,101	GRMZM2G017868	Unknown
CRR	PZE-101160089	1	202,300,686	GRMZM2G315375	br2 (brachytic2)
CRR	SYN35026	5	1,946,471	GRMZM2G415491	rh3 (RNA helicase3)
CRR	SYN18315	1	252,377,691	GRMZM2G165042	bhlh43 (bHLH-transcription factor 43)
CRR	PZE-106008760	6	25,291,385	GRMZM2G168441	Putative HLH DNA-binding domain superfamily protein
CRR	SYN28088	5	68,653,887	GRMZM2G168603	MDIS1-interacting receptor like kinase 1
CRR	PZE-106000504	6	1,234,387	GRMZM2G137894	Pentatricopeptide repeat-containing protein At2g33760
CRR	SYN18708	1	21,466,619	GRMZM2G004397	pco148373a Syntaxin/t-SNARE family protein
GCR	PZE-104024889	4	28,985,737	GRMZM2G130442	ocl5a (outer cell layer5a)
GCR	PZE-104068814	4	136,958,099	GRMZM2G034697	Phosphatidyl-N-methylethanolamine N-methyltransferase
GCR	PZE-108010908	8	11,504,308	GRMZM2G112968	Unknown
GCR	SYN37974	2	10,782,867	GRMZM2G068982	Methionine aminopeptidase
GCR	PZE-103108199	3	169,053,554	GRMZM2G028252	Hypothetical protein
GCR	PZE-104069507	4	138,153,696	GRMZM2G133226	Nucleotide/sugar transporter family protein
GCR	PZE-101106628	1	110,914,630	GRMZM2G368632	Cysteine-rich receptor-like protein kinase 10
GCR	PZE-101223466	1	274,722,612	GRMZM2G001869	Unknown
GCR	PZE-108021239	8	20,231,393	GRMZM2G168933	Hypothetical protein

The traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate).

homeodomain. Another candidate gene, GRMZM2G412611, which was correlated with CDR was annotated as an alpha-glucan water dikinase 1, chloroplastic-like (**Table 6**). In wheat, the suppression of glucan water dikinase in the endosperm altered the wheat grain properties, germination, and coleoptile growth (Bowerman et al., 2016). The CDR-associated gene, GRMZM2G066749, was annotated to *dek 35* (defective kernel 35) (**Table 6**). Clark and Sheridan (1988) demonstrated that *dek 35* is pleiotropic when affecting endosperm, gametophyte, or embryo development by using two non-allelic defective-kernel mutants of maize. These findings indicate that the above genes probably control the callus regenerative capacity by affecting cell fate determination or development of somatic embryo.

CONCLUSIONS

In this study, four new multi-locus GWAS methods were used to identify traits related to regenerative capacity. A total of

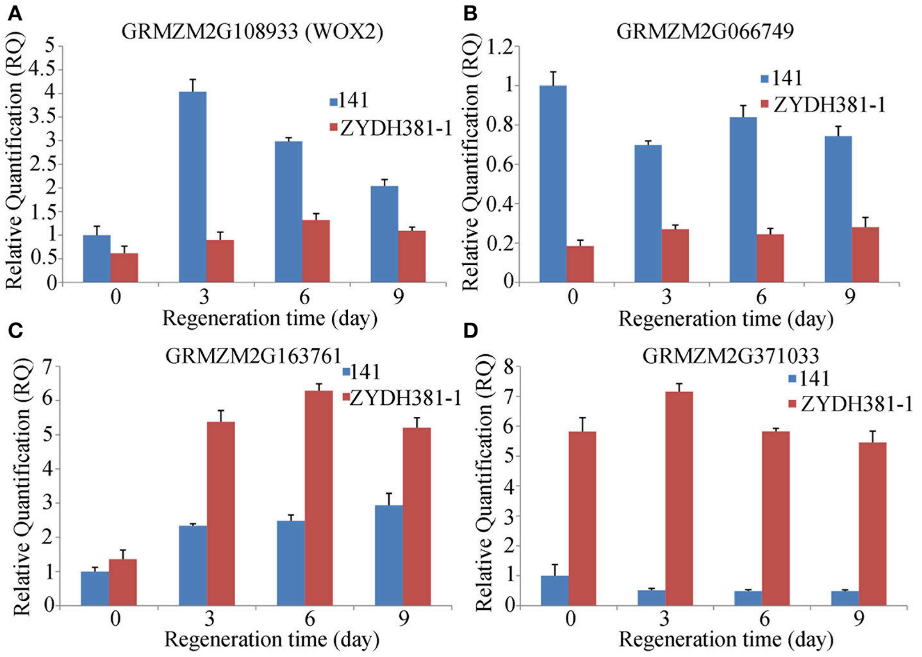


FIGURE 6 | Expression levels of candidate genes at different regeneration stages. Here, 141 is a line with high regeneration and ZYDH381-1 IS the one with low regeneration. **(A–D)** Represents the relative expression levels of GRMZM2G108933 (WOX2), GRMZM2G066749, GRMZM2G163761 and GRMZM2G371033, respectively.

127, 56, 160, and 130 significant QTNs, respectively, were identified in mrMLM, FASTmrEMMA, ISIS EM-BLASSO, and pLARmEB for five traits across three environments and the BLUP model. Among these QTNs, 63 were commonly detected in multiple environments or using multiple methods. In total, 40 candidate genes were obtained based on the common QTNs, with several previously reported to correlate with transgenic callus regeneration, auxin transport, or embryo development. For the common QTNs, the percentages of superior alleles ranged from 0.00 to 96.67% within the 31 elite inbred lines. Further analysis revealed that these superior alleles exhibit an additive effect on the regenerative capability of the related traits. These findings suggest that an improvement of the maize callus regenerative ability can be achieved by integrating more superior alleles into maize lines by MAS.

AUTHOR CONTRIBUTIONS

YS and GP: Designed the experiments; LM, ML, XZ, YY, CQ, YZ, YL, LW, LP, CZ, ZL, YW, and HP: Conducted the experiments and performed the analysis. LM, ML, and YS: Drafted the manuscript. All of the authors provided final approval of this manuscript.

FUNDING

This work was supported by the Young Scientists Fund of Sichuan Province (2016JQ0008), the National Basic Research Program of China (2014CB138200), and the Key Fund of the Department of Education of Sichuan Province (15ZA0004).

ACKNOWLEDGMENTS

We would like to thank the Maize Research Institute of Sichuan Agricultural University for providing the platform. We also extend thanks to other graduate students who attended this project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00561/full#supplementary-material

Figure S1 | Distribution of 43,427 SNPs on 10 chromosomes.

Figure S2 | Predictions of the five traits in three environments by BLUP (Zhang et al., 2017b). The traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate). CZ, JH, and YJ denote the population planted in Chongzhou (2015), Jinghong (2014), and Yuanjiang (2015), respectively. R^2 denotes the correlation coefficient between the phenotype value mean and corresponding BLUP-value.

Figure S3 | Linkage disequilibrium decay in the mapping population. A cutoff of $r^2 = 0.2$ was utilized.

Figure S4 | Manhattan plot of multi-locus GWAS for the five traits. The plots show all of the significant QTNs (LOD score > 3) across three environments and the BLUP model for the four methods (mrMLM, FASTmrEMMA, ISIS EM-BLASSO, and pLARmEB). CZ, JH, and YJ denote the population planted in Chongzhou (2015), Jinghong (2014), and Yuanjiang (2015), respectively. Panels **(A–E)** denotes the significant QTNs for CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate), respectively.

Figure S5 | Running times for the four methods using 43,427 SNPs.

Table S1 | Inbred lines and population Q matrix.

Table S2 | Significant QTNs for the five traits across three environments and the BLUP model using the mrMLM method. The traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate). JH, CZ, and YJ denote the population planted in Jinghong (2014), Chongzhou (2015), and Yuanjiang (2015), respectively. r^2 (%), phenotypic variation of traits explained by each QTN.

Table S3 | Significant QTNs for the five traits across three environments and the BLUP model using the FASTmrEMMA method. The traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate). JH, CZ, and YJ denote the population planted in Jinghong (2014), Chongzhou (2015), and Yuanjiang (2015), respectively. r^2 (%), phenotypic variation of traits explained by each QTN.

Table S4 | Significant QTNs for the five traits across three environments and the BLUP model using the ISIS EM-BLASSO method. The traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate). JH, CZ, and YJ denote the population planted in Jinghong (2014), Chongzhou (2015), and Yuanjiang (2015), respectively. r^2 (%), phenotypic variation of traits explained by each QTN.

Table S5 | Significant QTNs for the five traits across three environments and the BLUP model using the pLARmEB method. The traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate). JH, CZ, and YJ denote the population planted in Jinghong (2014), Chongzhou (2015), and Yuanjiang (2015), respectively. r^2 (%), phenotypic variation of traits explained by each QTN.

Table S6 | Phenotypic value and superior allele numbers for the 31 elite lines for each trait. The traits include CBR (callus browning rate), CDR (callus differentiating rate), CPN (callus plantlet number), CRR (callus rooting rate), and GCR (green callus rate).

Table S7 | SNPs significantly associated with traits detected by GWAS using statistical models MLM and FarmCPU. The traits include CPN (callus plantlet number) and GCR (green callus rate).

REFERENCES

Abdel-Ghani, A. H., Kumar, B., Reyesmatamoros, J., Gonzalezportilla, P. J., Jansen, C., Jps, M., et al. (2013). Genotypic variation and relationships between seedling and adult plant traits in maize (*Zea mays* L.) inbred lines grown under contrasting nitrogen levels. *Euphytica* 189, 123–133. doi: 10.1007/s10681-012-0759-0

Barker, T. C. (2005). *Inbred Maize Line PH4CV: U.S. Patent No. 6,897,363*. Washington, DC: U.S. patent and Tademark Office.

Bowerman, A. F., Newberry, M., Dielen, A. S., Whan, A., Larroque, O., Pritchard, J., et al. (2016). Suppression of glucan, water dikinase in the endosperm alters wheat grain properties, germination and coleoptile growth. *Plant Biotechnol. J.* 14, 398–408. doi: 10.1111/pbi.12394

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. doi: 10.1093/bioinformatics/btm308

Cassani, E., Villa, D., Durante, M., Landoni, M., and Pilu, R. (2011). The brachytic 2 and 3 maize double mutant shows alterations in plant growth and embryo development. *Plant Growth Regul.* 64, 185–192. doi: 10.1007/s10725-010-9556-8

Cheng, Y., Liu, H., Cao, L., Wang, S., Li, Y., Zhang, Y., et al. (2015). Down-regulation of multiple CDK inhibitor ICK/KRP genes promotes cell proliferation, callus induction and plant regeneration in Arabidopsis. *Front. Plant Sci.* 6:825. doi: 10.3389/fpls.2015.00825

Clark, J. K., and Sheridan, W. F. (1988). Characterization of the two maize embryo-lethal defective kernel mutants rgh*-1210 and fl*-1253b: effects on embryo and gametophyte development. *Genetics* 120, 279–290.

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Ge, F., Luo, X., Huang, X., Zhang, Y., He, X., Liu, M., et al. (2016). Genome-wide analysis of transcription factors involved in maize embryonic callus formation. *Physiol. Plant.* 158, 452–462. doi: 10.1111/ppl.12470

Hardy, O. J., and Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Resour.* 2, 618–620. doi: 10.1046/j.1471-8286.2002.00305.x

Ingram, G. C., Boisnard-Lorig, C., Dumas, C., and Rogowsky, P. M. (2000). Expression patterns of genes encoding HD-ZipIV homeo domain proteins define specific domains in maize embryos and meristems. *Plant J.* 22, 401–414. doi: 10.1046/j.1365-313X.2000.00755.x

Iwase, A., Harashima, H., Ikeuchi, M., Rymen, B., Ohnuma, M., Komaki, S., et al. (2017). WIND1 promotes shoot regeneration through transcriptional activation of enhancer of shoot regeneration1 in Arabidopsis. *Plant Cell* 29, 54–69. doi: 10.1105/tpc.16.00623

Iwase, A., Mita, K., Nonaka, S., Ikeuchi, M., Koizuka, C., Ohnuma, M., et al. (2015). WIND1-based acquisition of regeneration competency in Arabidopsis and rapeseed. *J. Plant Res.* 128, 389–397. doi: 10.1007/s10265-015-0714-y

Jiménez, V. M., and Bangerth, F. (2001). Hormonal status of maize initial explants and of the embryogenic and non-embryogenic callus cultures derived from them as related to morphogenesis *in vitro*. *Plant Sci.* 160, 247–257. doi: 10.1016/S0168-9452(00)00382-4

Kareem, A., Durgaprasad, K., Sugimoto, K., Du, Y., Pulianmackal, A. J., Trivedi, Z. B., et al. (2015). PLETHORA genes control regeneration by a two-step mechanism. *Curr. Biol.* 25, 1017–1030. doi: 10.1016/j.cub.2015.02.022

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. doi: 10.1186/1746-4811-9-29

Krakowsky, M., Lee, M., Garay, L., Woodman-Clikeman, W., Long, M., Sharopova, N., et al. (2006). Quantitative trait loci for callus initiation and totipotency in maize (*Zea mays* L.). *Theor. Appl. Genet.* 113, 821–830. doi: 10.1007/s00122-006-0334-y

Kump, K. L., Bradbury, P. J., Wisser, R. J., Buckler, E. S., Belcher, A. R., Oropezarosas, M. A., et al. (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* 43, 163–168. doi: 10.1038/ng.747

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45, 43–50. doi: 10.1038/ng.2484

Li, K., Wang, H., Hu, X., Ma, F., and Wu,Y., Wang, Q., et al. (2017). Genetic and quantitative trait locus analysis of cell wall components and forage digestibility in the Zheng 58× HD568 maize RIL population at anthesis stage. *Front. Plant Sci.* 8:1472. doi: 10.3389/fpls.2017.01472

Li, X., Zhou, Z., Ding, J., Wu, Y., Zhou, B., Wang, R., et al. (2016). Combined linkage and association mapping reveals QTL and candidate genes for plant and ear height in maize. *Front. Plant Sci.* 7:833. doi: 10.3389/fpls.2016.00833

Lowe, K., Wu, E., Wang, N., Hoerster, G., Hastings, C., Cho, M.-J., et al. (2016). Morphogenic regulators Baby boom and Wuschel improve monocot transformation. *Plant Cell* 28, 1998–2015. doi: 10.1105/tpc.16.00124

Ma, D. L., Xie, R. Z., Niu, X. K., Li, S. K., Long, H. L., and Liu, Y. E. (2014). Changes in the morphological traits of maize genotypes in China between the 1950s and 2000s. *Eur. J. Agron.* 58, 1–10. doi: 10.1016/j.eja.2014.04.001

Mellbye, B., and Schuster, M. (2014). Physiological framework for the regulation of quorum sensing-dependent public goods in *Pseudomonas aeruginosa*. *J. Bacteriol. Mycol.* 196, 1155–1164. doi: 10.1128/JB.01223-13

Nardmann, J., Zimmermann, R., Durantini, D., Kranz, E., and Werr, W. (2007). WOX gene phylogeny in Poaceae: a comparative approach addressing leaf and embryo development. *Mol. Biol. Evol.* 24, 2474–2484. doi: 10.1093/molbev/msm182

Nishimura, A., Ashikari, M., Lin, S., Takashi, T., Angeles, E. R., Yamamoto, T., et al. (2005). Isolation of a rice regeneration quantitative trait loci gene and its application to transformation systems. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11940–11944. doi: 10.1073/pnas.0504220102

Pace, J., Gardner, C., Romay, C., Ganapathysubramanian, B., and Lübberstedt, T. (2015). Genome-wide association analysis of seedling root development in maize (*Zea mays* L.). *BMC Genomics* 16:47. doi: 10.1186/s12864-015-1226-9

Pace, J., Lee, N., Naik, H. S., Ganapathysubramainian, B., and Lübberstedt, T. (2014). Analysis of maize (*Zea mays*) seedling roots with the high-throughput image analysis tool ARIA (Automatic Root Image Analysis). *PLoS ONE* 9:e108255. doi: 10.1371/journal.pone.0108255

Pasternak, T. P., Prinsen, E., Ayaydin, F., Miskolczi, P., Potters, G., Asard, H., et al. (2002). The role of auxin, pH, and stress in the activation of embryogenic cell division in leaf protoplast-derived cells of alfalfa. *Plant Physiol.* 129, 1807–1819. doi: 10.1104/pp.000810

Ping, H., Lishuang, S., Chaofu, L., Ying, C., and Lihuang, Z. (1998). Genetic analysis and mapping the anther calture response genes in rice (*Oryza sativa* L.). *Acta Genet. Sin.* 25, 337–344.

Sarkar, A. K., Luijten, M., Miyashima, S., Lenhard, M., Hashimoto, T., Nakajima, K., et al. (2007). Conserved factors regulate signalling in *Arabidopsis thaliana* shoot and root stem cell organizers. *Nature* 446, 811–814. doi: 10.1038/nature05703

Schefe, J. H., Lehmann, K. E., Buschmann, I. R., Unger, T., and Funke-Kaiser, H. (2006). Quantitative real-time RT-PCR data analysis: current concepts and the novel "gene expression's C_T difference" formula. *J. Mol. Med.* 84, 901–910. doi: 10.1007/s00109-006-0097-6

Schlappi, M., and Hohn, B. (1992). Competence of immature maize embryos for Agrobacterium-mediated gene transfer. *Plant Cell* 4, 7–16. doi: 10.1105/tpc.4.1.7

Shen, Y., Jiang, Z., Lu, S., Lin, H., Gao, S., Peng, H., et al. (2013). Combined small RNA and degradome sequencing reveals microRNA regulation during immature maize embryo dedifferentiation. *Biochem. Biophys. Res. Commun.* 441, 425–430. doi: 10.1016/j.bbrc.2013.10.113

Shen, Y., Jiang, Z., Yao, X., Zhang, Z., Lin, H., Zhao, M., et al. (2012). Genome expression profile analysis of the immature maize embryo during dedifferentiation. *PLoS ONE* 7:e32237. doi: 10.1371/journal.pone.0032237

Smith, H. M., Boschke, I., and Hake, S. (2002). Selective interaction of plant homeodomain proteins mediates high DNA-binding affinity. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9579–9584. doi: 10.1073/pnas.092271599

Szakács, E., Kovács, G., Pauk, J., and Barnabás, B. (1988). Substitution analysis of callus induction and plant regeneration from anther culture in wheat (*Triticum aestivum* L.). *Plant Cell Rep.* 7, 127–129. doi: 10.1007/BF00270121

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flintgarcia, S., et al. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* 43, 159–162. doi: 10.1038/ng.746

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform. bbw* 18:908. doi: 10.1093/bib/bbx028

Yang, X., Zhang, X., Yuan, D., Jin, F., Zhang, Y., and Xu, J. (2012). Transcript profiling reveals complex auxin signalling pathway and transcription regulation involved in dedifferentiation and redifferentiation during somatic embryogenesis in cotton. *BMC Plant Biol.* 12:110. doi: 10.1186/1471-2229-12-110

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017a). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity (Edinb).* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Q., Zhang, L., Evers, J., Werf, W. V. D., Zhang, W., and Duan, L. (2014). Maize yield and quality in response to plant density and application of a novel plant growth regulator. *Field Crops Res.* 164, 82–89. doi: 10.1016/j.fcr.2014.06.006

Zhang, X., Long, Y., Ge, F., Guan, Z., Zhang, X., Wang, Y., et al. (2017b). A genetic study of the regeneration capacity of embryonic callus from the maize immature embryo culture. *Hereditas* 39, 143–155. doi: 10.16288/j.yczz.16-323

Zhang, X., Zhang, H., Li, L., Lan, H., Ren, Z., Liu, D., et al. (2016). Characterizing the population structure and genetic diversity of maize breeding germplasm in Southwest China using genome-wide SNP markers. *BMC Genomics* 17:697. doi: 10.1186/s12864-016-3041-3

Zhang, Y. M., Mao, Y., Xie, C., Smith, H., Luo, L., and Xu, S. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics* 169, 2267–2275. doi: 10.1534/genetics.104.033217

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 07 May 2018
doi: 10.3389/fpls.2018.00611

Multi-Locus Genome-Wide Association Study Reveals the Genetic Architecture of Stalk Lodging Resistance-Related Traits in Maize

*Yanling Zhang[1], Peng Liu[1], Xiaoxiang Zhang[1], Qi Zheng[1], Min Chen[1], Fei Ge[1], Zhaoling Li[1], Wenting Sun[1], Zhongrong Guan[1,2], Tianhu Liang[1], Yan Zheng[1], Xiaolong Tan[1], Chaoying Zou[1], Huanwei Peng[3], Guangtang Pan[1] and Yaou Shen[1]**

[1] Key Laboratory of Biology and Genetic Improvement of Maize in Southwest Region, Maize Research Institute, Sichuan Agricultural University, Chengdu, China, [2] Research Center of Tumofous Stem Mustard, Chongqing Yudongnan Academy of Agricultural Sciences, Chongqing, China, [3] Institute of Animal Nutrition, Sichuan Agricultural University, Chengdu, China

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Xiaohong Yang,
China Agricultural University, China
Yang Xu,
Yangzhou University, China

****Correspondence:***
Yaou Shen
shenyaou@sicau.edu.cn

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *31 January 2018*
Accepted: *18 April 2018*
Published: *07 May 2018*

Citation:
Zhang Y, Liu P, Zhang X, Zheng Q, Chen M, Ge F, Li Z, Sun W, Guan Z, Liang T, Zheng Y, Tan X, Zou C, Peng H, Pan G and Shen Y (2018) Multi-Locus Genome-Wide Association Study Reveals the Genetic Architecture of Stalk Lodging Resistance-Related Traits in Maize. Front. Plant Sci. 9:611.
doi: 10.3389/fpls.2018.00611

Stalk lodging resistance, which is mainly measured by stem diameter (SD), stalk bending strength (SBS), and rind penetrometer resistance (RPR) in maize, seriously affects the yield and quality of maize (*Zea mays* L.). To dissect its genetic architecture, in this study multi-locus genome-wide association studies for stalk lodging resistance-related traits were conducted in a population of 257 inbred lines, with tropical, subtropical, and temperate backgrounds, genotyped with 48,193 high-quality single nucleotide polymorphisms. The analyses of phenotypic variations for the above traits in three environments showed high broad-sense heritability (0.679, 0.720, and 0.854, respectively). In total, 423 significant Quantitative Trait Nucleotides (QTNs) were identified by mrMLM, FASTmrEMMA, ISIS EM-BLASSO, and pLARmEB methods to be associated with the above traits. Among these QTNs, 29, 34, and 48 were commonly detected by multiple methods or across multiple environments to be related to SD, SBS, and RPR, respectively. The superior allele analyses in 30 elite lines showed that only eight lines contained more than 50% of the superior alleles, indicating that stalk lodging resistance can be improved by the integration of more superior alleles. Among sixty-three candidate genes of the consistently expressed QTNs, GRMZM5G856734 and GRMZM2G116885, encoding membrane steroid-binding protein 1 and cyclin-dependent kinase inhibitor 1, respectively, possibly inhibit cell elongation and division, which regulates lodging resistance. Our results provide the further understanding of the genetic foundation of maize lodging resistance.

Keywords: maize, stalk lodging resistance, multi-locus GWAS, QTNs, candidate gene

INTRODUCTION

Lodging is one of the most important factors threatening grain yield in maize, and can result in reduced photosynthesis, nutrient transportation, and grain quality (Remison and Dele Akinleye, 1978). The annual yield losses caused by lodging are approximately 5–20% globally (Flintgarcia et al., 2003). In some areas where strong wind and heavy rain occur frequently, the risk of lodging will significantly increase (Adelana, 1980). Some properties of the stem itself are also strongly associated with lodging, such as the structure and mechanical strength of the stem, and the number of vascular bundles (Xu et al., 2017). In addition, Tesso and Ejeta (2011) showed that stalk rot disease can reduce stem strength, which further leads to lodging.

The most direct way to improve breeding populations for quantitative traits is phenotypic selection, where the frequency of favorable alleles is increased within a population over cycles of selection. Previous studies on crop morphological traits showed that plant stem diameter (SD), stalk bending strength (SBS), and rind penetrometer resistance (RPR) are closely associated with stalk lodging in the field (Kashiwagi et al., 2008; Hu et al., 2012, 2013). Furthermore, these three traits are significantly negatively correlated with stalk lodging rate in the field (Ling, 2008). The method for testing RPR involves the use of an electronic rind penetrometer to penetrate the rind of the maize stalk, and the maximum value of penetration is then indicated on the screen of the instrument (Sibale et al., 1992). This method does not affect maize seedling growth.

Genome-Wide Association Study (GWAS) is a very powerful tool for dissecting the genetic basis of complex traits (Korte and Farlow, 2013). To date, GWAS has been used to analyze many agronomic traits such as leaf architecture, maize kernel composition, plant height, oil biosynthesis in maize kernels (Tian et al., 2011; Weng et al., 2011; Cook et al., 2012; Li et al., 2013), and other traits, i.e., Some genetic research on crop lodging has also been conducted using GWAS. Hu et al. (2013) detected a complex polygenic inheritance for SBS-related traits, including the maximum load exerted to breaking (F_{max}), the breaking moment (M_{max}), and critical stress (σ_{max}). A total of seven quantitative trait loci (QTLs) explaining 65.7% of the genotypic variance for these three traits. Ookawa et al. (2010) used chromosome segment substitution lines (CSSL) to identify an effective QTL, *SCM2,* for culm strength in rice, and the near-isogenic line (NIL) carrying *SCM2* showed enhanced culm strength. Moreover, Lin et al. (2005) detected another six QTLs for stem strength, culm wall thickness, pith diameter, and stem diameter using a doubled-haploid (DH) population. Conversely, GWAS for maize lodging has rarely been reported, and the molecular mechanisms of variation for maize lodging-related traits remain poorly understood.

Currently, the Bonferroni correction is applied to control the false positive rate for single-marker GWAS, and some important loci with small effects could be excluded by this stringent correction. Multi-locus GWAS methodologies, such as FASTmrEMMA, ISIS EM-BLASSO, mrMLM, pLARmEB, and FarmCPU, have been shown to effectively resolve this issue. The first four methods have higher power and accuracy for quantitative trait nucleotide (QTN) detection and are more suitable for genetic models (Liu et al., 2016; Wang et al., 2016; Tamba et al., 2017; Wen et al., 2017; Zhang J. et al., 2017). Additionally, a combination of various methods for multi-locus GWAS has also been used to control the false positive rate (Wu et al., 2016; Misra et al., 2017).

Our objectives were to (i) estimate the genetic variance and heritability of SD, SBS, and RPR; (ii) estimate the correlations between these three traits; (iii) detect significant quantitative trait nucleotides (QTNS) for SD, RPR, and SBS in multiple environments; (iv) dissect the genetic basis of variation of lodging-related traits in maize, and (v) identify candidate genes controlling maize stalk lodging-related traits.

MATERIALS AND METHODS

Phenotyping of Maize Lodging-Related Traits

The SD, SBS, and RPR tests were conducted in an association-mapping panel of 257 diverse inbred lines, which were collected from tropical or subtropical and temperate regions (Li et al., 2013). The names and pedigree information for this association panel are presented in **Table S1**. The 257 inbred lines were planted in three locations: Xishuangbanna (XSBN, N22°0, E100°79′, Yunnan province, China, 2014), Bijie at Guizhou (GZ, N27°32′, E105°29′, Guizhou province, China, 2014), and Wenjiang (WJ, N30°97, E103°81′, Chengdu, Sichuan province, China, 2014). The 257 inbred lines were sown in a randomized complete block design in two replications. Each plot consisted of a single row (14 plants) that was 3 m in length and 0.75 m from the next row, and the plant density was approximately 62,000 individuals per hectare. Each line was grown in a single-row.

At the flowering stage, 10 plants from each line from each replication were randomly selected for phenotyping and their mean values were computed for the three traits: SD, SBS, and RPR, as detailed in Wang L. M. et al. (2012). Briefly, a Vernier caliper was used to measure the SD (mm) of the 15-cm region above ground. A plant stalk strength appliance SY-S03 with a measuring range from 5 to 500 N and a resolution ratio of 0.1 N (Shijiazhuang Shiya Technology Co., Ltd) was used to measure RPR and SBS, and the units of RPR and SBS are N/mm^2 and N, respectively.

Statistical Analysis of the Phenotype

SPSS version 21.0 (IBM, Armonk, NY, 2012) was used to analyze the phenotypic data, including descriptive statistics (mean, range, standard deviation, skewness, kurtosis) and the correlation analysis. To obtain the best linear unbiased prediction (BLUP) of the three lodging-related traits, the R package lme4 (version 3.4.2, https://www.r-project.org/) was fitted to each genotype: Phenotype ∼ (1|Genotype) + (1|Repeat%in%Environment) + (1|Genotype&Environment). Broad-sense heritability (h^2) for each trait was estimated as described by Knapp (Knapp et al., 1985) as: $h^2 = \sigma_g{}^2/(\sigma_g{}^2+\sigma_{gy}{}^2/r+\sigma_e{}^2/yr)$, where $\sigma_g{}^2$, $\sigma_{gy}{}^2$, and $\sigma_e{}^2$ are genetic, genotype-by-environment interaction and residual error variances, respectively, r is the number of replications, and y is the number of environments. All the variances were calculated using a general linear model in SPSS.

Genotyping and ML-GWAS

Using publicly available genotypic data from previous studies, all the 257 lines of the association panel were genotyped using the Maize SNP50 BeadChip (Illumine, San Diego, CA), which contains 56,110 SNP loci (Ganal et al., 2011; Yang et al., 2011; Li et al., 2012). A total of 48,193 high-quality SNPs with a minor allele frequency (MAF) ≥0.05 were used in this study (http://www.maizego.org/Resources.html). A total of 500 SNPs for each chromosome were randomly selected to calculate population structure, as described by (Pritchard et al., 2009). Briefly, five independent simulations with 500,000 Markov Chain Monte Carlo (MCMC) replications and 5,000 SNPs were performed with the number of subpopulations (k) ranging from 1 to 12. The results calculated by STRUCTURE software were submitted to the website http://taylor0.biology.ucla.edu/structureHarvester/, and the optimal k was inferred. The relative kinship (K matrix) between the lines was calculated as previously described in Wang et al. (2016) and Zhang J. et al. (2017). Four multi-locus GWAS methods including mrMLM, FASTmrEMMA, pLARmEB, and ISIS EM-BLASSO were used in this study. All parameters were set at default values (Wang et al., 2016; Tamba et al., 2017; Wen et al., 2017; Zhang J. et al., 2017).

Annotation of Candidate Genes and Pathway Enrichment Analysis

Those genes with common SNPs in the GWAS result were selected as candidate genes. The maize inbred line B73 assembly v2 that was used as the reference genome for the candidate gene analyses was publicly available on the MaizeGDB genome browser (Andorf et al., 2010). The methods of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for these candidate genes were annotated as described by Zhang Y. et al. (2017).

RESULTS

Diversity and Heritability of the Three Lodging-Related Traits

The phenotypic characteristics for SD, SBS, and RPR across the three environments are shown in **Table 1** and **Figure 1**. As shown in **Table 1**, the skewness and kurtosis were less than 1 for SD and RPR, indicating that SD and RPR followed a normal distribution. SBS was slightly skewed to the left (**Figures 1A,C,E**). For the above three traits, the means of phenotypic values decreased from XSBN, WJ, to GZ; the coefficients of variation ranged from 11.87∼14.2, 36.42∼49.29, and 19.43∼23.45 (%), respectively (**Table 1**, **Figures 1B,D,F**).

The results of correlation analysis were showed in **Table 2**. Significant correlations between the traits across three environments were observed. For example, the correlation coefficients between SD and SBS in GZ, WJ, and XSBN were 0.762, 0.615 and 0.668 (*P*-values <0.01), respectively; the correlations between SD and RPR across three environments were relatively smaller ($0.219 < r < 0.308$, $P < 0.01$) than those between SBS and RPR ($0.507 < r < 0.652$, $P < 0.01$). In addition, a significant correlation between different environments was observed for each of the three traits (**Table 2**).

In the analysis of variance for the three traits, highly significant variations for genotypes (G) and environments (E) and significant variation for genotype-by-environment interaction were found (**Table 3**). This indicates the important roles of both environment and G × E interaction. The broad-sense heritabilities (h^2) for SD, SBS, and RPR across the three environments in the 257 inbred lines ranged from 0.679 (SD) to 0.854 (RPR), indicating the predominant role of genetic factors for these traits (**Table 3**).

TABLE 1 | Phenotypic performance of the three lodging resistance-related traits in 257 inbred lines under three environments.

Trait	Env.	Mean	Range	SDD	CV (%)	Skew	Kurt
SD	GZ	13.62	9.23–19.44	1.93	14.20	0.43	−0.16
	WJ	15.80	12.03–21.56	1.88	11.87	0.40	−0.26
	XSBN	18.19	12.11–25.19	2.22	12.18	0.20	−0.06
SBS	GZ	21.26	6.43–64.94	10.48	49.29	1.22	1.65
	WJ	25.33	7.11–70.58	10.68	42.17	1.15	1.77
	XSBN	32.25	7.91–70.84	11.74	36.42	0.70	0.32
RPR	GZ	39.86	20.24–73.79	9.35	23.45	0.73	0.62
	WJ	41.07	23.90–67.46	7.98	19.43	0.64	0.22
	XSBN	45.03	27.54–79.79	8.87	19.70	0.82	1.39

SD (stalk diameter) is measured in the unit of millimeter (mm), SBS (stalk bending strength) is measured in the unit of newton (N) and RPR (rind penetrometer resistance) is measured in the unit of newton per square millimeter (N/mm^2).
Env. Represents environments; GZ, WJ, and XSBN represent Guizhou, Wenjiang and Xishuangbanna, respectively. SDD, standard deviation.
CV, coefficient of variation.

QTNs Identified by ML-GWAS

The ΔK calculation of STRUCTURE indicated a peak ($K = 2$) in the broken line graph reflecting the number of subpopulations (K) (**Figures S1A,B**), indicating that the 257 maize inbred lines could be divided into two subpopulations. Owning to significant variations for each of the three lodging-related traits in 257 maize inbred lines across the three locations, BLUP values across the three locations were also used for the GWAS. In total, 423 significant QTNs were identified at the critical logarithm of odds (LOD) score (≥3) for these traits in the three environments using mrMLM, FASTmrEMMA, PLARmEB and ISIS EM-BLASSO (**Table S2**, **Figure S2**).

A total of 126 significant QTNs, mainly distributed on chromosomes 1, 2, 3, 5, 6, 8, and 9, were detected to be associated with SD (**Table S2**, **Figure S2A**). Among them, 29 QTNs were common across the methods or the locations. The LOD of these 32 QTNs identified by mrMLM ranged from 3.03 to 6.25, and the percentage of phenotypic variation explained by each QTN (PVE) in GZ, WJ, XSBN, and BLUP was 30.96, 40.90, 44.21, and 54.38 (%), respectively. The LOD scores of the significant 21 QTNs identified by FASTmrEMMA ranged from 3.08 to 6.21, and the PVE in GZ, WJ, XSBN, and BLUP for SD was 19.51, 20.51, 22.31, and 21.25 (%), respectively. For PLARmEB, the LOD scores of the 43 QTNs ranged from 3.01 to 7.83 in GZ, WJ, XSBN, and BLUP, and PVE was 13.84, 35.20, 31.17, and 36.84 (%), respectively. The LOD scores of the 66 QTNs detected by ISIS EM-BLASSO ranged from 3.00 to 14.08, and the PVE in GZ, WJ, XSBN and BLUP was 51.15, 41.37, 48.99, and 44.37 (%), respectively.

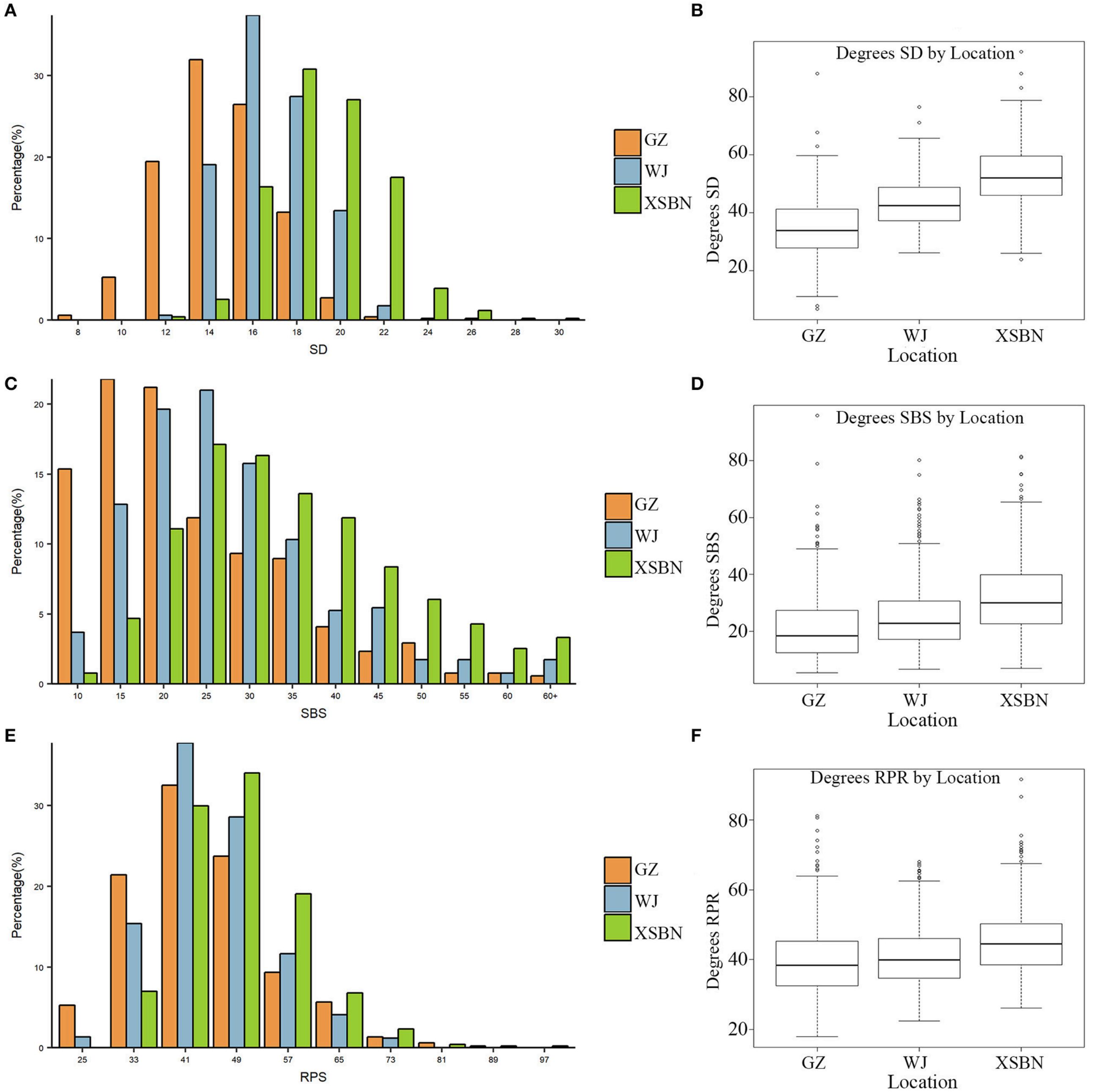


FIGURE 1 | Frequency distributions of SD **(A)**, SBS **(C)**, RPR **(E)** in 257 maize inbred lines and the boxplots for SD **(B)**, SBS **(D)**, RPR **(F)** in the three environments.

In total, 148 significant QTNs were correlated with SBS, and were evenly distributed on 10 chromosomes under the environments and BLUP model. Among them, 35 QTNs were common across the methods or the locations. The LOD values of the 148 QTNs identified by mrMLM, FASTmrEMMA, PLARmEB, and ISIS EM-BLASSO were in the range of 3.01∼8.78, 3.32∼10.75, 3.09∼8.69, and 3.05∼12.18, respectively (**Table S2**, **Figure S2B**). Among these QTNs, 43 identified by mrMLM explained 58.31, 53.47, 52.02, and 57.69 (%) of the phenotypic variation in GZ, WJ, XSBN, and BLUP for SBS, respectively. Conversely, 26.98, 43.87, 17.02, and 26.10 (%) of the phenotypic variation was separately explained by 28 QTNs using FASTmrEMMA. Using PLARmEB, the PVE was 28.77, 30.40, 17.31, and 49.24 (%) in the different environments, respectively. The PVE in GZ, WJ, XSBN, and BLUP for SBS was 53.88, 64.19, 56.12, and 49.64 (%), respectively, for the 73 QTNs by ISIS EM-BLASSO.

TABLE 2 | Phenotypic correlation coefficients between lodging resistance-related traits across three environments.

Trait		SD			SBS			RPR		
	Environment	**GZ**	**WJ**	**XSBN**	**GZ**	**WJ**	**XSBN**	**GZ**	**WJ**	**XSBN**
SD	GZ	1								
	WJ	0.314**	1							
	XSBN	0.356**	0.568**	1						
SBS	GZ	0.762**	0.297**	0.255**	1					
	WJ	0.349**	0.615**	0.319**	0.524**	1				
	XSBN	0.300**	0.387**	0.668**	0.385**	0.485**	1			
RPR	GZ	0.219**	0.238**	0.131*	0.507**	0.381**	0.352**	1		
	WJ	0.078	0.308**	0.099	0.324**	0.652**	0.307**	0.661**	1	
	XSBN	0.036	0.279**	0.274**	0.283**	0.391**	0.614**	0.688**	0.644**	1

,Indicate significance level at $P < 0.05$ and 0.01, respectively.*
Env. Represents environments; GZ, WJ, and XSBN represent Guizhou, Wenjiang, and Xishuangbanna, respectively.

TABLE 3 | Analysis of variance (ANOVA) for lodging resistance-related traits of 257 lines in three environments.

Trait	Source of variation	Mean square	F	Significance	H^2
SD	Environment (E)	2,679.898	1,046.302	<0.01**	0.679
	Genotype (G)	14.811	5.783	<0.01**	
	Replication	9.878	3.857	0.051	
	G × E	4.761	1.859	<0.01**	
	Residual Error	2.561		<0.01**	
SBS	Environment (E)	15,870.661	288.548	<0.01**	0.720
	Genotype (G)	463.779	8.432	<0.01**	
	Replication	35.159	0.639	0.424	
	G × E	129.899	2.362	<0.01**	
	Residual Error	55.002		<0.01**	
RPR	Environment (E)	3,761.979	127.496	<0.01**	0.854
	Genotype (G)	355.790	12.058	<0.01**	
	Replication	123.123	4.173	0.042*	
	G × E	51.886	1.758	<0.01**	
	Residual Error	29.507		<0.01**	

,Indicate significance level at $P < 0.05$ and 0.01, respectively.*

We detected a total of 149 RPR-associated QTNs with LODs ranging from 3.01 to 14.39 in the three environments and BLUP model, and were mainly located on chromosomes 1, 2, 4, 5, 7, 8, and 9 (**Table S2**, **Figure S2C**). And 47 QTNs were common across the methods or the locations. Among these, four QTNs were also detected in SBS traits. Of the 149 RPR-associated QTNs, 54, 31, 57, and 74 QTNs were separately identified by mrMLM, FASTmrEMMA, PLARmEB, and ISIS EM-BLASSO, which explained 60.91∼67.76, 23.53∼35.38, 30.90∼56.86, and 45.28∼63.77 (%) of the phenotypic variation, respectively.

Verification of the Common QTNs by Multi-Methods or Across Environments

A total of 107 QTNs were co-identified by at least two of the methods or across different environments, among which 29, 34, and 48 were associated with SD, SBS, and RPR, respectively (**Table S3** and **Figure 2**). To verify the significance of each common QTN, we divided the population into two groups according to their allele types and compared the mean phenotypic values between the two groups. For SD, the average of the group containing the superior alleles was significantly greater than the group containing inferior alleles, with the exception of the SNPs SYN35339, SYN6428, PZE-102085765, and PZE-101121408 (**Table S4**). As for SBS and RPR, the group with the superior alleles showed a significantly larger mean than the group with inferior alleles for every common SNP (**Table S4**). These results verified the reliability and significance of the common QTNs identified by these ML-GWAS methods.

Utilization of Superior Alleles in Elite Maize Lines

Thirty elite inbred lines from China and America that have excellent agronomic traits and serve as the parents of commercialized hybrid varieties were included in the maize population, which enabled us to evaluate the utilization of the superior alleles for maize breeding. The results indicated that the percentage of SD superior alleles in the elite lines ranged from 27.59 to 55.17% (**Table S5**). The lines with >15 superior alleles indicated a significantly higher SD phenotypic value, with an average of 14.50 in GZ, 16.75 in WJ, and 20.72 in XSBN, whereas the lines with 0∼10 superior alleles had average SD values of 12.78, 14.13, and 16.61 in GZ, WJ, and XSBN, respectively (**Table S5**, **Figure 3A**). The utilization of the SBS superior alleles in the elite lines ranged from 25.71 to 65.71% (**Table S5**). The phenotypic averages of the lines with >20 superior alleles were 30.56, 53.68, and 44.83 in GZ, WJ, and XSBN, respectively, whereas those with 15∼20 superior alleles had a lower average of 15.14, 15.87, and 25.85 in GZ, WJ, and XSBN, respectively (**Table S5**, **Figure 3B**). As for RPR, the elite lines contained 29.17∼66.67% of the superior alleles (**Table S5**). The average RPR in the lines with >30 superior alleles were 45.06, 50.69, and 49.01 in GZ, WJ, and XSBN, respectively; however, those lines with <20 superior alleles had average RPR values of 29.05, 32.43, and 35.97 in GZ, WJ, and XSBN, respectively (**Table S5**, **Figure 3C**).

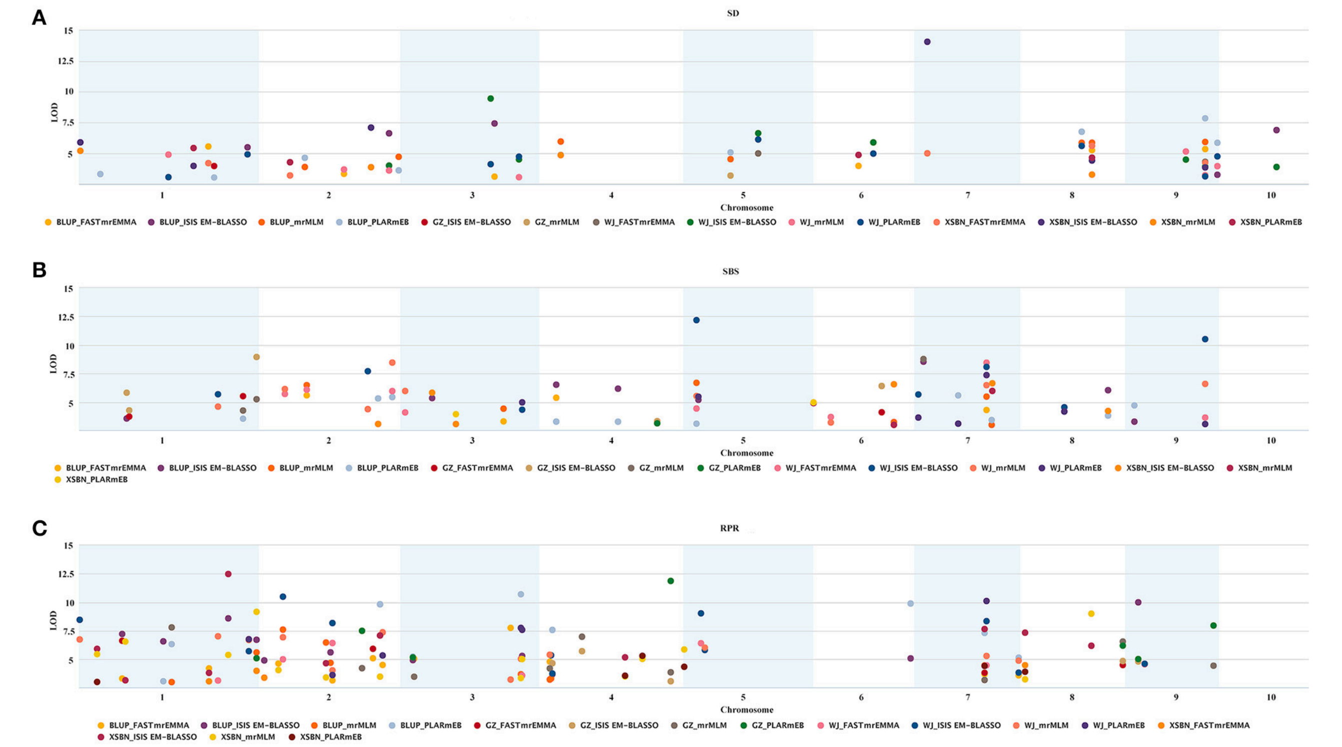


FIGURE 2 | Repeatability and significance of the SNPs associated with the three lodging resistance-related traits in the three environments and BLUP. The significance threshold is LOD = 3.0. **(A–C)** Represent SD, SBS, and RPR, respectively.

The results suggested that these superior alleles had an additive effect on the lodging resistance-related traits. Further analysis indicated that only eight of the 30 elite inbred lines had more than 50% utilization of all the superior alleles (**Table S5**, **Figure 4**), implying that the superior alleles were not efficiently selected during maize breeding. In future work, integrated utilization of the superior alleles would be an efficient approach for lodging-resistance breeding in maize by marker-assisted selection (MAS).

Candidate Genes Associated With Common QTNs

To further understand the molecular basis of lodging-related traits, we focused on the candidate genes that were directly associated with the common QTNs. As a result, 19, 17, and 30 candidate genes around their common QTNs were found to be associated with SD, SBS, and RPR, respectively. The annotations for the candidate genes are displayed in **Table S3**, with seven transcription factors, eight kinase-related proteins, and four transport proteins involved. These genes mainly participate in metabolic pathway, genetic information processing, environmental information processing, cellular processes, and organismal systems (**Table 4**).

DISCUSSION

According to previous studies, the strength of the maize stalk depends on the tissue and morphology, and the morphology of the stalk is largely determined by the mechanical stresses in maize (Von et al., 2015). SD, SBS, and RPR were demonstrated to show potential as selective breeding indexes for improving lodging resistance (Liu et al., 2011; Xiang et al., 2016). The heritability and genetic models vary among different studies since the calculations depend on the experimental populations, design, and conditions (Lynch and Walsh, 1998). The genetic architecture of lodging resistance-related traits has been illustrated in diverse maize populations by linkage mapping. (Kashiwagi et al., 2008; Hu et al., 2012, 2013). However, the genetic basis and the molecular pathways underlying lodging resistance-related traits, as well as the major genes associated with the traits, remain largely unknown. In this study, we interpreted the natural variation and revealed the genetic architecture of three lodging resistance-related traits based on 257 maize inbred lines by ML-GWAS analysis. And identified the candidate genes and their possible pathways for stalk lodging resistance.

Genetic Basis of Lodging-Related Traits

In this study, the three lodging-related traits exhibited wide phenotypic variation and were normally distributed. ANOVA showed that the genetic effects and interactive effects between the genetics and environment were both significant for these traits, and the heritability (h^2) was very high for SD, SBS, and RPR. Previous studies on SD in different crops mainly focused on the phenotypic correlations with stalk mechanical strength and the identification of QTLs for SD, whereas the heritability of SD has

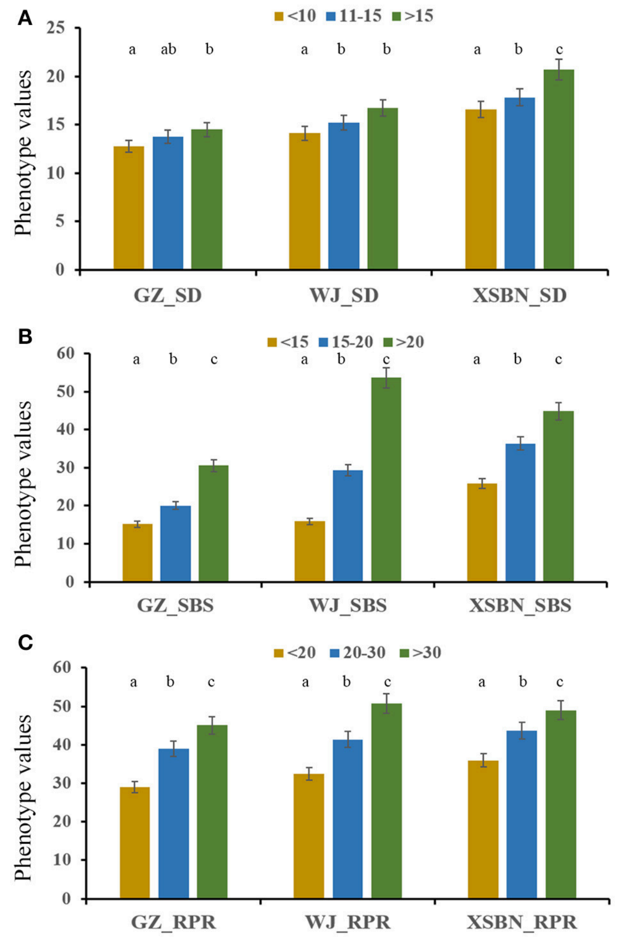


FIGURE 3 | The phenotypic values in the maize elite inbred lines with different numbers of superior alleles for SD **(A)**, SBS **(B)**, and RPR **(C)**.

not been investigated (Lin et al., 2005; Kashiwagi et al., 2008). In our study, h^2 was 0.679 across the three environments for SD. The $F_{\max}$, $M_{\max}$, and $\sigma_{\max}$ can be used as tools to determine SBS accreditation (Timoshenko and Gere, 1972). An h^2 of 0.84 for $F_{\max,}$ was reported in rice, and in maize the h^2 for $F_{\max}$, $M_{\max}$, and $\sigma_{\max}$ were 0.81, 0.79, and 0.75, respectively (Sun, 1987; Hu et al., 2013). Both these estimates are in close agreement with the estimates of SBS ($h^2 = 0.720$) in our study (**Table 3**). The h^2 estimates were previously found to range from 0.81 to 0.92 for RPR in different maize populations (Flintgarcia et al., 2003; Hu et al., 2012), which corroborates our value of 0.854 across the three environments. In combination with previous results, our findings suggest that all the measured lodging-related traits showed high precision and that the three lodging resistance-related traits generally exhibited high heritability.

Phenotypic correlations were observed among the three lodging-related traits. For instance, the correlation coefficient between SBS and RPR was 0.507 in GZ, 0.652 in WJ, and 0.614 in XSBN, respectively (**Table S2**). Meanwhile, we identified four QTNs (PZE-101187823, SYN31353, PZE-105036664, and PZE-107063605), all of which were associated with both SBS and RPR (**Table S3**). The above results suggested that some genetic factors were shared among these lodging resistance-related traits.

Common Candidate Genes Reveal the Possible Molecular Basis of Lodging Resistance

No previous studies have reported on GWAS for SD, SBS, and RPR in maize. However, some studies have evaluated the QTLs. Hu et al. (2013) detected two, three, and two QTLs for $F_{\max}$, $M_{\max}$, and $\sigma_{\max}$, respectively, using 216 recombinant inbred lines and 129 SSR markers. Among them, a QTL of $\sigma_{\max,}$ an important parameter for characterizing SBS, was located in markers umc1993 and bulg1450. In the present research, a QTN on Chr10 (position: 137282081 bp) for SBS locates exactly in the interval of the $\sigma_{\max}$ QTL reported by Hu et al. (2013) (**Table S2**). The remaining QTNs in the present study are the first to be reported as associated with lodging resistance-related traits in maize.

Furthermore, we identified 63 common candidate genes in total that were around common QTNs for lodging-related traits. Notably, GRMZM5G856734 encodes Membrane steroid-binding protein 1 (MSBP1) (**Table S3**), whose homologous gene *MSBP1* in *Arabidopsis thaliana* was proven to be involved in the inhibition of cell elongation (Yang et al., 2005). Interestingly, the candidate gene GRMZM2G116885 that encodes cyclin-dependent kinase inhibitor 1 was associated with both SBS and RPR. The homologous gene of GRMZM2G116885 in *Arabidopsis* was reported to be involved in coordinated cell growth or cell division (Bemis and Torii, 2007). It is generally known that cell elongation and cell wall thickening regulate plant lodging resistance (Fan et al., 2017). According to RNA-Seq data from the previous study, the candidate genes GRMZM5G856734 and GRMZM2G116885 had high expression levels in maize stems, with the FPKM are 115.5 and 58.0, respectively (Sekhon et al., 2012). In addition, more than 90% of the candidate genes found in our study were expressed in maize stems, especially the expression levels of GRMZM2G038126, GRMZM2G073934, GRMZM2G058584, and GRMZM2G084181 were extremely high (**Table S3**). The functional validation of these genes should be addressed in future work.

Additionally, seven candidate genes were classified into transcription factors based on their functional annotation, including ethylene-responsive transcription factor 12, bHLH-transcription factor 105, bHLH-transcription factor 65, GRAS transcription factor, transcription factor VOZ1, and MYB 9 transcription factor (**Table S3**). Transcription factors are a group of proteins that regulate targeted gene expression in particular cells at a certain time, and are vital for cell division, growth, and death (Latchman, 1997; Riechmann and Meyerowitz, 1998; Guilfoyle and Hagen, 2007).

The Superiority of the New ML-GWAS

Previous studies demonstrated that the single-locus GWAS was useful to dissect complex agronomic trait by using general linear models (GLMs) and mixed linear models (MLMs) (Zhang et al., 2010; Wang M. et al., 2012). High false positive rates are an

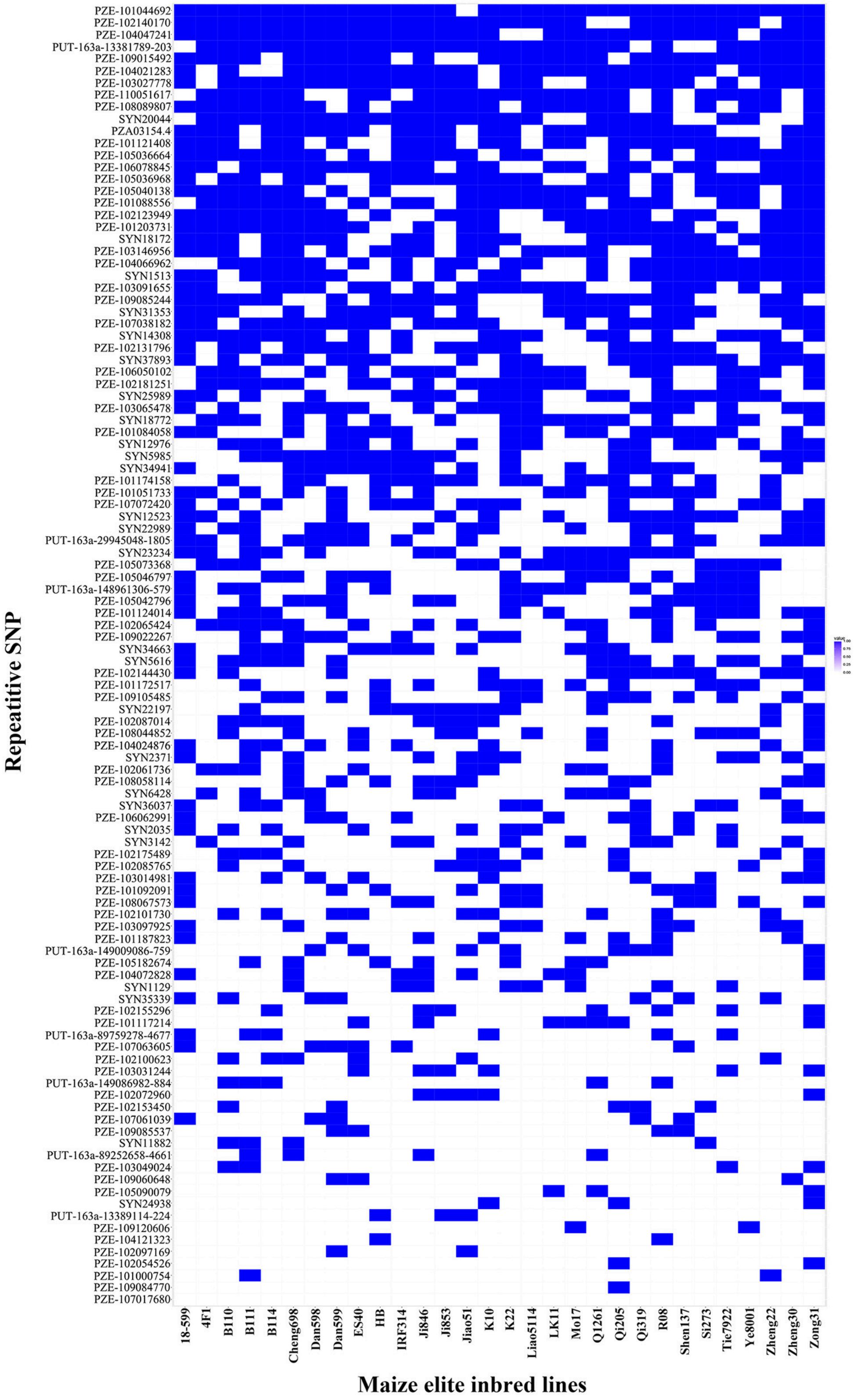


FIGURE 4 | The superior allele SNP distributions in the 30 maize elite inbred lines. Blue and white colors represent superior and inferior alleles, respectively.

TABLE 4 | SNPs, chromosomal position and the pathway of candidate genes significantly associated with three lodging resistance-related traits identified by multi GWAS methods across all environments.

SNPs	Traits	Genes	Genotype	Chr	Gene interval (bp)	Annotation	Pathway class A	Pathway class B
PZE-103065478	SBS	GRMZM2G018447	A/G	3	93187188~93205169	Ubiquitin-conjugating enzyme 15	Genetic Information Processing	Folding, sorting and degradation
PZE-105090079	SD	GRMZM2G038126	A/G	5	125802058~125817596	26S protease regulatory subunit 6B homolog	Genetic Information Processing	Folding, sorting and degradation
PZE-104021283	RPR	GRMZM2G047800	A/G	4	22835928~22842145	NAD(P)-binding Rossmann-fold superfamily protein	Metabolism	Metabolism of terpenoids and polyketides
SYN37893	RPR	GRMZM2G051101	T/G	4	219995778~219999760	Putative seven in absentia domain family protein	Genetic Information Processing	Folding, sorting and degradation
SYN20044	RPR	GRMZM2G058584	A/G	1	285124791~285129572	Histidinol dehydrogenase chloroplastic	Metabolism	Amino acid metabolism
							Metabolism	Global and Overview
PZE-102123949	RPR	GRMZM2G067514	A/G	2	172399126~172403574	Phosphoglycerate mutase family protein	Genetic Information Processing	Folding, sorting and degradation
							Genetic Information Processing	Transcription
PZE-102144430	RPR	GRMZM2G083504	A/G	2	191378248~191380615	Transcription factor bHLH62	Environmental Information Processing	Signal transduction
							Organismal Systems	Environmental adaptation
SYN1129	RPR	GRMZM2G084181	A/G	2	9114721~9126810	ABC transporter C family member 2	Environmental Information Processing	Membrane transport
PZE-104047241	RPR	GRMZM2G103721	A/G	4	71631328~71640766	Phosphatidylinositol 3-kinase VPS34	Cellular Processes	Transport and catabolism
							Environmental Information Processing	Signal transduction
							Metabolism	Carbohydrate metabolism
PZE-101172517	SD	GRMZM2G119357	T/C	1	216471619~216479224	Chromatin remodeling protein EBS	Cellular Processes	Transport and catabolism
							Genetic Information Processing	Folding, sorting and degradation
							Genetic Information Processing	Transcription
SYN34663	RPR	GRMZM2G135341	A/G	7	174359079~174365275	BADH-like protein	Metabolism	Amino acid metabolism
SYN18172	RPR	GRMZM2G138255	T/G	8	8857140~8860877	ARM repeat superfamily protein	Cellular Processes	Transport and catabolism
SYN11882	SBS	GRMZM2G155312	A/G	1	275140650~275150229	Leucine-rich repeat protein kinase family protein	Organismal Systems	Environmental adaptation
PZE-107017680	SBS	GRMZM2G156692	T/G	7	15316520~15323155	proline-rich family protein	Genetic Information Processing	Transcription
SYN2035	SBS	GRMZM2G304638	A/G	1	297169418~297180338	BEACH domain-containing protein C2	Metabolism	Amino acid metabolism
							Metabolism	Global and Overview
PZE-101203731	RPR	GRMZM2G324276	T/C	1	250294172~250296186	Core-2/I-branching beta-16-N-acetylglucosaminyltransferase family protein	Cellular Processes	Transport and catabolism
PZE-108089807	SBS	GRMZM2G375975	T/C	8	146830973~146836666	Putative MAP kinase family protein	Human Diseases	Endocrine and metabolic diseases
SYN5616	RPR	GRMZM2G431309	A/G	2	207837259~207841334	GRAS transcription factor	Environmental Information Processing	Signal transduction

obvious shortcoming for GLMs, because there is no kinship among materials as covariate (Pace et al., 2015). The screening criteria of significance for single-locus GWAS is $P = 0.05/m$ (m is the number of markers) (Perneger, 1998). Owning to large number of SNPs (Gordon et al., 2016), some important loci might be excluded under the stringent criteria in MLM. To remedy the shortcomings of the methods mentioned above, ML-GWAS methods have recently been explored, including mrMLM (Wang et al., 2016), pLARmEB (Zhang J. et al., 2017), ISIS EM-BLASSO (Tamba et al., 2017), and FASTmrEMMA (Wen et al., 2017). Several studies have individually analyzed published data using the multi-locus methods, and have indicated that these methods constituted effective approaches with high detection power and less stringent criteria (Wang et al., 2016; Tamba et al., 2017; Wen et al., 2017; Zhang J. et al., 2017). In our study, a total of 126, 77, 176, and 230 significant QTNs were detected for three lodging-related traits using mrMLM, FASTmrEMMA, pLARmEB, and ISIS EM-BLASSO, respectively (**Figure S2**, **Table S2**). A comparison of the four methods showed that ISIS EM-BLASSO was more powerful than the other three multi-locus methods in the identification of QTNs for lodging resistance-related traits (**Table S2**, **Figure S2**). Furthermore, some stably expressed QTNs were detected in the multi-environment and BLUP model using multi-methods (**Tables s2**, **s3**). Notably, the candidate genes GRMZM5G856734 and GRMZM2G116885, were proven to inhibit cell elongation and division, which regulates lodging resistance. However, only 4, 4, and 7 SNPs were detected for SD, SBS, and RPR, respectively, from FarmCPU (R packages FarmCPU, K and PCA calculated by SPAGeDi software and GAPIT package, respectively. The threshold is P-value = 0.05/48193) (Table S6). In addition, six of these SNPs were also be detected by ML-GWAS methods. Using GAPIT (R packages GAPIT) method, only one SD-associated SNP was found in XSBN, which was also detected in the ML-GWAS methods (Table S6). Our study demonstrated that improved efficiency and accuracy could be achieved by a combination of the new multi-locus methods for identification of lodging resistance-related QTNs in maize.

CONCLUSIONS

SD, SBS, and RPR were used in this study to dissect the genetic basis of stalk lodging resistance in maize using ML-GWAS methods. Among all the significantly associated QTNs for the three traits, 107 were commonly identified across multiple methods or environments. Around these common QTNs, sixty-three candidate genes were responsive for maize lodging resistance. These QTNs provide the important information for the marker-assisted selection, and these candidate genes should improve our understanding of the molecular basis of maize lodging resistance.

AUTHOR CONTRIBUTIONS

YS and PL designed the experiments. YLZ, XZ, QZ, MC, FG, ZL, WS, and YZ performed the analysis. YLZ, YS, ZG, TL, YZ, XT, CZ, HP, and GP drafted the manuscript. All the authors critically revised and approval the final version of this manuscript.

FUNDING

This work was supported by the National Basic Research Program of China (2014CB138200) and the Young Scientists Fund of Sichuan Province (2016JQ0008).

ACKNOWLEDGMENTS

We thank Jianbing Yan (Huazhong Agricultural University) for providing the maize population.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00611/full#supplementary-material

Figure S1 | Population structure of the 257 maize inbred lines based on 48,193 SNP markers. **(A)** Plot of delta.K against putative K ranging from 1 to 12. **(B)** Stacked bar plot of ancestry relationship of the natural population.

Figure S2 | Manhattan plots showing all the significant SNPs associated with lodging resistance-related traits using four ML-GWAS methods across three environments and BLUP. **(A–C)** represent SD, SBS, and RPR, respectively. Points of different colors represent different methods and environments.

Table S1 | Pedigree information of the maize accessions used in this study.

Table S2 | ML-GWAS detected significant signals associated with SD, SBS, and RPR across the three environments and BLUP.

Table S3 | Repetitive SNPs and their information by ML-GWAS consistently identified in multiple methods or environments.

Table S4 | Distribution of the important SNPs superior alleles in the 257 inbred lines.

Table S5 | Distribution and utilization percentage of the important SNPs superior alleles in the 30 maize elite inbred lines.

Table S6 | Significant signals associated with SD, SBS, and RPR, detected by FarumCPU and GAPIT.

REFERENCES

Adelana, B. O. (1980). Relationship between lodging, morphological characters and yield of tomato cultivars. *Sci. Hortic.* 13, 143–148. doi: 10.1016/0304-4238(80)90078-3

Andorf, C. M., Lawrence, C. J., Harper, L. C., Schaeffer, M. L., Campbell, D. A., and Sen, T. Z. (2010). The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps. *Bioinformatics* 26, 434. doi: 10.1093/bioinformatics/btp556

Bemis, S. M., and Torii, K. U. (2007). Autonomy of cell proliferation and developmental programs during Arabidopsis aboveground organ morphogenesis. *Dev. Biol.* 304, 367. doi: 10.1016/j.ydbio.2006.12.049

Cook, J. P., McMullen, M. D., Holland, J. B., Tian, F., Bradbury, P., Ross-Ibarra, J., et al. (2012). Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiol.* 158, 824. doi: 10.1104/pp.111.185033

Fan, C., Li, Y., Hu, Z., Hu, H., Wang, G., Li, A., et al. (2017). Ectopic expression of a novel *OsExtensin-like* gene consistently enhances plant lodging resistance by regulating cell elongation and cell wall thickening in rice. *Plant Biotechnol. J.* 16, 254–263. doi: 10.1111/pbi.12766

Flintgarcia, S. A., Jampatong, C., Darrah, L. L., and Mcmullen, M. D. (2003). Quantitative Trait Locus Analysis of Stalk Strength in Four Maize Populations. *Crop Sci.* 43, 13–22. doi: 10.2135/cropsci2003.0013

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., et al. (2011). A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS ONE* 6:e28334. doi: 10.1371/journal.pone.0028334

Gordon, D., Huddleston, J., Chaisson, M. J., Hill, C. M., Kronenberg, Z. N., Munson, K. M., et al. (2016). Long-read sequence assembly of the gorilla genome. *Science* 352:aae0344. doi: 10.1126/science.aae0344

Guilfoyle, T. J., and Hagen, G. (2007). Auxin response factors. *J. Plant Growth Regul.* 10:453. doi: 10.1016/j.pbi.2007.08.014

Hu, H., Liu, W., Fu, Z., Homann, L., Technow, F., Wang, H., et al. (2013). QTL mapping of stalk bending strength in a recombinant inbred line maize population. *Theor. Appl. Genet.* 126, 2257–2266. doi: 10.1007/s00122-013-2132-7

Hu, H., Meng, Y., Wang, H., Hai, L., and Chen, S. (2012). Identifying quantitative trait loci and determining closely related stalk traits for rind penetrometer resistance in a high-oil maize population. *Theor. Appl. Genet.* 124, 1439–1447. doi: 10.1007/s00122-012-1799-5

Kashiwagi, T., Togawa, E., Hirotsu, N., and Ishimaru, K. (2008). Improvement of lodging resistance with QTLs for stem diameter in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 117, 749–757. doi: 10.1007/s00122-008-0816-1

Knapp, S. J., Stroup, W. W., and Ross, W. M. (1985). Exact Confidence Intervals for Heritability on a Progeny Mean Basis. *Crop Sci.* 25, 192–194. doi: 10.2135/cropsci1985.0011183X002500010046x

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. doi: 10.1186/1746-4811-9-29

Latchman, D. S. (1997). Transcription factors: an overview. *Int. J. Exp. Pathol.* 29:1305. doi: 10.1016/S1357-2725(97)00085-X

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45, 43–50. doi: 10.1038/ng.2484

Li, Q., Yang, X., Xu, S., Cai, Y., Zhang, D., Han, Y., et al. (2012). Genome-wide association studies identified three independent polymorphisms associated with α-tocopherol content in maize kernels. *PLoS ONE* 7:e36807. doi: 10.1371/journal.pone.0036807

Lin, H., Guo, H., Xiao, S., Jiang, G., Zhang, X., Yan, C., et al. (2005). Quantitative trait loci (QTL) of stem strength and related traits in a doubled-haploid population of wheat (*Triticum aestivum* L.). *Euphytica* 141, 1–9. doi: 10.1007/s10681-005-4713-2

Ling, G. (2008). Bending mechanical properties of stalk and lodging-resistance of maize (*Zea mays* L.): bending mechanical properties of stalk and lodging-resistance of Maize (*Zea mays* L.). *Acta Agron. Sin.* 34, 653–661. doi: 10.3724/SP.J.1006.2008.00653

Liu, W. W., Zhao, H. J., Hong-Qi, L. I., Zhao, X. J., Yuan, L. G., and Wei-Wei, H. U. (2011). Effects of planting densities and modes on stem lodging resistance of summer maize. *J. Henan Agric. Sci.* 40, 75–78. doi: 10.15933/j.cnki.1004-3268

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12:e1005767. doi: 10.1371/journal.pgen.1005767

Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits.* Sunderland, MA: Sinauer Associates.

Misra, G., Badoni, S., Anacleto, R., Graner, A., Alexandrov, N., and Sreenivasulu, N. (2017). Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. *Sci. Rep.* 7:12478. doi: 10.1038/s41598-017-12778-6

Ookawa, T., Hobo, T., Yano, M., Murata, K., Ando, T., Miura, H., et al. (2010). New approach for rice improvement using a pleiotropic QTL gene for lodging resistance and yield. *Nat. Commun.* 1:132. doi: 10.1038/ncomms1132

Pace, J., Yu, X., and Lübberstedt, T. (2015). Genomic prediction of seedling root length in maize (*Zea mays* L.). *Plant J. Cell Mol. Biol.* 83:903. doi: 10.1111/tpj.12937

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *Br. Med. J.* 316, 1236–1238. doi: 10.1136/bmj.316.7139.1236

Pritchard, J. K., Wen, X., and Falush, D. (2009). *Documentation for STRUCTURE software: version 2.3.* The University of Chicago Press.

Remison, ,U., and Dele Akinleye S., (1978). Relationship between lodging, morphological characters and yield of varieties of maize (*Zea mays* L.). *J. Agric. Sci.* 91, 633–638. doi: 10.1017/S0021859600060019

Riechmann, J. L., and Meyerowitz, E. M. (1998). The AP2/EREBP family of plant transcription factors. *Biol. Chem.* 379:633.

Sekhon, R. S., Childs, K. L., Santoro, N., Foster, C. E., Buell, C. R., Leon, N. D., et al. (2012). Transcriptional and metabolic analysis of senescence induced by preventing pollination in maize. *Plant Physiol.* 159, 1730–1744. doi: 10.1104/pp.112.199224

Sibale, E. M., Darrah, L. L., and Zuber, M. S. (1992). Comparison of two rind penetrometers for measurement of stalk strength in maize. *Maydica* 37, 111–114.

Sun, X. (1987). Studies on the resistance of the culm of rice to lodging. *Sci. Agric. Sin.* 20, 32–37.

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tesso, T., and Ejeta, G. (2011). Stalk strength and reaction to infection by Macrophomina phaseolina of brown midrib maize (Zea mays) and sorghum (*Sorghum bicolor*). *Fuel Ener. Abstr.* 120, 271–275. doi: 10.1016/j.fcr.2010.10.015

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., et al. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* 43, 159–162. doi: 10.1038/ng.746

Timoshenko, S. P., and Gere, J. M. (1972). Mechanics of Materials. Daejeon-si: Van Nostrand Reinhold Company.

Von, F. G., Robertson, D., Lee, S. Y., and Cook, D. D. (2015). Preventing lodging in bioenergy crops: a biomechanical analysis of maize stalks suggests a new approach. *J. Exp. Bot.* 66:4367. doi: 10.1093/jxb/erv108

Wang, L. M., Jian-Sheng, L. I., and Yao, G. Q. (2012). Characterizations of resistance to stalk and root lodging in maize. *J. Maize Sci.* 20, 69–74,81. doi: 10.13597/j.cnki.maize.science.2012.02.015

Wang, M., Yan, J., Zhao, J., Song, W., Zhang, X., Xiao, Y., et al. (2012). Genome-wide association study (GWAS) of resistance to head smut in maize. *Plant Sci. Int. J. Exper. Plant Biol.* 196:125. doi: 10.1016/j.plantsci.2012.08.004

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* doi: 10.1093/bib/bbw145

Weng, J., Xie, C., Hao, Z., Wang, J., Liu, C., Li, M., et al. (2011). Genome-wide association study identifies candidate genes that affect plant height in Chinese elite maize (*Zea mays* L.) inbred lines. *PLoS ONE* 6:e29229. doi: 10.1371/journal.pone.0029229

Wu, X., Li, Y., Shi, Y., Song, Y., Zhang, D., Li, C., et al. (2016). Joint-linkage mapping and GWAS reveal extensive genetic loci that regulate male inflorescence size in maize. *Plant Biotechnol. J.* 14, 1551–1562. doi: 10.1111/pbi.12519

Xiang, D. B., Zhao, G., Wan, Y., Tan, M. L., Song, C., and Song, Y. (2016). Effect of planting density on lodging-related morphology, lodging rate, and yield of tartary buckwheat (*Fagopyrum tataricum*). *Plant Prod. Sci.* 19, 1–10. doi: 10.1080/1343943X.2016.1188320

Xu, C., Gao, Y., Tian, B., Ren, J., Meng, Q., and Wang, P. (2017). Effects of EDAH, a novel plant growth regulator, on mechanical strength, stalk vascular bundles and grain yield of summer maize at high densities. *Field Crops Res.* 200, 71–79. doi: 10.1016/j.fcr.2016.10.011

Yang, X., Gao, S., Xu, S., Zhang, Z., Prasanna, B. M., Li, L., et al. (2011). Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. *Mol. Breed.* 28, 511–526. doi: 10.1007/s11032-010-9500-7

Yang, X. H., Xu, Z. H., and Xue, H. W. (2005). Arabidopsis membrane steroid binding protein 1 is involved in inhibition of cell elongation. *Plant Cell* 17:116. doi: 10.1105/tpc.104.028381

Zhang, J., Feng, J., Ni, Y., Wen, Y., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity (Edinb)*. 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y., Ge, F., Hou, F., Sun, W., Zheng, Q., Zhang, X., et al. (2017). Transcription factors responding to Pb stress in maize. *Genes* 8:231. doi: 10.3390/genes8090231

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42:355. doi: 10.1038/ng.546

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Plant Science

ORIGINAL RESEARCH
published: 14 August 2018
doi: 10.3389/fpls.2018.01196

Genome-Wide Association Studies of Free Amino Acid Levels by Six Multi-Locus Models in Bread Wheat

Yanchun Peng[1†], Hongbo Liu[2†], Jie Chen[1], Taotao Shi[1], Chi Zhang[3], Dongfa Sun[1], Zhonghu He[4], Yuanfeng Hao[4] and Wei Chen[1]**

[1] College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China, [2] National Key Laboratory of Crop Genetic Improvement, National Center of Plant Gene Research, Huazhong Agricultural University, Wuhan, China, [3] School of Chemical Science and Engineering, Royal Institute of Technology, Stockholm, Sweden, [4] Institute of Crop Science, National Wheat Improvement Center, Chinese Academy of Agricultural Sciences, Beijing, China

OPEN ACCESS

Edited by:
Zhenyu Jia,
University of California, Riverside,
United States

Reviewed by:
Jin-Ying Gou,
Fudan University, China
Yang Zhang,
Sichuan University, China

****Correspondence:***
Yuanfeng Hao
haoyuanfeng@caas.cn
Wei Chen
chenwei0609@mail.hzau.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: *27 May 2018*
Accepted: *26 July 2018*
Published: *14 August 2018*

Citation:
Peng Y, Liu H, Chen J, Shi T, Zhang C, Sun D, He Z, Hao Y and Chen W (2018) Genome-Wide Association Studies of Free Amino Acid Levels by Six Multi-Locus Models in Bread Wheat. Front. Plant Sci. 9:1196. doi: 10.3389/fpls.2018.01196

Genome-wide association studies (GWAS) have been widely used to dissect the complex biosynthetic processes of plant metabolome. Most studies have used single-locus GWAS approaches, such as mixed linear model (MLM), and little is known about more efficient algorithms to implement multi-locus GWAS. Here, we report a comprehensive GWAS of 20 free amino acid (FAA) levels in kernels of bread wheat (*Triticum aestivum* L.) based on 14,646 SNPs by six multi-locus models (FASTmrEMMA, FASTmrMLM, ISISEM-BLASSO, mrMLM, pKWmEB, and pLARmEB). Our results showed that 328 significant quantitative trait nucleotides (QTNs) were identified in total (38, 8, 92, 45, 117, and 28, respectively, for the above six models). Among them, 66 were repeatedly detected by more than two models, and 155 QTNs appeared only in one model, indicating the reliability and complementarity of these models. We also found that the number of significant QTNs for different FAAs varied from 8 to 41, which revealed the complexity of the genetic regulation of metabolism, and further demonstrated the necessity of the multi-locus GWAS. Around these significant QTNs, 15 candidate genes were found to be involved in FAA biosynthesis, and one candidate gene (*TraesCS1D01G052500*, annotated as tryptophan decarboxylase) was functionally identified to influence the content of tryptamine *in vitro*. Our study demonstrated the power and efficiency of multi-locus GWAS models in crop metabolome research and provided new insights into understanding FAA biosynthesis in wheat.

Keywords: wheat, free amino acid (FAA), genome-wide association studies, multi-locus models, QTNs

INTRODUCTION

Genome-wide association studies (GWAS) have largely been applied to the genetic dissection of complex traits in plants. With the landmark GWAS study of 107 phenotypes in *Arabidopsis* (Atwell et al., 2010), numerous other studies have been successfully performed, including those addressing the flowering time and grain yield in rice (Huang et al., 2012; Yang W. et al., 2014), salinity tolerance in barley (Fan et al., 2016), male inflorescence size in maize (Wu et al., 2016), floret fertility in wheat (Guo et al., 2017), and the reducing levels of cucurbitacin in cucumber domestication (Shang et al., 2014). Of these studies, the mixed linear model (MLM) has been adopted most frequently owing to its effective control of spurious associations (Yu et al., 2006). However, as a single-locus GWAS approach, MLM leads to missing some significant loci because of the conservative Bonferroni correction ($0.05/m_e$, where m_e is the number of effective markers) and the stringent criterion of the significance test (Wang et al., 2016). To address this issue, several multi-locus models have been developed, such as Bayesian LASSO (Hoggart et al., 2008), ISISEM-BLASSO (Tamba et al., 2017), pLARmEB (Zhang et al., 2017), and pKWmEB (Ren et al., 2018). Because of the multi-locus nature, the obvious superiority of these approaches is that no Bonferroni correction is demanded, hence, a looser significance criterion can be adopted, and more-powerful quantitative trait nucleotides (QTNs) can be detected (Wang et al., 2016).

Plants produce a vast array of metabolites that provide nutrition and medicine for humans (Saito and Matsuda, 2010; Chae et al., 2014). Unraveling the diversity of the plant metabolome and its underlying mechanism has attracted increasing research interest in the past decade (Schwab, 2003; De Luca et al., 2012). Recent research showed that GWAS coupled with metabolome analysis (mGWAS) exhibited great potential to dissect the genetic and biochemical bases of metabolome diversity (Chan et al., 2011; Chen et al., 2014; Wen et al., 2014). Similar to complex traits such as plant height and grain weight, which are usually controlled by several loci with small effects (Huang et al., 2010; Yang W. et al., 2014), the production of plant metabolites is often controlled by pathways composed of multiple genes. For instance, levels of primary metabolites, such as amino acids, fatty acids and saccharides, tend to be controlled by small effects loci (Angelovici et al., 2013; Matsuda et al., 2015). Whereas, in contrast to primary metabolites, the contents of secondary metabolites are always controlled not only by a few major loci with large effects but also by additional numerous loci with small effects (Chan et al., 2010; Riedelsheimer et al., 2012). Although the single-locus mGWAS models have succeeded in identifying a number of genetic variants associated with thousands of metabolites, this methodology ignores the joint effects of multiple genetic markers on metabolites (Chan et al., 2010; Tamba et al., 2017). Therefore, multi-locus models are a valuable alternative method for mGWAS analysis.

Bread wheat or common wheat (*Triticum aestivum* L.) is one of the most important crops worldwide and provides approximately 20% of the energy, protein and dietary fiber consumed for human (Ling et al., 2013). The improvement of kernel quality has been a major target in breeding for a long time (Nelson et al., 2006; Jin et al., 2016). Although the seed amino acids are mainly present as components of storage proteins, free amino acids (FAAs) can contribute significantly to be the contents of limited essential amino acids in wheat kernels (Angelovici et al., 2013). To improve the amino acid compositions, both traditional plant breeding techniques and new biotechnologies can be utilized (Fernie and Schauer, 2009). Recently, with the rapid development of the next-generation sequencing technologies, some key genes influencing FAA concentrations have been identified in rice (Chen et al., 2016), maize (Deng et al., 2017), and *Arabidopsis* (Angelovici et al., 2013) via mGWAS, which showed great potential to accelerate breeding for balanced AA compositions. However, to our knowledge, no studies of dissecting genetic associations with FAA levels in wheat have been reported.

Here, to understand the genetic bases underlying the natural variation and the biosynthesis of FAAs in wheat kernels, we detected the levels of 20 FAAs with an LC-MS platform (Chen et al., 2013) from a highly diverse association panel of 182 accessions. We identified 328 significant QTNs (LOD $>$ 3.0) with six multi-locus mGWAS models and assigned 15 candidate genes involved in FAA biosynthesis. As a proof of concept, we functionally identified *TraesCS1D01G052500 in vitro*. Our study proved the efficiency of multi-locus GWAS models in metabolome research and provided new insights into understanding of FAA biosynthesis in wheat, which may facilitate metabolomics-based breeding for quality improvement.

MATERIALS AND METHODS

Plant Material

A highly diverse association panel of 182 *Triticum aestivum* L. accessions, including both landraces and elite varieties (**Supplementary Table S1**), was described as before (Liu J. et al., 2017). All accessions were grown at Gaoyi in Hebei province and Dezhou in Shandong province during the 2016–2017 cropping season. Field trials were conducted in randomized complete blocks with three replicates at each location. Each plot contained three 2 m rows spaced 20 cm apart. Field trials followed standard agronomic wheat management practice. Ten mature seeds were randomly collected and pooled for metabolic profiling analysis.

Genotyping

Total genomic DNA was extracted from young leaves for SNP arrays. The 182 accessions were genotyped using the Illumina wheat 90 K SNP by Capital Bio Corporation, Beijing, China[1]. Accuracy of SNP clustering was validated visually step by step. Of the 81,587 SNPs, those with minor allele frequencies (MAFs) < 0.05 and missing data $>20\%$ were excluded from further analysis (Liu J. et al., 2017) to avoid spurious MTAs, finally, a total of 14,646 SNPs were employed in the association panel for GWAS analysis (Dong et al., 2016). The physical positions of SNPs were obtained from the International Wheat Genome Sequencing Consortium website (IWGSC)[2].

Determination of AA Levels

A widely targeted metabolomic platform was applied to quantify the FAA contents in mature wheat kernel samples as described previously (Chen et al., 2013). The dried kernels were crushed using a mixer mill (MM 400, Retsch) for 1.2 min at 29 Hz. Then, 100 mg powder was weighted and extracted for 8 h at 4°C with 1.0 ml 70% aqueous methanol containing 0.1 mg/l lidocaine (internal standard). Extracts were centrifuged at 10,000 *g* for 10 min, and filtrated (SCAA-104, 0.22 μm pore size; ANPEL,

[1] http://www.capitalbiotech.com/

[2] http://www.wheatgenome.org/

Shanghai, China[3] before LC–MS analysis. The HPLC conditions as follow: column, shim-pack VP-ODS C18; solvent system, water with 0.04% acetic acid and acetonitrile with 0.04% acetic acid; gradient program, 0 min, 100:0 V/V, 20.0 min, 5:95 V/V, 22.0, 5:95 V/V, 22.1, 95:5 V/V, 25.0, 95:5 V/V; flow rate, 0.25 ml min^{-1}; temperature, 40°C; Injection volume, 5 μl. The MS parameters as follow: ion spray voltage (IS) 5,500 V; source temperature 500°C; ion source gas I (GSI), gas II (GSII), curtain gas (CUR) were set at 55, 60, and 25.0 psi, respectively, the collision gas (CAD) was high. A specific set of MRM (multiple reaction monitoring) transitions were monitored for each FAA (**Supplementary Table S2**), each MRM transition was obtained with a 5 ms pause time and 5 ms Dwell time, data were processed by Analyst 1.5.1 software, peak areas were integrated using a IntelliQuan algorithm. Endogenous concentrations of FAAs were quantified by calculating the peak area in comparison to standard curves obtained from authenticated standards (purchased from Sigma-Aldrich). Calibration curves were drawn by plotting at least four different concentrations of each FAA standard according to the peak area (Dong et al., 2014). Finally, to eliminate environmental effects, BLUPs (best linear unbiased predictor) across two environments were used as the phenotypic values for all subsequent analyses (Liu J. et al., 2017).

GWAS Mapping

Free amino acid levels were simultaneously studied with a single-locus GWAS model (MLM) and six multi-locus GWAS models. The single-locus model was implemented by FaST-LMM program (Lippert et al., 2011), while multi-locus models were implemented by mrMLM (Wang et al., 2016), FASTmrMLM (Tamba, 2017), FASTmrEMMA (Wen et al., 2017), ISISEM-BLASSO (Tamba et al., 2017), pLARmEB (Zhang et al., 2017), and pKWmEB (Ren et al., 2018). The critical threshold for significantly associated SNPs was set at $LOD > 3.0$ for the six multi-locus models, and $P = 0.05/14{,}646 = 3.41 \times 10^{-6}$ (or $-\log_{10} P -$ value = 5.5, Bonferroni correction) for MLM.

Statistical Analysis

We used $s/\bar{y} \times 100$ to calculate the values of coefficient variation (CV, %) for each FAA, where s and $\bar{y}$ are the standard deviation (SD) and the mean of each FAA in the population, respectively. Spearman's rank correlation coefficient was used to calculate the correlation between each pair of FAAs, and statistical significance was obtained by using Student's t-test.

In vitro Validation of Candidate Genes

Full-length cDNA of *TraesCS1D01G052500* was amplified with the primer using cDNA from Huaimai20 as a template. Clones were digested with *Bam*H I/*Eco*R I and directionally ligated to the pre-digested *pGEX-6p-1* vector. Error-free recombinant proteins were expressed in BL-21 (DE3) competent cells after induced by adding 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and growing continually for 12 h at 16°C. Cells were harvested and suspended in the lysis buffer [contains 500 mM NaCl, 50 mM Tris-HCl (pH 8.0), 10% glycerol, 5 mM β-mercaptoethanol and 1 mM PMSF] and lysed by high pressure. The crude extract was collected and clarified by centrifugation at 14,000 *g* for 1 h at 4°C, and the supernatant was stored at −80°C for future experiments.

The standard *in vitro* enzyme assay for the role of TraesCS1D01G052500 (tryptophan as substrate) was performed in a total volume of 20 μl containing 100 ppm PLP and 50 μM substrate in 50 mM Tris-HCl buffer (pH 8.0). After incubating at 37°C for 30 min, the reaction was stopped by adding 60 μl of methanol. The reaction mixture was then filtered through a 0.2 μm filter (Millipore) before being used for LC-MS analysis.

Phylogenetic Analysis of Different Gene Families

We use the CLUSTALW (version 1.83) program to align the amino acid sequences and construct the neighbor-joining tree by MEGA5. Bootstrap values from 1,000 times are indicated at each node. Bar = 0.1 amino acid substitutions per site.

Enzyme Kinetics

To determine the kinetic difference between TraesCS1D01G052500 and its homologs in rice (OsTDC1 and OsTDC3), their activities were measured using 50 ng of purified protein expressed from *E. coli*, with 10–1,250 μM different tryptophan (Sigma) as substrates and a fixed concentration of 50 ppm PLP (Sigma) as co-factor. The kinetic parameters were calculated using Michaelis–Menten model (SigmaPlot software, version 14.0). All reactions were run in duplicate and repeated twice.

RESULTS

Natural Variation of Free Amino Acids in Wheat Kernel

To assess the phenotypic variation for FAAs in dry, mature wheat kernels, the absolute levels of 20 FAAs (alanine, arginine, asparagine, aspartic acid, glutamic acid, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, serotonin, threonine, tryptamine, tryptophan, tyramine, tyrosine, and valine in nmol/mg dry wheat kernels) were quantified using LC-MS/MS as previously described (Chen et al., 2013). Visualization of the FAA profiling was performed by hierarchical cluster analysis (HCA), and accumulation of FAAs displayed a distinct phenotypic variation according to their abundance (**Figure 1**). Aspartic acid, glutamic acid, alanine and serine were the most highly abundant FAAs, with average concentrations of 0.37, 0.31, 0.30, 0.30 nmol/mg, respectively, while tyramine, threonine, and tryptamine were the less abundant, with average concentrations of 0.005, 0.02, 0.03 nmol/mg, respectively (**Supplementary Table S2**). The content of each FAA varied widely within the association panel, with variation ranging from a 2.30-fold difference in tyrosine to a 30.36-fold difference in proline and with the genetic coefficient variation (CV, %) ranging

[3] www.anpel.com.cn/

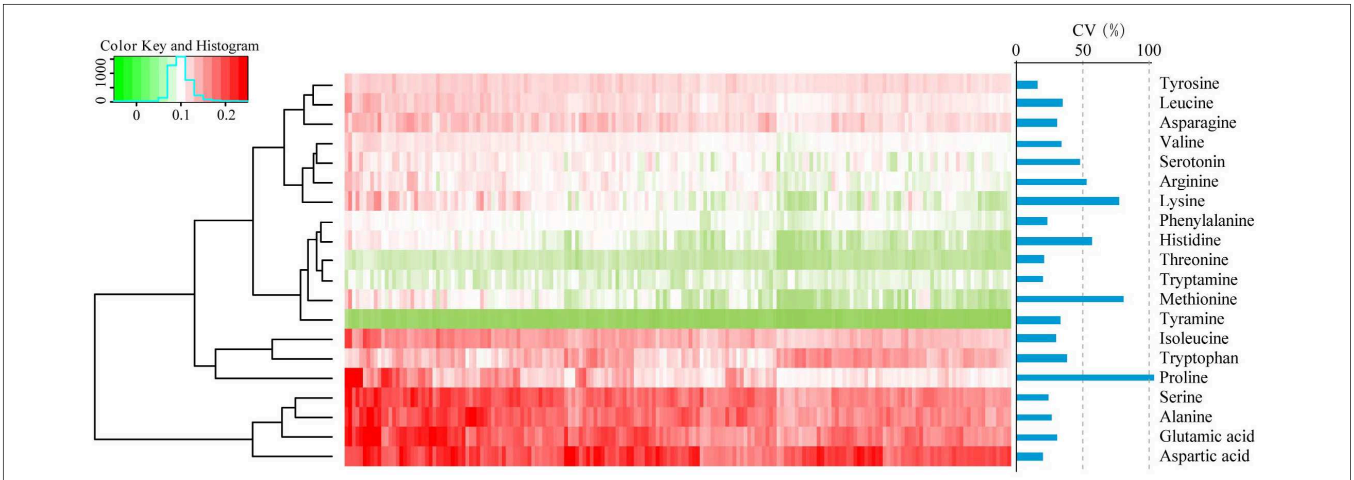


FIGURE 1 | Hierarchical cluster analysis (HCA) and the coefficient variation (CV, %) of the levels of FAAs in 182 wheat accessions. Each accession is visualized in a single column, and each FAA is represented by a single row. Red indicates high level, whereas low FAA contents are shown in green.

TABLE 1 | Summary of significant QTNs identified by different models.

Model	FASTmrEMMA	FASTmrMLM	ISISEM-BLASSO	mrMLM	pKWmEB	pLARmEB	MLM
Number of traits with significant QTNs	18	6	20	15	19	11	4
Number of QTNs	38	8	92	45	117	28	4
Average QTNs per trait	2.1	1.3	4.6	3.0	6.2	2.5	1

from 15.9 to 103.2, respectively (**Figure 1** and **Supplementary Table S2**). The relationships between 20 FAA values were evaluated by Spearman's rank correlation, and strong positive correlations were identified between most of these FAAs, with the exceptions of tryptamine and tryptophan (**Supplementary Table S3**).

Associated Loci Mapped by Different Models

To dissect the genetic basis of natural variation for FAA levels in mature wheat kernels, GWAS was performed using seven different models simultaneously. In total, 328 significant QTNs were identified by six multi-locus models (FASTmrEMMA, FASTmrMLM, ISISEM-BLASSO, mrMLM, pKWmEB, and pLARmEB) at a critical threshold of $LOD > 3.0$ (**Supplementary Table S4**), and the numbers of QTNs for the above six models were 38, 8, 92, 45, 117, and 28 (**Table 1**), respectively. Of these QTNs, 66 were detected by at least two different models; some QTNs, such as the association between lysine and SNP BS00003585_51 on chromosome 2B (747,603,047 bp), were simultaneously mapped by five different models (**Supplementary Table S4**). Only four significant SNP-trait associations were identified by the single-locus model (MLM) (**Table 1**), and could be also detected by some multi-locus models. Although 18 FAAs were found by FASTmrEMMA to be significantly associated with QTNs, the total number of QTNs is only 38, with an average of 2.1 QTNs per FAA. Comparatively, for the pKWmEB and ISISEM-BLASSO models, the average QTNs per trait reached 6.2 and 4.6, respectively (**Table 1**). The phenotypic variation explained by different loci varied from 0.1% (tyramine in pKWmEB) to 21.4% (aspartic acid in mrMLM), with an average of 5.6%. We also found that the same QTN shows different effects to explain the phenotypic variation in different models; for instance, the association between arginine and SNP BS00022811_51 on chromosome 7A (709,639,589 bp) with the r^2 ranged from 0.1% in FASTmrEMMA to 19.7% in pKWmEB (**Supplementary Table S4**).

The number of significant QTNs also varied widely among different FAAs, ranging from 8 for tryptophan to 41 for tyramine (**Figure 2**), indicating the complex genetic regulation of FAAs. The chromosomal distribution of all identified QTNs revealed that A genome had the greatest number of significant associations, while only few QTNs were detected in the D genome (**Figure 2**). Since QTNs were not distributed evenly on the chromosomes (Deng et al., 2017), five QTN hotspots were observed on chromosomes 2A, 4A, 6A, 7A, and 7B, with the most obvious one being that more than 18 QTNs can be detected between 7 FAAs and SNP RAC875_c1022_3059 (located at 595,984,457 bp on chromosome 4A) (**Figure 2** and **Supplementary Table S4**). The candidate genes underlying these QTN hotspots could include transcriptional factors, transporters or some other rate-limiting enzymes of the amino acid metabolic pathway.

Candidate Genes Underlying QTNs

Notably, the 328 significantly QTNs facilitated the assignment of candidate genes. To identify them, the flanking sequences corresponding to the SNP markers significantly associated with

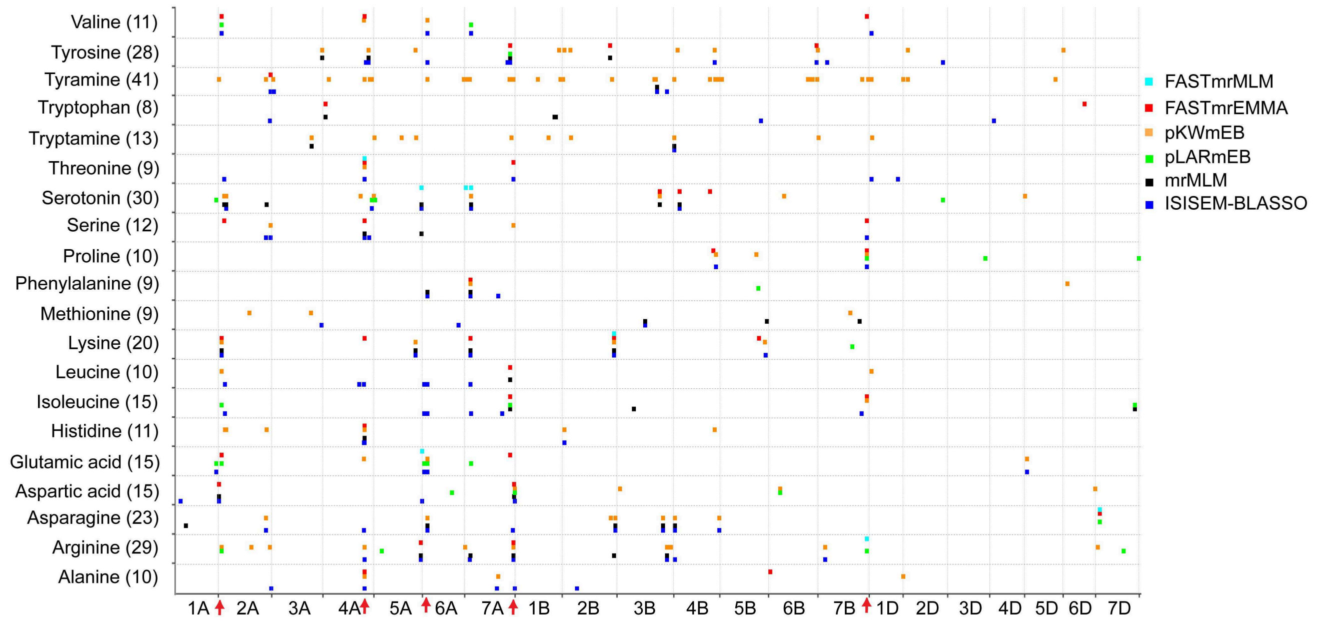


FIGURE 2 | Chromosomal distribution of QTNs identified in this study. The *x*-axis indicates genomic locations by chromosomal order, and the significant QTNs are plotted against genome location. Each row represents one QTN identified by a different model. The red arrows show the QTN hotspots.

TABLE 2 | Summary of 15 candidate genes significantly associated with FAA levels.

Traits	Chr	Lead SNP position (bp)	r^2 (%)[a]	LOD	Candidate gene[b]	Annotation
Glutamic acid	1A	558,490,011	6.9	3.5	TraesCS1A01G390300	Glutamate receptor
Alanine	4A	595,984,457	15.6	5.7	TraesCS4A01G294100	Aminopeptidase
Asparagine	4B	14,124,082	6.1	4.1	TraesCS4B01G020000	Aminopeptidase
Tryptamine	1D	34,621,416	6.7	3.8	TraesCS1D01G052500	Tryptophan decarboxylase
Tyramine	3B	543,718,678	1.9	5.5	TraesCS3B01G340000	Tyrosine decarboxylase
Glutamic acid	5D	31,273,563	10.1	4.2	TraesCS5D01G031800	Amino acid transporter
Isoleucine	7A	660,464,837	15.9	5.7	TraesCS7A01G464900	Amino acid transporter
Tyrosine	2B	689,871,912	8.3	4.5	TraesCS2B01G493000	Amino acid permease
Arginine	7B	105,558,975	4.4	3.9	TraesCS7B01G093200	Amino acid permease
Methionine	3B	408,354,812	14.4	4.3	TraesCS3B01G253600	Amino acid transporter
Valine	4A	593,337,515	9.6	3.6	TraesCS4A01G287900	Peptide transporter
Tyramine	3B	582,466,573	17.0	4.9	TraesCS3B01G369800	Aminotransferase
Lysine	2A	41,237,242	9.2	4.5	TraesCS2A01G088600	Pyruvate decarboxylase
Histidine	2B	26,581,220	16.1	7.3	TraesCS2B01G053600	Pyruvate dehydrogenase
Lysine	2B	747,603,047	9.4	5.8	TraesCS2B01G553300	Shikimate kinase

[a] *The phenotypic variance explained by the corresponding locus.* [b] *A possible biological candidate gene in the locus or the nearest annotated gene to the lead SNP. More information is listed in* ***Supplementary Table S4****.*

FAA levels were used in BLASTx search against NCBI database[4]. In most cases, the chemical structure combining with the existing knowledge of the biosynthetic pathway of the amino acids allowed the tentative assignment of a protein sequence that is biochemically related to the associated FAAs. Notably, 15 candidate genes involved in FAAs anabolism or catabolism were identified by mGWAS in this study (**Table 2**), based on the wheat reference genome information (see footnote 2).

A significant QTN between the levels of glutamic acid and the SNP Excalibur_c35310_375 was identified on chromosome 1A; this SNP is located 0.5 Mb away from *TraesCS1A01G390300* (encoding a putative glutamate receptor). The high homology (58% identity at amino acid level) between *TraesCS1A01G390300* and the glutamate receptor gene *AtGLR3.5* (Teardo et al., 2015) suggests that *TraesCS1A01G390300* is likely the candidate gene underlying this locus. The SNP RAC875_c1022_3059 was significantly associated with 7 FAAs (**Supplementary Table S4**), which is comprised a hotspot on chromosome 4A as mentioned

[4] http://www.ncbi.nlm.nih.gov/

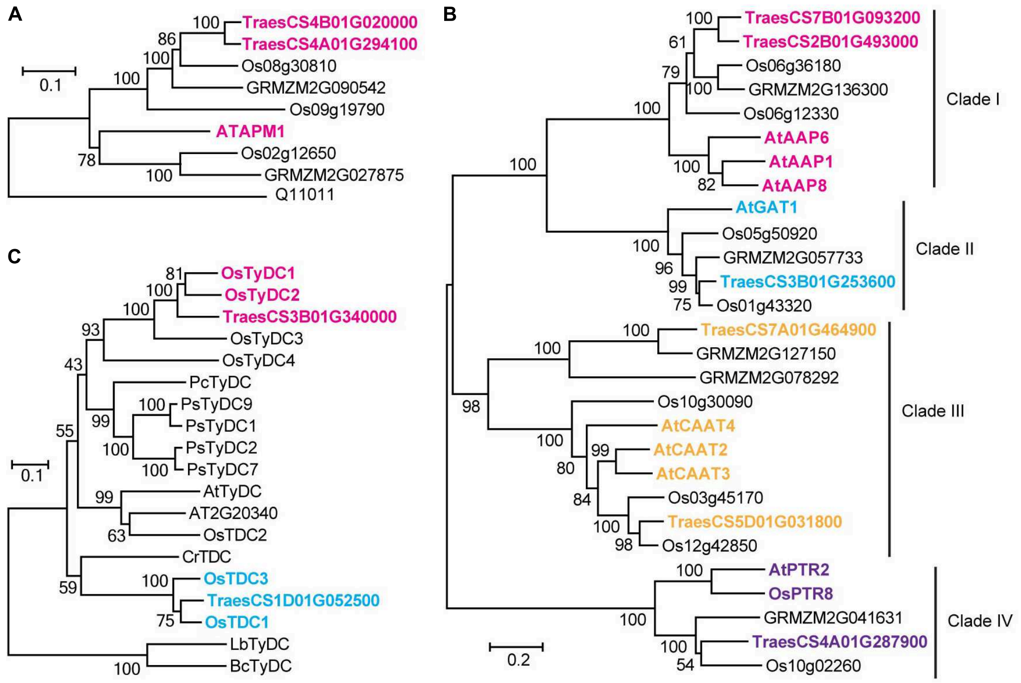


FIGURE 3 | Homologous amino acid sequences of aminopeptidase gene family **(A)**, tyrosine decarboxylase and tryptophan decarboxylase gene families **(B)**, and amino acid permease, amino acid transporter and peptide transporter gene families **(C)** from multiple species were collected and aligned. The neighbor-joining trees were constructed using MEGA software and tested using bootstrap method at replication number of 1000. Phylogenetic analysis of different gene families assigned in the study. Os, *Oryza sativa*; At, *Arabidopsis thaliana*; Pc, *Petroselinum crispum*; Ps, *Papaver somniferum*; Cr, *Catharanthus roseus*; Lb, *Lactobacillus brevis*; Bc, *Bacillus cereus*.

above. The high sequence identity (61% at the amino acid level) between adjacently located gene *TraesCS4A01G294100* (0.4 Mb to SNP RAC875_c1022_3059) and *AtAPM1* (Murphy et al., 2002), an aminopeptidase in *Arabidopsis*, suggests that *TraesCS4A01G294100* is likely the candidate gene underlying this QTN. Similarly, *TraesCS4B01G020000* (also encoding a putative aminopeptidase), was assigned as the candidate gene underlying the content of asparagine. The associations were further supported by phylogenetic analysis (**Figure 3A**).

Levels of tryptamine were significantly associated (LOD = 3.8) with the SNP BS00012936_51 on chromosome 1D that is 1.0 Mb away from *TraesCS1D01G052500*, which encodes a protein annotated as tryptophan decarboxylase, suggesting that TraesCS1D01G052500 catalyzes the key step of tryptamine biosynthesis. Similarly, *TraesCS3B01G340000* (encoding a putative tyrosine decarboxylase) was assigned as the candidate gene underlying the levels of tyramine. The high sequence identities between *TraesCS1D01G052500* and *OsTDC1* (88% at the amino acid level, Kanjanaphachoat et al., 2012), *TraesCS3B01G340000* and *OsTyDC2* (79% at the amino acid level, Kang et al., 2007) further supported the realness of these QTNs (**Figure 3B**).

Six candidate genes putatively annotated as amino acid transporters (AATs) or amino acid permeases (AAPs) were identified by mGWAS (**Table 2**). We investigated the phylogenetic relationships among the AATs (or AAPs) by constructing the phylogenetic tree with a neighbor-joining algorithm based on the amino acid sequences of these candidate genes and a collection of nine reported genes (Dietrich et al., 2004; Hirner et al., 2006; Meyer et al., 2006; Lee et al., 2007; Yang H. et al., 2014; Santiago and Tegeder, 2016). As a result, characterized AATs (or AAPs) were sorted into four major clades (**Figure 3C**). Closer examination of the phylogeny in clade III reveled that *TraesCS5D01G031800* lies next to *AtCAAT2*, *AtCAAT3*, and *AtCAAT4*, three cationic amino acid transporters from *Arabidopsis* (Yang H. et al., 2014), consistent with the significant QTN between the levels of glutamic acid (a typical cationic amino acid) and *TraesCS5D01G031800* locus (**Figure 3C** and **Supplementary Table S4**). Our analysis also placed *TraesCS2B01G493000* and *TraesCS7B01G093200* close to *AtAAP1*, *AtAAP6*, and *AtAAP8* (Hirner et al., 2006; Lee et al., 2007; Santiago and Tegeder, 2016) within clade I, strongly supporting the annotation of these candidates as AAPs in wheat (**Figure 3C**). Moreover, the high sequence identities between *TraesCS3B01G253600* and *AtGAT1* (63% at the amino acid level, Meyer et al., 2006), *TraesCS4A01G287900* and *AtPTR2* (44% at the amino acid level, Dietrich et al., 2004) provide further evidence for these assignments (**Figure 3C**).

Functional Identification of Candidate Genes

Although experimental validation of all candidate genes disclosed by our mGWAS analyses is beyond the scope of a single study, we nevertheless tried to show that such confirmation is possible. For

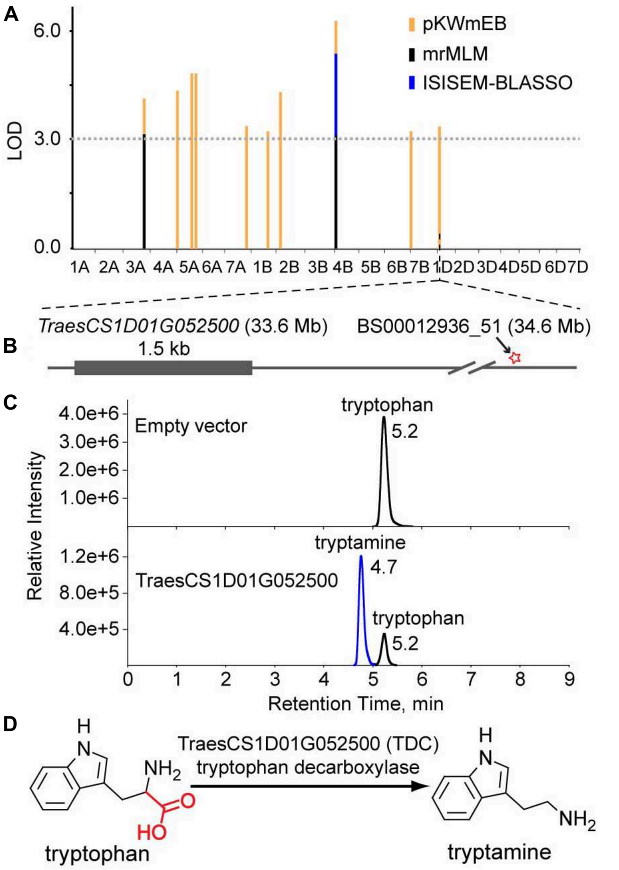


FIGURE 4 | Functional identification of *TraesCS1D01G052500 in vitro*. **(A)** The multi-locus GWAS results for the tryptamine level in different models. **(B)** Gene model of *TraesCS1D01G052500*. The filled gray box represents coding sequence, and the star represents the associated site. **(C)** LC-MS/MS chromatograms of *in vitro* enzyme assays showing the enzyme activity of recombinant TraesCS1D01G052500 (Down). Protein extract from *E. coli* containing empty vector were used as a negative control (Up). **(D)** The proposed pathway of tryptamine biosynthesis in wheat.

this purpose, we further characterized one candidate gene and provided novel biochemical insight into the FAA biosynthesis in wheat.

As mention above, the association between *TraesCS1D01G052500* and tryptamine levels suggests that TraesCS1D01G052500 is the decarboxylase that catalyzes the biosynthesis of tryptamine (**Figures 3B**, **4A,B**). To characterize the enzymatic properties of TraesCS1D01G052500, recombinant protein was expressed with an *N*-terminal glutathione *S*-transferase (GST) tag in *E. coli* BL-21 and the reaction product was confirmed by commercial standard with LC-MS (**Figure 4C**). An obvious TDC activity showed for tryptophan, and its activity was not inhibited by tyrosine, indicating a high level of substrate specificity toward tryptophan (**Supplementary Table S5**). We further investigated the enzyme kinetics of TraesCS1D01G052500 and its rice homologs (OsTDC1 and OsTDC3), all of them displayed similar K_{cat} values for tryptophan (**Supplementary Table S5**), suggesting that the three proteins have similar TDC activities. Based on these results, we functionally identified TraesCS1D01G052500 as a decarboxylase that catalyzes the biosynthesis of tryptamine from tryptophan in wheat (**Figure 4D**), which further confirmed the correctness of our GWAS results and the candidate gene assignment.

DISCUSSION

By coupling with the rapid development of LC-MS strategies, more accurate contents of metabolites can be obtained, and larger phenotypic variation can be observed (Chen et al., 2014). In this study, most of the FAAs varied widely across the association panel, such as proline with range of 30.4-fold (**Supplementary Table S2**), indicating the complexity of the biosynthetic processes of FAAs (**Figure 1**). The levels of lysine (an essential amino acid) have huge phenotypic variation, with a CV (%) of 77.2, implying the existence of a large number of alleles with high genetic diversity in the wheat germplasms (Liu Y. et al., 2017). Thus, identification of the favorable alleles and dissection of the genetic architecture underlying the levels of FAA is beneficial for improving the amino acid compositions in the future.

Dissecting the natural variation and the underlying genetic bases of metabolism is essential for the improvement of crop nutritional quality (Luo, 2015). Due to recent advances in both high-throughput metabolic profiling and sequencing technologies, mGWAS has been employed as a powerful strategy to reveal the genetic and biochemical basis of crop metabolism (Riedelsheimer et al., 2012; Wen et al., 2014; Matsuda et al., 2015). So far, most of these studies have been carried out on maize and rice. What's more important, hundreds of significant loci were identified for various metabolites of nutritional importance, both of large effects and at high resolution, which facilitated the identification of the candidate genes (Luo, 2015). Advanced in developing the genomic toolbox (Jia et al., 2013; Ling et al., 2013; Avni et al., 2017), Matros et al. (2017) quantified 76 leaf metabolites from 135 winter wheat lines and identified several significant associations for six metabolic traits based on 17,372 SNP markers. This confirmed the potential of the mGWAS approach and provided the opportunity for a further understanding of metabolic diversity in wheat. In our study, we also mapped hundreds of QTNs for the levels of 20 FAAs in a wheat diverse association panel, however, most of them had very small effects, explaining the phenotypic variation with an average of 5.6% (**Supplementary Table S4**). Obviously, the limitations of mGWAS in wheat relate in part to the large size of the genome and in part to the limited availability of sets of genetic markers (Zhou et al., 2018), which leads to great difficulties to confirm the candidate genes. These constraints could be gradually complemented by applying new sequencing technologies and developing additional genomic markers (Liu Y. et al., 2017), and also, utilizing larger number of accessions and choosing more comprehensive choices of germplasms can enhance the power of mGWAS approaches, as demonstrated in rice and maize (Huang et al., 2012; Riedelsheimer et al., 2012).

As usual, variation of primary metabolites tends to be controlled by many small-effect loci. To increase the detection power of mGWAS, six multi-locus models were applied in this study. Totally, 328 significant QTNs were identified, however, only 4 SNP-trait associations were found with the single-locus model (MLM) at $P \leq 3.41 \times 10^{-6}$ (**Table 1** and **Supplementary Table S3**). These results indicated the power of these multi-locus methods. Furthermore, the common QTNs appeared in different models confirming the credibility of these multi-locus GWAS approaches.

Based on these QTNs identified by the six multi-locus methodologies, candidates that have not been identified previously can be explored by searching for a protein or protein cluster that is biochemically related to the associated FAAs encoded at these loci. As a result, our mGWAS has allowed the assignment of 15 candidate genes underlying FAA levels (**Table 2**). The existing knowledge of plant FAA pathways, the high sequence identities between them and known functions in rice and *Arabidopsis* further confirmed these candidate genes. Notably, the validation of *TraesCS1D01G052500* was detected only by the pKWmEB model (**Figure 4**), further demonstrating the reliability and effectiveness of these multi-locus methods.

CONCLUSION

In this study, a comprehensive GWAS of 20 FAA levels based on 14,646 SNPs in bread wheat was performed by six multi-locus models. Among 328 significant QTNs, 66 were detected by at least two models, and 155 QTNs appeared only in one model. Fifteen candidate genes were assigned to FAA biosynthesis, and one candidate gene was functionally identified *in vitro*. This study proved the power and reliability of multi-locus GWAS models in plant metabolome research and provided new insights into understanding FAA biosynthesis in wheat, which may facilitate metabolomics-based breeding for quality improvement.

AUTHOR CONTRIBUTIONS

WC, YH, and ZH conceived the project and supervised this study. YP, HL, and JC performed most of the experiments. TS, DS, and CZ participated in preparation of the materials. WC and YP analyzed the data. WC wrote the paper. All the authors discussed the results and commented on the manuscript.

FUNDING

This work was supported by Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (Program No. 2017RC006), National Key Research and Development Program of China (2016YFE0108600 and 2016YFD0101802).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01196/full#supplementary-material

REFERENCES

Angelovici, R., Lipka, A. E., Deason, N., Gonzalez-Jorge, S., Lin, H., Cepela, J., et al. (2013). Genome-wide analysis of branched-chain amino acid levels in *Arabidopsis* seeds. *Plant Cell* 25, 4827–4843. doi: 10.1105/tpc.113.119370

Atwell, S., Huang, Y. S., Vilhjalmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465, 627–631. doi: 10.1038/nature08800

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., et al. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* 357, 93–97. doi: 10.1126/science.aan0032

Chae, L., Kim, T., Nilo-Poyanco, R., and Rhee, S. Y. (2014). Genomic signatures of specialized metabolism in plants. *Science* 344, 510–513. doi: 10.1126/science.1252076

Chan, E. K., Rowe, H. C., Corwin, J. A., Joseph, B., and Kliebenstein, D. J. (2011). Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. *PLoS Biol.* 9:e1001125. doi: 10.1371/journal.pbio.1001125

Chan, E. K., Rowe, H. C., Hansen, B. G., and Kliebenstein, D. J. (2010). The complex genetic architecture of the metabolome. *PLoS Genet.* 6:e1001198. doi: 10.1371/journal.pgen.1001198

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46, 714–721. doi: 10.1038/ng.3007

Chen, W., Gong, L., Guo, Z., Wang, W., Zhang, H., Liu, X., et al. (2013). A novel integrated method for large-scale detection, identification, and quantification of widely targeted metabolites: application in the study of rice metabolomics. *Mol. Plant* 6, 1769–1780. doi: 10.1093/mp/sst080

Chen, W., Wang, W., Peng, M., Gong, L., Gao, Y., Wan, J., et al. (2016). Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nat. Commun.* 7:12767. doi: 10.1038/ncomms12767

De Luca, V., Salim, V., Atsumi, S. M., and Yu, F. (2012). Mining the biodiversity of plants: a revolution in the making. *Science* 336, 1658–1661. doi: 10.1126/science.1217410

Deng, M., Li, D., Luo, J., Xiao, Y., Liu, H., Pan, Q., et al. (2017). The genetic architecture of amino acids dissection by association and linkage analysis in maize. *Plant Biotechnol. J.* 15, 1250–1263. doi: 10.1111/pbi.12712

Dietrich, D., Hammes, U., Thor, K., Suter-Grotemeyer, M., Fluckiger, R., Slusarenko, A. J., et al. (2004). AtPTR1, a plasma membrane peptide transporter expressed during seed germination and in vascular tissue of *Arabidopsis*. *Plant J.* 40, 488–499. doi: 10.1111/j.1365-313X.2004.02224.x

Dong, X., Chen, W., Wang, W., Zhang, H., Liu, X., and Luo, J. (2014). Comprehensive profiling and natural variation of flavonoids in rice. *J. Integr. Plant Biol.* 56, 876–886. doi: 10.1111/jipb.12204

Dong, Y., Liu, J., Zhang, Y., Geng, H., Rasheed, A., Xiao, Y., et al. (2016). Genome-wide association of stem water soluble carbohydrates in bread wheat. *PLoS One* 11:e0164293. doi: 10.1371/journal.pone.0164293

Fan, Y., Zhou, G., Shabala, S., Chen, Z., Cai, S., Li, C., et al. (2016). Genome-wide association study reveals a new QTL for salinity tolerance in barley (*Hordeum vulgare* L.). *Front. Plant Sci.* 7:946. doi: 10.3389/fpls.2016.00946

Fernie, A. R., and Schauer, N. (2009). Metabolomics-assisted breeding: a viable option for crop improvement? *Trends Genet.* 25, 39–48. doi: 10.1016/j.tig.2008.10.010

Guo, Z., Chen, D., Alqudah, A. M., Roder, M. S., Ganal, M. W., and Schnurbusch, T. (2017). Genome-wide association analyses of 54 traits identified multiple loci for the determination of floret fertility in wheat. *New Phytol.* 214, 257–270. doi: 10.1111/nph.14342

Hirner, A., Ladwig, F., Stransky, H., Okumoto, S., Keinath, M., Harms, A., et al. (2006). *Arabidopsis* LHT1 is a high-affinity transporter for cellular amino acid uptake in both root epidermis and leaf mesophyll. *Plant Cell* 18, 1931–1946. doi: 10.1105/tpc.106.041012

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* 4:e1000130. doi: 10.1371/journal.pgen.1000130

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967. doi: 10.1038/ng.695

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., et al. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 44, 32–39. doi: 10.1038/ng.1018

Jia, J. Z., Zhao, S. C., Kong, X. Y., Li, Y. R., Zhao, G. Y., He, W. M., et al. (2013). *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* 496, 91–95. doi: 10.1038/nature12028

Jin, H., Wen, W., Liu, J., Zhai, S., Zhang, Y., Yan, J., et al. (2016). Genome-wide QTL mapping for wheat processing quality parameters in a gaocheng 8901/zhoumai 16 recombinant inbred line population. *Front. Plant Sci.* 7:1032. doi: 10.3389/fpls.2016.01032

Kang, S., Kang, K., Lee, K., and Back, K. (2007). Characterization of rice tryptophan decarboxylases and their direct involvement in serotonin biosynthesis in transgenic rice. *Planta* 227, 263–272. doi: 10.1007/s00425-007-0614-z

Kanjanaphachoat, P., Wei, B., Lo, S., Wang, I., Wang, C., Yu, S., et al. (2012). Serotonin accumulation in transgenic rice by over-expressing tryptophan decarboxylase results in a dark brown phenotype and stunted growth. *Plant Mol. Biol.* 78, 525–543. doi: 10.1007/s11103-012-9882-5

Lee, Y. H., Foster, J., Chen, J., Voll, L. M., Weber, A. P., and Tegeder, M. (2007). AAP1 transports uncharged amino acids into roots of *Arabidopsis*. *Plant J.* 50, 305–319. doi: 10.1111/j.1365-313X.2007.03045.x

Ling, H., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., et al. (2013). Draft genome of the wheat a-genome progenitor *Triticum urartu*. *Nature* 496, 87–90. doi: 10.1038/nature11997

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8, 833–835. doi: 10.1038/Nmeth.1681

Liu, J., He, Z., Rasheed, A., Wen, W., Yan, J., Zhang, P., et al. (2017). Genome-wide association mapping of black point reaction in common wheat (*Triticum aestivum* L.). *BMC Plant Biol.* 17:220. doi: 10.1186/s12870-017-1167-3

Liu, Y., Lin, Y., Gao, S., Li, Z., Ma, J., Deng, M., et al. (2017). A genome-wide association study of 23 agronomic traits in Chinese wheat landraces. *Plant J.* 91, 861–873. doi: 10.1111/tpj.13614

Luo, J. (2015). Metabolite-based genome-wide association studies in plants. *Curr. Opin. Plant Biol.* 24, 31–38. doi: 10.1016/j.pbi.2015.01.006

Matros, A., Liu, G., Hartmann, A., Jiang, Y., Zhao, Y., Wang, H., et al. (2017). Genome-metabolite associations revealed low heritability, high genetic complexity, and causal relations for leaf metabolites in winter wheat (*Triticum aestivum*). *J. Exp. Bot.* 68, 415–428. doi: 10.1093/jxb/erw441

Matsuda, F., Nakabayashi, R., Yang, Z., Okazaki, Y., Yonemaru, J., Ebana, K., et al. (2015). Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. *Plant J.* 81, 13–23. doi: 10.1111/tpj.12681

Meyer, A., Eskandari, S., Grallath, S., and Rentsch, D. (2006). AtGAT1, a high affinity transporter for gamma-aminobutyric acid in *Arabidopsis thaliana*. *J. Biol. Chem.* 281, 7197–7204. doi: 10.1074/jbc.M510766200

Murphy, A. S., Hoogner, K. R., Peer, W. A., and Taiz, L. (2002). Identification, purification, and molecular cloning of N-1-naphthylphthalmic acid-binding plasma membrane-associated aminopeptidases from *Arabidopsis*. *Plant Physiol.* 128, 935–950. doi: 10.1104/pp.010519

Nelson, J. C., Andreescu, C., Breseghello, F., Finney, P. L., Gualberto, D. G., Bergman, C. J., et al. (2006). Quantitative trait locus analysis of wheat quality traits. *Euphytica* 149, 145–159. doi: 10.1007/s10681-005-9062-7

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2018). pKWmEB: integration of kruskal-wallis test with empirical bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., et al. (2012). Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8872–8877. doi: 10.1073/pnas.1120813109

Saito, K., and Matsuda, F. (2010). Metabolomics for functional genomics, systems biology, and biotechnology. *Annu. Rev. Plant Biol.* 61, 463–489. doi: 10.1146/annurev.arplant.043008.092035

Santiago, J. P., and Tegeder, M. (2016). Connecting source with sink: the role of *Arabidopsis* AAP8 in phloem loading of amino acids. *Plant Physiol.* 171, 508–521. doi: 10.1104/pp.16.00244

Schwab, W. (2003). Metabolome diversity: too few genes, too many metabolites? *Phytochemistry* 62, 837–849.

Shang, Y., Ma, Y., Zhou, Y., Zhang, H., Duan, L., Chen, H., et al. (2014). Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* 346, 1084–1088. doi: 10.1126/science.1259215

Tamba, C. L. (2017). *A Fast mrMLM Algorithm Improves Statistical Power, Accuracy and Computational Efficiency of Multi-locus Genome-Wide Association Studies.* Doctoral dissertation, Nanjing Agricultural University, Nanjing.

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Teardo, E., Carraretto, L., De Bortoli, S., Costa, A., Behera, S., Wagner, R., et al. (2015). Alternative splicing-mediated targeting of the *Arabidopsis* GLUTAMATE RECEPTOR3.5 to mitochondria affects organelle morphology. *Plant Physiol.* 167, 216–227. doi: 10.1104/pp.114.242602

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., et al. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat. Commun.* 5, 3438. doi: 10.1038/ncomms4438

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* doi: 10.1093/bib/bbw145 [Epub ahead of print].

Wu, X., Li, Y., Shi, Y., Song, Y., Zhang, D., Li, C., et al. (2016). Joint-linkage mapping and GWAS reveal extensive genetic loci that regulate male inflorescence size in maize. *Plant Biotechnol. J.* 14, 1551–1562. doi: 10.1111/pbi.12519

Yang, H., Krebs, M., Stierhof, Y. D., and Ludewig, U. (2014). Characterization of the putative amino acid transporter genes AtCAT2, 3 & 4: the tonoplast localized AtCAT2 regulates soluble leaf amino acids. *J. Plant Physiol.* 171, 594–601. doi: 10.1016/j.jplph.2013.11.012

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., et al. (2014). Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.* 5:5087. doi: 10.1038/ncomms6087

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhou, Y., Chen, Z., Cheng, M., Chen, J., Zhu, T., Wang, R., et al. (2018). Uncovering the dispersion history, adaptive evolution and selection of wheat in China. *Plant Biotechnol. J.* 16, 280–291. doi: 10.1111/pbi.12770

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Plant Science

ORIGINAL RESEARCH
published: 16 August 2018
doi: 10.3389/fpls.2018.01169

Multi-Locus Genome-Wide Association Studies of Fiber-Quality Related Traits in Chinese Early-Maturity Upland Cotton

Junji Su [1,2]*, ***Qi Ma*** [2], ***Mei Li*** [3], ***Fushun Hao*** [4] ***and Caixiang Wang*** [1]*

[1] State Key Laboratory of Cotton Biology, Institute of Cotton Research of CAAS, Anyang, China, [2] Cotton Research Institute, Xinjiang Academy of Agricultural and Reclamation Science, Shihezi, China, [3] College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China, [4] State Key Laboratory of Cotton Biology, Henan Key Laboratory of Plant Stress Biology, College of Life Science, Henan University, Kaifeng, China

OPEN ACCESS

Edited by:
Zhenyu Jia,
University of California, Riverside,
United States

Reviewed by:
Yang-Jun Wen,
College of Sciences, Nanjing
Agricultural University, China
Jia Wen,
University of North Carolina
at Charlotte, United States

***Correspondence:**
Junji Su
xjnkysujj1980@163.com
Caixiang Wang
wangcx8050@163.com

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: 13 June 2018
Accepted: 23 July 2018
Published: 16 August 2018

Citation:
Su J, Ma Q, Li M, Hao F and Wang C
(2018) Multi-Locus Genome-Wide
Association Studies of Fiber-Quality
Related Traits in Chinese
Early-Maturity Upland Cotton.
Front. Plant Sci. 9:1169.
doi: 10.3389/fpls.2018.01169

Early-maturity varieties of upland cotton are becoming increasingly important for farmers to improve their economic benefits through double cropping practices and mechanical harvesting production in China. However, fiber qualities of early-maturing varieties are relatively poor compared with those of middle- and late- maturing ones. Therefore, it is crucial for researchers to elucidate the genetic bases controlling fiber-quality related traits in early-maturity cultivars, and to improve synergistically cotton earliness and fiber quality. Here, multi-locus genome-wide association studies (ML-GWAS) were conducted in a panel consisting of 160 early-maturing cotton accessions. Each accession was genotyped by 72,792 high-quality single nucleotide polymorphisms (SNPs) using specific-locus amplified fragment sequencing (SLAF-seq) approach, and fiber quality-related traits under four environmental conditions were measured. Applying at least three ML-GWAS methods, a total of 70 significant quantitative trait nucleotides (QTNs) were identified to be associated with five objective traits, including fiber length (FL), fiber strength (FS), fiber micronaire (FM), fiber uniformity (FU) and fiber elongation (FE). Among these QTNs, D11_21619830, A05_28352019 and D03_34920546 were found to be significantly associated with FL, FS, and FM, respectively, across at least two environments. Among 96 genes located in the three target genomic regions (A05: 27.95 28.75, D03: 34.52 35.32, and D11: 21.22 22.02 Mbp), six genes (*Gh_A05G2325*, *Gh_A05G2329*, *Gh_A05G2334*, *Gh_D11G1853*, *Gh_D11G1876,* and *Gh_D11G1879*) were detected to be highly expressed in fibers relative to other eight tissues by transcriptome sequencing method in 12 cotton tissues. Together, multiple favorable QTN alleles and six candidate key genes were characterized to regulate fiber development in early-maturity cotton. This will lay a solid foundation for breeding novel cotton varieties with earliness and excellent fiber-quality in the future.

Keywords: upland cotton, fiber quality, early maturity, multi-locus GWAS, candidate genes

INTRODUCTION

Upland cotton (*Gossypium hirsutum* L.), a tetraploid plant, is the most important natural-fiber crop. It is widely cultivated in the world and supplies more than 95% of the global fiber yield due to its extensive adaptive ability and high productivity (Chen et al., 2007). Upland cotton cultivars can be divided into early-, middle- and late- maturity varieties, according to the duration of growth period. Early-maturity (short-season) cotton is an ecological type with a relatively short growing period (Yu et al., 2005; Song et al., 2015). It is suited for wheat-cotton, barley-cotton and rape-cotton double cropping patterns in cotton growing areas of Yellow River Region (YRR) and Yangzi River Region (YZRR), and is also fit for single cropping production in the early-maturity areas of Northwest Inland Region (NIR) and the Northern Specific Early-Maturity Region (NSEMR), with the short frost-free period in China (Yu et al., 2005; Song et al., 2015). Additionally, mechanized harvesting of cotton after good ripening is very common in NIR. The cotton varieties appropriate for mechanical harvesting should have earlier maturing characteristics, especially for early and concentrated boll-opening traits, compared with those suitable for manual harvesting (Bao et al., 2014; Feng et al., 2017). Therefore, the early-maturity upland cotton varieties are becoming more and more important in Chinese cotton production.

Currently, farmers gained increased economic benefits after using new production patterns of double cropping and mechanical harvesting; whereas application of these cultivation measures need early-maturity cotton (Du et al., 2015; Dai et al., 2017; Lu et al., 2017). Owing to their great necessity, a series of early-maturity cotton varieties, such as "Liaomian," "Zhongmiansuo," and "Xinluzao," were developed and released in recent 40 years in China. However, their fiber qualities were relatively poor compared with those of middle- and late- maturity cotton varieties. Therefore, it is crucial to improve fiber quality of early-maturity cotton varieties.

To meet human higher needs for improving textile products, it is also essential for researchers to focus on fiber-quality improvement of early-maturity cotton in future. However, it is difficult to improve fiber quality of early-maturity cotton by means of traditional breeding strategy because of the significant negative correlation between earliness and excellent-quality fiber (Song et al., 2005; Fan et al., 2006). The rapid development of genotyping techniques based on simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers provided an alternative method to improve the efficiency of crop breeding. Generally, marker-assisted selection (MAS) is a high-efficiency and economical approach for modern breeding, compared with the traditional phenotyping breeding (Lande and Thompson, 1990). Researchers have spent a great amount of time and effort on mapping quantitative trait loci (QTL) by using linkage analysis. Over the last two decades, a number of cotton earliness-related QTL have been identified via linkage mapping (Fan et al., 2006; Li et al., 2013; Jia et al., 2016). Compared to the studies evaluating cotton early maturity, far too many investigations have been conducted to identify genetic signatures for fiber quality. A recent meta-QTL analysis suggested that approximately one thousand QTL for fiber-related traits have been detected in intraspecific upland cotton populations (Said et al., 2015), and a few near-term studies have added new QTL for cotton fiber quality (Shang et al., 2015; Tan et al., 2015; Tang et al., 2015; Fang X. et al., 2017).

A genome-wide association study (GWAS) is a wonderful supplement to QTL mapping, and it has been widely used in upland cotton in recent years (Su et al., 2016a, 2018; Fang L. et al., 2017; Huang et al., 2017; Sun et al., 2017; Ma Z. et al., 2018). Although there are a lot of reports on GWAS for cotton earliness and fiber-quality related traits in the past ten years (Zeng et al., 2009; Zhang et al., 2013; Cai et al., 2014; Nie et al., 2016; Su et al., 2016a,b; Sun et al., 2017; Ma Z. et al., 2018), few GWAS investigations have been conducted on fiber-quality related traits in early-maturity upland cotton. In the previous studies, the majority of QTL or quantitative trait nucleotides (QTNs) for fiber quality are mainly derived from germplasms of *G. barbadense* and late-maturity *G. hirsutum*, they are not convenient for use in fiber-quality improvement of early-maturity cotton. Therefore, it is needed to identify QTNs and candidate genes associated with fiber quality in the panel consisting of early-maturity upland cotton accessions.

To date, a lot of single-locus GWAS (SL-GWAS) have been reported in upland cotton (Zeng et al., 2009; Zhang et al., 2013; Nie et al., 2016; Su et al., 2016a,b,c; Sun et al., 2017; Ma Z. et al., 2018). The SL-GWAS methods are involved in multiple testing, and Bonferroni correction is frequently adopted to control the false positive rate. However, this correction is very stringent, thus some important loci cannot be detected, especially for large error in the phenotypic measurement in field experiments (Tamba et al., 2017). To overcome this issue, multi-locus GWAS (ML-GWAS) methodologies have been developed. They include mrMLM (Wang et al., 2016), FASTmrMLM (Tamba and Zhang, 2018), ISIS EM-BLASSO (Tamba et al., 2017), FASTmrEMMA (Wen et al., 2017), pLARmEB (Zhang et al., 2017), and pKWmEB (Ren et al., 2018). Additionally, to decrease the false positive rate, a combination of several ML-GWAS methods have been applied in previous studies (Wu et al., 2016; Misra et al., 2017; Ma L. et al., 2018).

In this study, ML-GWAS for fiber-quality related traits were conducted in a panel composed of 160 early-maturing cotton accessions. The main objective of our study was to discover the favorable QTN allelic variations and some potential candidate genes controlling fiber quality in the early-maturity upland cotton. This investigation will lay a foundation for breeding new cotton varieties with earliness and excellent fiber quality in the future.

MATERIALS AND METHODS

Plant Materials

A natural population consisting of 160 Chinese early-maturity upland cotton accessions were generated (**Table S1**). These accessions were sampled from the germplasm gene bank of the Cotton Research Institute of the Chinese Academy of

Agricultural Sciences (CRI-CAAS). The germplasms fell into three groups based on cotton-planting regions in China. Specifically, 81, 58, and 21 accessions were from the YRR, NIR and NSEMR, respectively. All the accessions have relatively short whole growing period (ranging from 100 to 120 days).

Field Experiments

A collection of 160 early-maturity upland cotton accessions was evaluated under four environmental conditions (2 locations × 2 years): Anyang, Henan, China (36.13°N, 114.80°E) in 2014 and 2015 (designated AY-2014 and AY-2015, respectively), and Shihezi (SHZ), Xinjiang, China (44.52°N, 86.02°E) in 2014 and 2015 (designated SHZ-2014 and SHZ-2015, respectively). The field experiments were arranged in a randomized complete block design with three replications. At AY, each accession was sown in a single-row plot with about 20 plants, while at SHZ, each accession was planted in double-row plots with about 30 plants. The field trials at SHZ were performed with drip irrigation under plastic film conditions, whereas the plots at AY were furrow irrigated as needed. The experimental field management measures were full accordance with local agronomic practices.

Phenotyping and Data Analysis

After mature, a total of 20 naturally opened bolls, as a cotton fiber sample, were handly picked from central part of the plants from each accession in each replicate every year. Fiber samples weighing 10∼15 g lint cotton were then measured for fiber property determination using an HVI-MF 100 instrument (User Technologies, Inc., USTER, Switzerland) at the Cotton Fiber Quality Inspection and Testing Center of the Ministry of Agriculture, Anyang, China. The following fiber-quality related traits were evaluated: 50% fiber span length (FL, mm), fiber strength (FS, cN.tex^{-1}), fiber micronaire (FM), fiber uniformity (FU, %) and fiber elongation (FE, %). The analysis of variance (ANOVA) for phenotypic data was conducted using the SPSS22.0 software.

SNP Genotyping

Genomic DNA was isolated from young leaf tissue of all accessions using a modified cetyltrimethylammonium bromide (CTAB) method as described by Paterson et al. (1993). Reduced-representation DNA sequences of 160 early-maturity cotton accessions have been obtained by specific-locus amplified fragment sequencing (SLAF-seq) approach with coverage of approximate 5.50×. To mine the SNPs with higher quality, the raw reads were mapped to the *G. hirsutum* L. TM-1 genome v 1.1 (Zhang et al., 2015) using BWA software (Li and Durbin, 2009). The GATK (McKenna et al., 2010), and SAMTools (Li et al., 2009) packages were used for SNP calling. The filtered SNPs, with a missing rate <10% and a minor allele frequency (MAF) ≥ 0.05, were reserved and used for the subsequent analysis.

Clustering Analysis, Population Structure and Linkage Disequilibrium (LD) Analysis

A neighbor-joining phylogenetic tree among 160 individuals was constructed using the filtered SNPs by the Tassel 5.2 software (Bradbury et al., 2007). The population structure was analyzed using a principal component analysis (PCA) approach with the Tassel 5.2 program (Bradbury et al., 2007). LDs between SNPs were estimated as the squared correlation coefficient (R^2) of alleles using the Tassel 5.2 tool (Bradbury et al., 2007). The R^2-values were calculated within a 0- to 10-cM window.

Genome-Wide Association Study and Allelic Variation Analysis

Six ML-GWAS methods, including the mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, pKWmEB, and ISIS EM-BLASSO, were used in this study. The mrMLM is a multi-locus model including markers selected from the rMLM method with a less stringent selection criterion (Wang et al., 2016). The FASTmrMLM reduces the running time in mrMLM by more than 50%, and also shows slightly high statistical power in QTN detection, high accuracy in QTN effect estimation and low false positive rate as compared to mrMLM (Tamba and Zhang, 2018). FASTmrEMMA is a fast multi-locus random-SNP-effect EMMA model, which is more powerful in QTN detection and model fit (Wen et al., 2017). The pLARmEB integrates least angle regression with empirical Bayes to perform ML-GWAS under polygenic background control (Zhang et al., 2017). The pKWmEB retains the high power of Kruskal–Wallis test, and provides QTN effect estimates and effectively controls false positive rate (Ren et al., 2018). The ISIS EM-BLASSO has the highest empirical power in QTN detection and the highest accuracy in QTN effect estimation, and it is the fastest, as compared with EMMA and mrMLM (Tamba et al., 2017). All parameters were set at default values, and the critical thresholds of significant association for all the above six methods were set at LOD = 3.00 (Wang et al., 2016; Tamba et al., 2017; Wen et al., 2017; Zhang et al., 2017; Ren et al., 2018; Tamba and Zhang, 2018).

The phenotypic-effect value of each allelic variation was calculated by the phenotypic data over the accessions with each type, and box plots of the relative phenotypic data were produced using the R software.

Prediction of Potential Candidate Genes

Putative candidate genes were identified by physical positions of significant trait-associated SNP loci in the *G. hirsutum* L. reference genomes v1.1 (Zhang et al., 2015). According to LD decay distance, the interval for the prediction of candidate genes for the significant SNP loci was determined. The genes distributed in these regions were collected. Transcriptome sequencing data from 12 upland cotton tissues (including fiber in 5, 10, 20, 25 DPA (days post anthesis), root, stem, leaf, torus, calycle, cotyledon, petal and pistil) were available on the cotton website (https://cottonfgd.org/). Heat maps of the putative candidate gene expression patterns were drawn using the R package "pheatmap." The biological functions of putative candidate genes were annotated by gene ontology (GO) items on the cotton website (https://cottonfgd.org/).

RESULTS

SNP Genotyping

To gain insight into the genetic bases of fiber-quality related traits, 160 early-maturity upland cotton accessions were performed using SLAF-seq, and a complete set of markers containing 72,792 high-quality SNPs was explored by filtering according to the stringent quality control. These detected markers consisted of 47,594 and 25,198 SNPs in the At and Dt chromosomes respectively, and were unevenly distributed on all the 26 chromosomes of upland cotton. Moreover, the SNP loci with maximal number were identified on chromosome A10 (5013), while those with the minimal number were detected on chromosome D04 (1479). The average marker density was about one SNP per 28.10 kb genomic regions. The greatest marker density was found on chromosome A10 with one SNP per 20.12 kb, while the smallest marker density was seen in chromosome D05, with one SNP per 39.93 kb (**Table 1**).

TABLE 1 | Statistics of SNPs.

Chr.	Number of SNPs	Density of SNP (kb/SNP)	Chr.	Number of SNPs	Density of SNP (kb/SNP)
A01	4410	22.65	D01	1873	32.81
A02	3460	24.12	D02	1846	36.45
A03	3367	29.78	D03	1655	28.21
A04	2485	25.32	D04	1479	34.79
A05	3320	27.73	D05	1551	39.93
A06	4414	23.37	D06	1991	32.29
A07	3388	23.10	D07	1658	33.36
A08	4703	22.03	D08	2699	24.41
A09	2443	30.70	D09	1635	31.19
A10	5013	20.12	D10	2240	28.29
A11	3817	24.45	D11	2466	26.80
A12	2996	29.20	D12	2016	29.32
A13	3778	21.16	D13	2089	28.98

Chr., chromosome, according to the upland cotton reference genome (Zhang et al., 2015).

Phenotypic Variation

To examine whether significant phenotypic variances exist in the fibers among the 160 upland cotton accessions, the five fiber-quality related traits including FL, FS, FU, FM and FE were examined. The results showed that the parameters of fibers from different accessions were quite diverse (**Table 2**). For instance, the FL ranged from 24.07 to 33.69 mm, with a mean of 28.09 mm, the FS had a great variation ranging from 22.70 to 40.65 cN.tex^{-1}, and the FM from four environments varied from 2.50 to 6.00, with an average value of 4.83. Additionally, the FU and FE had wide distributions and variations (**Table 2**, **Figure 1**). These results indicate that early-maturity cotton varieties had broad variation in fiber-quality related traits under different planting conditions.

We observed that the phenotypic values of FL and FS at Anyang (AY) were significantly lower than those at Shihezi (SHZ). By contrast, there were no significant differences in FM and FU between the two locations. The FE values in AY-2014 were also strikingly lower than those in other environments (**Table 2**, **Figure 1**). Furthermore, the statistically significant differences ($P < 0.001$) were observed among genotypes, environments, and the genotype × environment interactions on all the five target traits (**Table S2**). These data suggest that the

TABLE 2 | Phenotypic distribution range of five fiber-quality related traits of 160 early-maturity upland cotton accessions.

Traits	Environments	Mean	Min	Max	*SD*	CV (%)
FL (mm)	AY-2014	29.08	26.31	33.69	1.25	4.31
	AY-2015	28.33	24.80	32.15	1.65	5.82
	SHZ-2014	27.86	25.79	31.53	1.08	3.88
	SHZ-2015	27.10	24.07	31.07	1.47	5.43
FS (cN.tex^{-1})	AY-2014	30.40	26.07	39.43	2.22	7.31
	AY-2015	28.75	22.70	40.65	3.15	10.97
	SHZ-2014	29.85	25.94	40.54	2.24	7.51
	SHZ-2015	26.07	22.87	34.47	2.44	9.35
FU (%)	AY-2014	84.60	82.00	86.53	0.86	1.02
	AY-2015	84.06	80.30	86.80	1.29	1.54
	SHZ-2014	84.04	80.80	86.43	1.11	1.32
	SHZ-2015	83.29	79.40	86.87	1.44	1.72
FM	AY-2014	4.68	3.28	5.78	0.50	10.61
	AY-2015	4.85	2.85	6.00	0.50	10.25
	SHZ-2014	4.73	3.78	5.39	0.32	6.68
	SHZ-2015	5.05	3.67	5.80	0.35	6.98
FE (%)	AY-2014	6.30	6.10	6.53	0.09	1.39
	AY-2015	6.78	6.50	7.05	0.12	1.77
	SHZ-2014	6.74	6.50	6.93	0.08	1.22
	SHZ-2015	6.68	6.40	6.93	0.12	1.79

FL, fiber length; FS, fiber strength; FM, fiber micronaire; FU, fiber uniformity; FE, fiber elongation; Max, maximum; Min, minimum; SD, standard deviation; CV, coefficient of variation; AY, Anyang; SHZ, Shihezi.

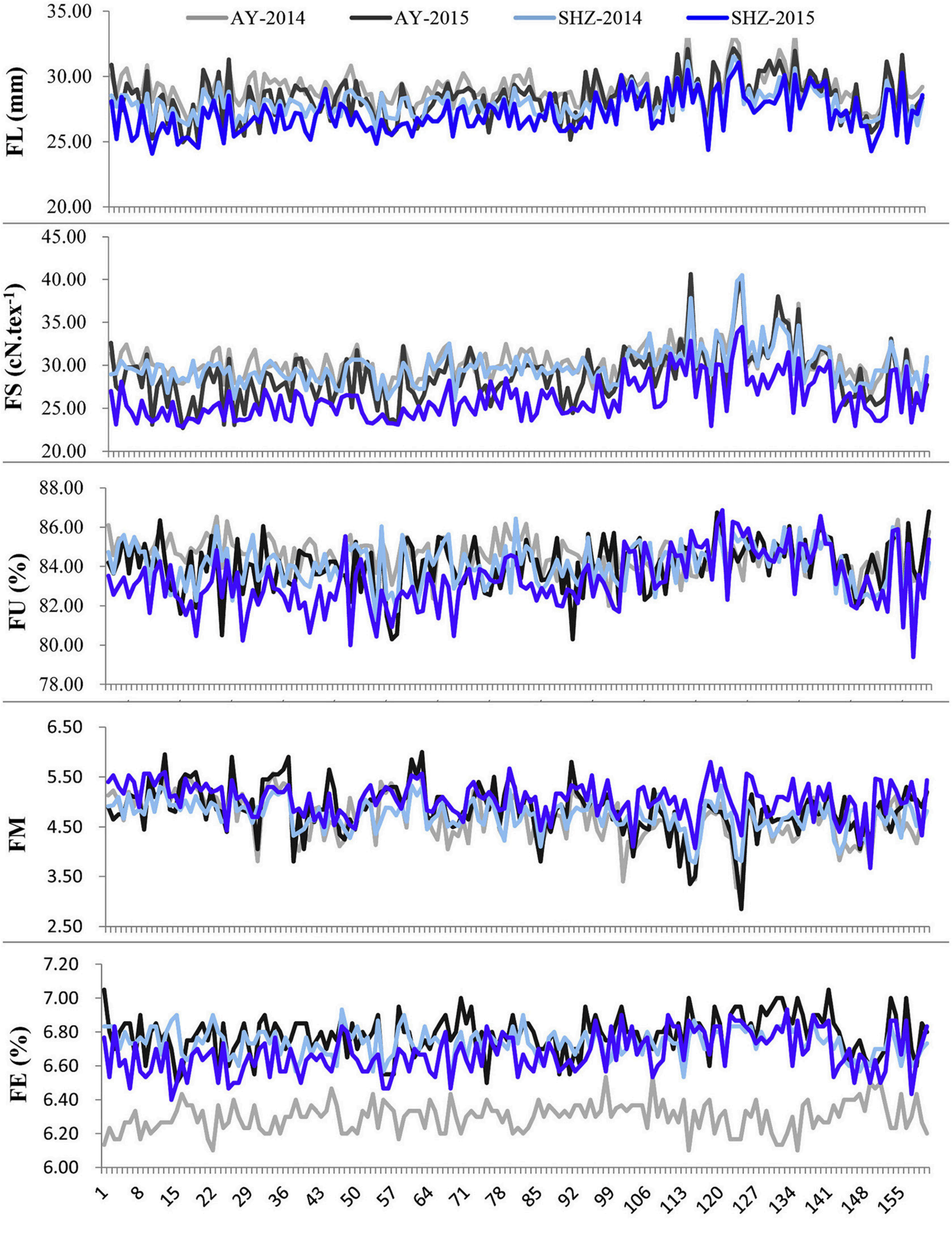


FIGURE 1 | Phenotypic distributions of five fiber-quality related traits of 160 early-maturity upland cotton accessions in four cultivating environments.

five fiber-quality related traits were significantly influenced by the environmental conditions.

Population Structure and Linkage Disequilibrium Analysis

To understand the phylogenetic relationship of the 160 upland cotton genotypes, a neighbor-joining phylogenetic tree was conducted based on their genetic distances, which derived from the SNP differences in these accessions. The population could be divided into three different groups, designated pop I (YRR, with 54 accessions), pop II (NIR, with 44 accessions) and pop III (YRR, NIR and NSEMR, with 62 accessions), respectively (**Figure 2A**). Furthermore, we found that there are an intimate genetic relationship between NSEMR accessions and the early varieties from YRR and NIR, which were mainly assigned to pop III, while the recent accessions from YRR and NIR belonged to pop I and pop II, respectively. These findings imply that early-maturity accessions in YRR and NIR might derive from NSEMR varieties in upland cotton.

Next, the population structure of the panel was analyzed using a PCA on the basis of the identified SNPs. Three conceivable subpopulations were separated by PC1 and PC2 (**Figure 2B**). Similarly, YRR, NIR and mixed group (YRR, NIR and NSEMR) were respectively distinguished via PCA. Based on the results from both the phylogenetic tree and PCA, the panel was separated into three groups (**Figures 2A,B**).

To examine the LD decay distance in the panel, its decay rate was estimated using the SNPs. The result showed that the genome-wide LD decay rate of the natural population was approximately 400 kb, where the R^2 drops to half of the maximum value (**Figure 2C**). Due to the average marker density with one SNP per 28.10 kb (**Table 1**), we concluded that these markers were sufficiently dense for detecting the associated QTNs.

Multi-Locus Genome-Wide Association Studies

A total of 70 significant QTNs were simultaneously detected to be associated with the above five objective traits by at least three multi-locus GWAS (ML-GWAS) methods (**Table 3**). Among these QTNs, 16, 20, 9, 16, and 9 were found to be associated with FL, FS, FM, FU, and FE, respectively. Among the 70 significant QTNs, three were simultaneously presented across at least two environments (**Table 3**). One (D11_21619830) for FL, with a high proportion of total phenotypic variance explained by the QTN (2.35∼11.07%), was found simultaneously in the two planting environments (AY-2014 and SHZ-2014) (**Figure 3A**, **Table 3**). Note that this QTN was detected by three methods (mrMLM, ISIS EM-BLASSO and pLARmEB) in AY-2014, and by four methods (mrMLM, FASTmrMLM, pLARmEB and ISIS EM-BLASSO) in SHZ-2014. Another QTN (A05_28352019) for FS was found by four methods (mrMLM, FASTmrMLM, pLARmEB and pKWmEB) in AY-2014 and by three methods (mrMLM, FASTmrMLM and pLARmEB) in SHZ-2014 (**Figure 3A**, **Table 3**). Most meaningfully, the QTN (D03_34920546) for FM was presented simultaneously in three environments (AY-2014, AY-2015 and SHZ-2015), and was detected at AY (2014 and 2015) by all the six ML-GWAS methods (**Figure 3A**, **Table 3**). In conclusion, the three identified QTNs (A05_28352019, D03_34920546 and D11_21619830), might be some steady major QTNs controlling the target traits.

Identification of Favorable Allelic Variations

To identify favorable alleles of QTNs for target traits, we focused on the above 3 steady QTNs, which exhibited the maximum LOD, –lg(*P*) value and phenotypic variation. The striking QTN D11_21619830 presented three types of allele (AA, AG and GG), and the accessions with the favorable allele AA ($n = 112$) showed significantly higher FL than those with the GG ($n = 26$) and AA ($n = 20$) alleles (**Figure 3B**). Moreover, we found that QTN A05_28352019 had three types of allelic variation AA, AG and GG, respectively, where the average FS of the favorable allele GG (28.89 cN.tex^{-1}) was higher than those of the AA (26.26 cN.tex^{-1}) and AG (26.98 cN.tex^{-1}) (**Figure 3B**). Additionally, the peak QTN (D03_34920546) had three allelic variations (AA, AG and GG), and the accessions with the GG variation showed higher FM than those with the alternate AA variation. Considering the most excellent level of FM (3.70∼4.20) for spinning, allele AA of the peak QTN could be regarded as the favorable allele with the mean FM value of 4.28, whereas the corresponding type GG was the unfavorable allele with the mean FM value of 4.89 (**Figure 3B**). These findings indicated that the fibers of the accessions with favorable allelic variations

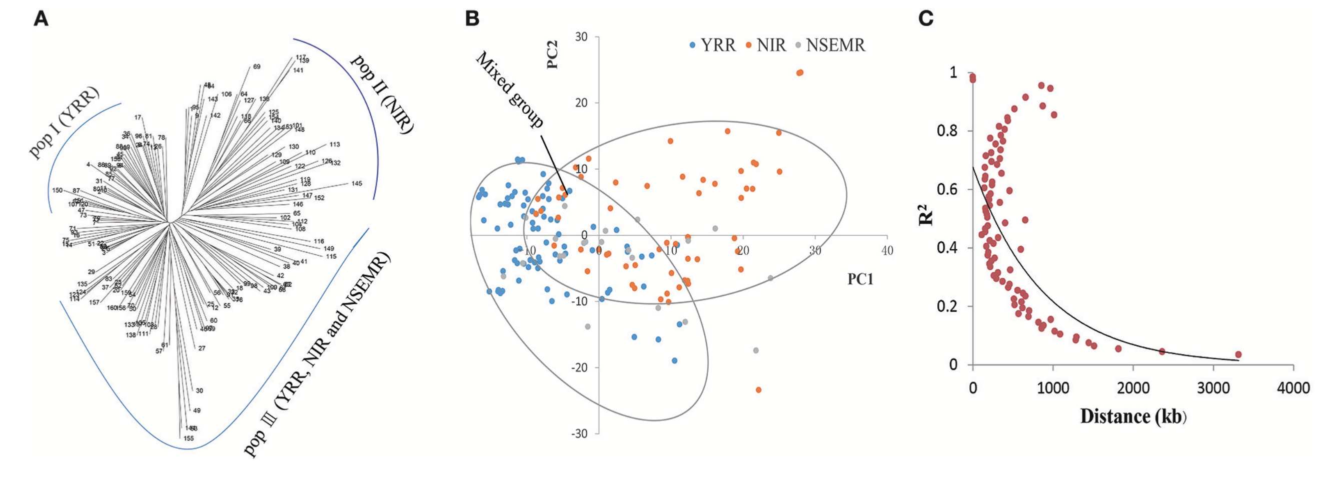


FIGURE 2 | Phylogenetic tree, population structure and LD decay of 160 early-maturity upland cotton accessions. **(A)** Neighbor-joining phylogenetic tree of all cotton accessions. The pop III was a mixed group including YRR, NIR and NSEMR. **(B)** Principal component analysis (PCA) of the association panel. **(C)** The entire genome LD decay of the population.

TABLE 3 | The significant QTNs for five fiber-quality related traits detected simultaneously by using three or more multi-locus GWAS methods.

Traits	QTN ID	Env.	Pos.(Mb)	Chr.	LOD	R^2 (%)	ML-GWAS Methods
FL	A02_8008893	AY-2014	8.01	A02	3.69–6.42	5.66–15.50	2, 3, 5
	A07_14808135	AY-2014	14.80	A07	3.10–5.22	6.97–12.22	1, 2, 5, 6
	A09_5999806	AY-2014	6.00	A09	3.23–7.46	2.47–12.33	1, 2, 4
	A11_77306905	AY-2014	77.31	A11	3.90–6.75	2.11–9.81	1, 2, 4, 5
	D01_38549510	AY-2014	38.55	D01	4.32–4.90	2.83–9.30	1, 2, 5
	D03_37661749	AY-2014	37.66	D03	3.56–5.11	1.71–6.70	1, 2, 4, 5, 6
	D08_44134483	AY-2014	44.13	D08	4.30–4.48	1.98–6.02	1, 2, 5
	D11_21619830*	AY-2014	21.62	D11	4.15–5.29	2.35–9.18	1, 3, 5
	D03_41720764	AY-2015	41.72	D03	3.00–5.09	7.08–13.15	1, 2, 4, 6
	A07_71351661	SHZ-2014	71.35	A07	4.95–5.94	2.84–11.07	1, 4, 5
	D11_21619830*	SHZ-2014	21.62	D11	3.99–5.45	2.79–11.32	1, 2, 3, 5
	A01_20468506	SHZ-2015	20.47	A01	3.88–4.72	8.71–9.84	1, 3, 5
	A07_5555999	SHZ-2015	5.56	A07	3.06–4.90	5.24–13.44	1, 2, 5, 6
	A09_56201893	SHZ-2015	56.20	A09	3.07–5.07	2.22–6.81	2, 3, 4, 5, 6
	D06_15590320	SHZ-2015	15.59	D06	3.23–5.64	2.31–6.62	2, 3, 4, 5, 6
	D13_20291732	SHZ-2015	20.29	D13	3.01–3.49	10.41–17.71	2, 5, 6
FS	A05_28352019*	AY-2014	28.35	A05	4.21–6.09	4.73–14.91	1, 2, 5, 6
	A01_3833158	AY-2015	3.83	A01	4.70–7.47	8.01–8.89	1, 5, 6
	A11_48101548	AY-2015	48.10	A11	3.60–4.21	7.86–8.44	1, 2, 6
	D01_53611999	AY-2015	53.61	D01	4.13–4.95	4.73–7.43	2, 5, 6
	D08_54727428	AY-2015	54.73	D08	3.58–10.36	2.23–9.81	1, 2, 6
	D08_63040058	AY-2015	63.04	D08	3.97–4.91	3.70–5.98	4, 5, 6
	A05_28352019*	SHZ-2014	28.35	A05	3.14–6.70	4.03–8.57	1, 2, 5
	A05_81758788	SHZ-2014	81.76	A05	3.37–6.73	2.59–6.47	1, 2, 5, 6
	A07_71351661	SHZ-2014	71.35	A07	3.14–7.25	3.36–6.33	1, 2, 4, 5
	A08_71454278	SHZ-2014	71.45	A08	3.27–5.88	6.75–11.73	2, 5, 6
	A09_30635120	SHZ-2014	30.64	A09	3.28–5.78	5.77–11.18	1, 2, 5
	A11_85908613	SHZ-2014	85.91	A11	4.32–9.06	15.66–18.87	2, 3, 5
	D01_17607059	SHZ-2014	17.61	D01	3.55–13.41	3.98–13.41	2, 5, 6
	D06_15477129	SHZ-2014	15.48	D06	5.27–6.94	5.60–7.21	1, 2, 5
	D07_45641817	SHZ-2014	45.64	D07	3.36–5.33	0.93–1.33	1, 2, 5
	A09_56201893	SHZ-2015	56.20	A09	3.14–4.33	1.80–4.39	3, 4, 5
	A13_60864258	SHZ-2015	60.86	A13	3.77–5.28	3.63–8.03	1, 2, 4
	D01_53662824	SHZ-2015	53.66	D01	4.35–5.55	2.69–7.99	3, 5, 6
	D06_15590320	SHZ-2015	15.59	D06	3.13–4.24	0.88–4.39	1, 2, 4, 5, 6
	D12_51137790	SHZ-2015	51.14	D12	3.14–5.64	4.51–14.69	3, 5, 6
FM	A03_97922050	AY-2014	97.92	A03	4.01–7.02	5.00–6.91	1, 2, 5, 6
	A05_31842417	AY-2014	31.84	A05	3.04–5.46	3.53–15.10	1, 2, 3, 4, 5, 6
	D03_34920546*	AY-2014	33.49	D03	4.18–9.66	5.58–13.87	1, 2, 3, 4, 5, 6
	D03_34920546*	AY-2015	34.92	D03	9.31–12.88	16.69–29.90	1, 2, 3, 4, 5, 6
	A06_5267401	SHZ-2014	35.27	A06	3.54–5.66	3.79–8.48	1, 3, 4, 6
	D03_34920546*	SHZ-2014	34.92	D03	3.86–5.26	5.89–17.95	1, 3, 5, 6
	D09_23399960	SHZ-2014	23.40	D09	3.99–5.78	5.12–11.08	1, 2, 3
	D11_1353254	SHZ-2014	1.35	D11	3.01–6.58	3.35–13.14	1, 2, 3, 4, 5
	D04_42032569	SHZ-2015	42.03	D04	3.59–4.24	4.81–11.96	2, 3, 5
FU	D07_43127704	AY-2014	43.13	D07	4.61–6.30	4.40–14.6	1, 2, 3, 4, 5
	A01_84353970	AY-2015	84.35	A01	3.12–3.77	4.24–4.68	1, 3, 5
	A05_65094242	AY-2015	65.09	A05	3.13–4.43	2.58–12.3	2, 3, 4
	A05_81758788	AY-2015	81.76	A05	3.82–8.50	5.32–9.88	1, 2, 5
	A09_577845	AY-2015	0.58	A09	3.57–5.04	2.34–4.94	1, 2, 3
	A11_90350675	AY-2015	90.35	A11	4.13–5.03	3.10–3.89	1, 2, 3

(Continued)

TABLE 3 | Continued

Traits	QTN ID	Env.	Pos.(Mb)	Chr.	LOD	R^2 (%)	ML-GWAS Methods
	D11_22135870	AY-2015	22.14	D11	3.00–6.25	4.17–6.68	1, 2, 5
	D12_29952510	AY-2015	29.95	D12	3.77–4.33	5.70–9.85	3, 4, 5
	A06_56710162	SHZ-2014	56.71	A06	3.51–5.14	4.32–12.49	1, 2, 3, 4, 5
	A07_71351661	SHZ-2014	71.35	A07	3.32–4.12	4.68–7.58	1, 2, 3
	A10_91235190	SHZ-2014	91.24	A10	4.00–8.37	6.79–12.89	1, 2, 3, 4, 5
	D08_44134483	SHZ-2014	44.13	D08	3.00–6.49	4.30–9.44	1, 2, 3, 4, 5
	D11_22044769	SHZ-2014	22.04	D11	3.34–7.46	9.50–23.35	1, 2, 4, 5, 6
	A04_57215161	SHZ-2015	57.22	A04	3.54–4.07	11.12–15.91	1, 2, 6
	D03_41720764	SHZ-2015	41.72	D03	4.05–6.07	6.67–9.88	1, 2, 3, 6
	D10_56039747	SHZ-2015	56.04	D10	4.41–7.04	7.97–12.76	2, 3, 4, 5, 6
FE	A11_77306905	AY-2014	77.31	A11	3.15–7.78	4.58–12.40	1, 2, 3, 4, 5, 6
	D07_21396263	AY-2014	21.40	D07	6.24–7.56	5.81–10.32	1, 2, 5
	D07_43127704	AY-2014	43.13	D07	4.15–6.40	1.47–6.99	1, 3, 5, 6
	A01_60422471	AY-2015	60.42	A01	3.67–5.88	3.65–17.78	1, 2, 4
	A05_22406870	AY-2015	22.41	A05	3.148–3.80	0.83–8.12	1, 2, 5
	A09_46286411	AY-2015	46.29	A09	3.84–6.68	7.93–10.45	1, 2, 4
	D01_53662824	AY-2015	53.66	D01	3.28–6.24	7.09–14.52	1, 3, 4, 6
	D04_32424503	AY-2015	32.42	D04	3.02–8.53	3.35–5.79	1, 2, 3, 5
	A09_56201893	SHZ-2015	56.20	A09	3.40–4.12	4.75–12.57	3, 4, 6

FL, fiber length; FS, fiber strength; FM, fiber micronaire; FU, fiber uniformity; and FE, fiber elongation; QTN, quantitative trait nucleotide; Env., Environment; Chr., chromosome; Pos., position; AY, Anyang; SHZ, Shihezi. LOD value indicates the significance levels and R^2 (%) indicates the percentage of phenotypic variation explained by each QTN.
mrMLM, FASTmrMLM, ISIS EM-BLASSO, FASTmrEMMA, pLARmEB and pKWmEB are marked by 1 to 6, respectively.
**Indicates the significant QTNs presented simultaneously across at least two environments*

were clearly improved compared to those of the accessions with unfavorable allelic variations.

Prediction of Candidate Genes

The genomic regions (±400 kb around the associated QTNs) of QTN-linked candidate genes were adopted according to the genome-wide LD decay distances (about 400 kb) in this study. Thus, three target regions of the candidate genes were determined as A05: 27.95–28.75, D03: 34.52–35.32, and D11: 21.22–22.02 Mbp, and a total of 29, 32 and 35 genes were presented respectively in the above regions, according to upland cotton reference genome v1.1 (Zhang et al., 2015; **Table S3**). Furthermore, we observed that the expression of 72 genes of them was clearly increased in 12 cotton tissues using RNA-Seq (**Figure 4**). Among these genes, *Gh_A05G2325*, *Gh_A05G2329*, *Gh_A05G2334*, *Gh_D11G1853*, *Gh_D11G1876*, and *Gh_D11G1879*, were highly expressed in the fiber. Notably, *Gh_A05G2334* was dominantly expressed in all the four fiber samples; *Gh_D11G1853* was mainly expressed in fibers of 20 and 25 DPA; and *Gh_D11G1876* and *Gh_A05G2325* was preferentially expressed in fiber of 25 DPA; whereas *Gh_A05G2329* and *Gh_D11G1879* had the maximum expression level in the fibers of 5 and 10 DPA, respectively. Also, the transcriptional abundances of *Gh_D03G1012* and *Gh_A05G2335* were slightly higher in fibers than in the other tissues (**Figure 4**). These results suggest that the six genes (*Gh_A05G2325*, *Gh_A05G2329*, *Gh_A05G2334*, *Gh_D11G1853*, *Gh_D11G1876*, and *Gh_D11G1879*) might play important roles in controlling fiber quality of early-maturity upland cotton.

To further understand thoroughly the above six putative candidate genes for target traits, their biological functions were annotated by gene ontology (GO) items. Three genes (*Gh_A05G2334*, *Gh_D11G1876*, and *Gh_D11G1879*) were annotated as transcription factors, such as sequence-specific DNA binding, DNA-binding transcription factor activity and regulation of transcription (**Table 4**). *Gh_A05G2334* encoded the agamous-like MADS-box protein AGL11 which likely plays roles in many aspects of plant growth and development (Rounsley et al., 1995). These results indicate that the putative candidate genes may regulate fiber development by DNA-binding transcription factors in early-maturity upland cotton.

DISCUSSION

The Origin and Domestication of Chinese Early-Maturity Upland Cotton

To exploit the limited natural resources and increase economic income of cotton producers, it is especially necessary to make use of the double cropping systems and mechanical harvesting in the major cotton growing regions in China. Thus, early-maturity cotton cultivars are needed. Indeed, early-maturity cotton varieties are attracting much attention from many cotton growers and breeders. Fiber characters are complicated and comprehensive traits regulated by a lot of QTL and influenced

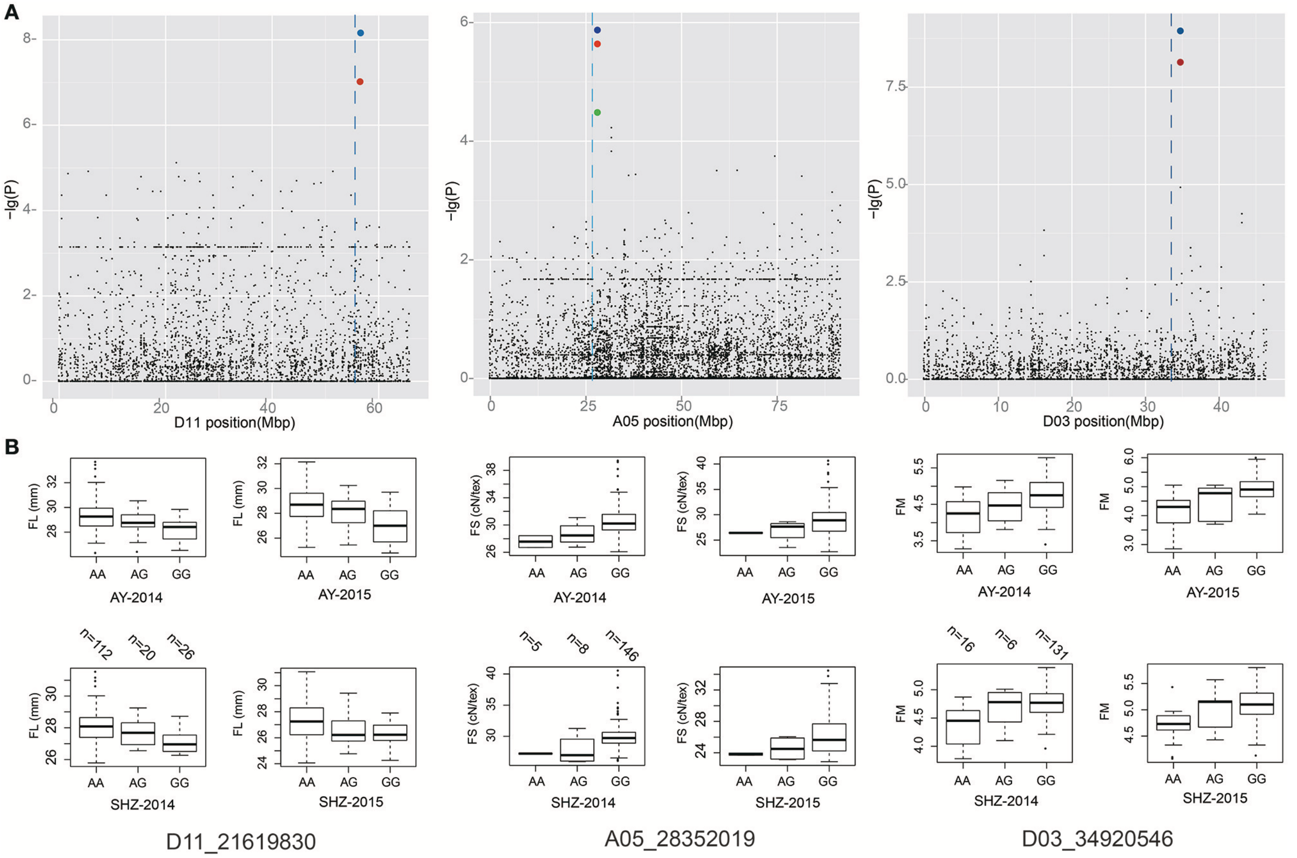


FIGURE 3 | Local Manhattan plot (top), and box plots for the fiber-quality related traits (bottom). **(A)** Manhattan plots of FL, FS, and FM on chromosome A05, D03, and D11, respectively. **(B)** Box plots of the significant QTNs (D11_21619830, A05_28352019, and D03_34920546). Each dot represents an SNP. The vertical dashed lines indicate the genomic region containing the significant QTNs. The red and blue circles mark the significant QTNs.

easily by many external factors (Ulloa and Meredith, 2000). Its related traits for example FL, FS, and FM are more important for the spinning industry. Previous investigations had shown that FL and FS have significant negative correlations with earliness in cotton. Thus, the early-maturity cotton varieties have much lower fiber quality than late-maturity ones. Sun et al. (2017) reported the association panel including early-, middle- and late-maturity cotton varieties have a big phenotypic variation of the FL (22.07∼35.56 mm) and FS (22.69∼36.80 cN.tex^{-1}). In this study, FL of the panel ranged from 24.07 to 33.69 mm, with a mean of 28.09 mm; while the FS had a great variation ranging from 22.70 to 40.65 cN.tex^{-1}. These findings indicate that FL of our association population of the early-maturity cotton has small distribution ranges compared with the previous results.

Although China is one of the largest nations producing and consuming cotton in the world, it is not an upland cotton domestication country (Zhang et al., 2013). The early cotton varieties were primarily developed by using introduced varieties (Zhang et al., 2013). King cultivar from America is the ancestor of the Chinese early-maturity upland cotton. Most of Chinese early-maturity cotton varieties of the early stage, such as "Jinmian1," "Heishanmian1," "Liaomian1," "Zhongmiansuo10," and "Xinluzao10," were all derived from "Guannong1," which had a breeding pedigree from the King cultivar. In this study, the association panel contained the above-mentioned core germplasms, and consisted of more than 80% of the Chinese early-maturity cotton varieties. Thus, it can represent the wide genetic diversity of Chinese early-maturity upland cotton. In the early stage, the Chinese early-maturity cotton varieties were developed by utilizing the core germplasms from NSEMR ("Jinmian1," "Heishanmian1" and "Liaomian1). On the basis of the clustering of phylogenetic tree and PCA of the study, along with breeding history, the early-maturity cotton could be divided into three groups, designated pop I (the recent accessions from YRR), pop II (the recent accessions from NIR) and pop III (the NSEMR varieties and the early germplasms from YRR and NIR), respectively. These findings suggest that early-maturity accessions in YRR and NIR might derive from the NSEMR early varieties in Chinese upland cotton.

Comparison of Our GWAS Results With QTL or QTNs Detected in Previous Studies

In the recent 30 years, many QTL have been mapped, and some fiber-quality QTL hotspots have been discovered by a comparative meta-analysis (Said et al., 2015). It has been shown that chromosome D11 (c24) has the most prominent cluster

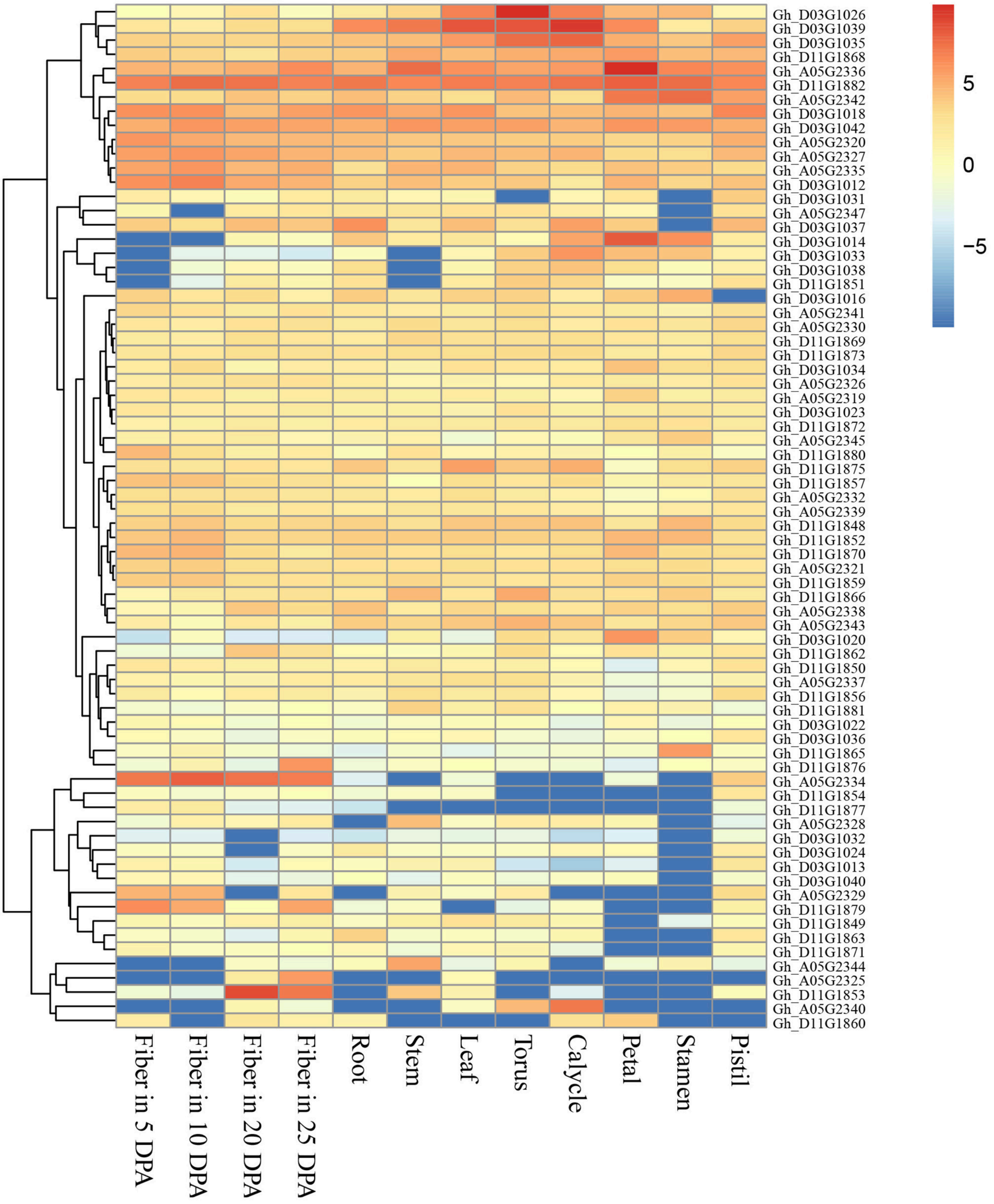


FIGURE 4 | Heat map of expression level of the 72 genes in 12 upland cotton tissues. The red indicates high expression, and the blue shows low expression.

TABLE 4 | The biological function annotations of the six putative candidate genes for five fiber-quality related traits.

Gene ID	Name	Chr.	Description	Function annotations
Gh_A05G2325		A05	Non-specific lipid-transfer protein 2	
Gh_A05G2329		A05		
Gh_A05G2334	*AGL11*	A05	Agamous-like MADS-box protein AGL11	DNA binding; DNA-binding transcription factor activity; regulation of transcription; protein dimerization activity
Gh_D11G1853	*ephx3*	D11	Epoxide hydrolase 3	
Gh_D11G1876	*GBF1*	D11	G-box-binding factor 1	DNA-binding transcription factor activity; regulation of transcription; sequence-specific DNA binding; nutrient reservoir activity
Gh_D11G1879	*ATHB-40*	D11	Homeobox-leucine zipper protein ATHB-40	DNA binding; DNA-binding transcription factor activity; regulation of transcription; transcription regulatory region sequence-specific DNA binding

Function annotations of putative candidate genes were conducted by gene ontology (GO) items on the cotton website (https://cottonfgd.org/); Chr.: Chromosome.

carrying FL, FE and FS QTL hotspots between CIR026 and NAU2407b. A hotspot cluster A07 (c7) carrying FL and FS QTL between E1M7_80 and CG05a has also been found (Said et al., 2015). Another cluster carrying FE, FL and FM QTL hotspots on D01 (c14) between CIR246 and G1012 has been identified; and the region between E5M4_480 and pAR544 harbors a hotspot cluster carrying FS QTL on chromosome D03 (c16) (Said et al., 2015). Additionally, some stable QTL for FS on A07 (Chr.07) have been identified by QTL mapping (Tan et al., 2015; Fang X. et al., 2017). Similarly, a few associated SNP loci with fiber quality have been detected via GWAS in upland cotton (**Table S4**). Among the identified FL-associated SNPs, most of markers were located on chromosome A10 and D11, such as A10_65694094, A10_65696540, D11_24030081 and D11_24030087. Recent reports have shown a number of cluster_A07 SNPs for FS are distributed in genome region A07: 71.99–72.25 Mbp (Sun et al., 2017; Ma Z. et al., 2018). In addition, we also found the major genomic region (D11:24.03–24.10 Mbp) consisting of nine SNP loci associated with FL, which was previously detected (Su et al., 2016a).

In the current study, we characterized the significant QTNs (D11_21619830, A05_28352019 and D03_34920546) for fiber-quality related traits. These QTNs were detected using several new ML-GWAS methods in at least two environments. Compared with the mapped QTL of the previous studies, the QTN D11_21619830 was located in the region of QTL hotspot clusters for fiber quality. Compared with the associated loci of previous GWAS, these associated QTNs were excluded in the genomic regions of the previous reports. Therefore, these identified SNPs may be novel QTNs controlling fiber quality in our association population of early-maturity cotton.

Superiority of the New Multi-Locus GWAS

Most of previous studies have focused on genetic bases of some complicated traits using general linear model (GLM) and mixed linear model (MLM) based on a single-locus GWAS (SL-GWAS) (Yu et al., 2006; Zhang et al., 2010). However, both of these models have certain shortcomings. A big false positive incidence is the uppermost disadvantage of GLM because polygenic kinship is not considered (Korte and Farlow, 2013). In MLM, the stringent P threshold ($P = 0.05/n$, n is the number of SNPs) leads to missing many significant QTNs, particularly small-effect QTNs (Wang et al., 2016). To make up for deficiencies of GLM and MLM, some multi-locus GWAS (ML-GWAS) methodologies have been developed, such as mrMLM (Wang et al., 2016), FASTmrMLM (Tamba and Zhang, 2018), FASTmrEMMA (Wen et al., 2017), ISIS EM-BLASSO (Tamba et al., 2017), pLARmEB (Zhang et al., 2017), and pKWmEB (Ren et al., 2018). Compared with the conventional SL-GWAS MLM methods, these ML-GWAS methods are more powerful and have the advantages of accuracy. Thus, we adopted the ML-GWAS methods in this study.

In addition, the significant threshold of these new ML-GWAS methods is set to a LOD score = 3, which is equal to $-\lg(P) = 3.70$ (Wang et al., 2016). Although the standards are less stringent in the ML-GWAS methods than in the SL-MLM ones, their false positive rates are effectively reduced (Wang et al., 2016; Tamba et al., 2017; Wen et al., 2017; Zhang et al., 2017; Ren et al., 2018; Tamba and Zhang, 2018). Thus, the ML-GWAS approaches are considered more effective, practical and alternative. In this study, 70 QTNs significantly associated with five fiber-quality related traits were simultaneously identified in three or more ML-GWAS methods (**Table 3**). Further investigation showed that three stably expressed QTNs were commonly detected in multiple environments (**Table 3**). However, no significantly associated QTN was found when using the Tassel 5.0 in MLM [PCs + K, $-\lg(P) = -\lg(0.05/72792) = 6.16$]. These data suggest that the ML-GWAS methods are more powerful and robust when applying to detect the small-effect QTNs for fiber-quality related traits of upland cotton.

CONCLUSION

In this study, a total of 70 significant QTNs were simultaneously detected to be associated with five objective traits by three or more methods. Among these QTNs, D11_21619830, A05_28352019 and D03_34920546, significantly associated with FL, FS, and FM, respectively, were simultaneously presented across at least two environments. Furthermore, favorable allelic variations of the three QTNs and 96 genes contained in the three target genomic range were mined. Among these, six genes highly expressed in the fibers might be candidate genes identified by RNA-Seq method. In summary, many favorable QTN alleles and six candidate genes were identified to modulate fiber development in early-maturity upland cotton. This will lay a solid basis for breeding earliness and excellent fiber-quality cotton varieties in the future.

AVAILABILITY OF SUPPORTING DATA

The sequence read data from the SLAF-seq analysis for the 160 sequenced upland cotton lines are available in the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA314284/) (SRP071133 under the accession number PRJNA314284).

AUTHOR CONTRIBUTIONS

JS and CW designed and supervised the research. JS and ML performed multi-locus GWAS. JS and QM conducted the field trials to evaluate the five fiber-quality related traits. JS, CW, and FH wrote and revised the manuscript. All of the authors read and approved the manuscript.

FUNDING

This research was funded by the Chinese National Natural Science Foundation (31660409).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01169/full#supplementary-material

REFERENCES

Bao, Y. Q., Liu, A. Q., Chen, Q. K., Li, C. P., Liu, Z. S., Zhang, D. W., et al. (2014). Development trend and variety choice of mechanical harvest of precocious cotton in the north of Xinjiang. *China Cotton* 41, 44–45.

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. *Bioinform.* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Cai, C., Ye, W., Zhang, T., and Guo, W. (2014). Association analysis of fiber quality traits and exploration of elite alleles in upland cotton cultivars/accessions (*Gossypium hirsutum* L.). *J. Integr. Plant Biol.* 56, 51–62. doi: 10.1111/jipb.12124

Chen, Z. J., Scheffler, B. E., Dennis, E., Triplett, B. A., Zhang, T. Z., Guo, W. Z., et al. (2007). Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* 145, 1303–1310. doi: 10.1104/pp.107.107672

Dai, J. L., Li, W. J., Zhang, D. M., Tang, W., Li, Z. H., Lu, H. Q., et al. (2017). Competitive yield and economic benefits of cotton achieved through a combination of extensive pruning and a reduced nitrogen rate at high plant density. *Field Crops Res.* 209, 65–72. doi: 10.1016/j.fcr.2017.04.010

Du, X. B., Chen, B. L., Shen, T. Y., Zhang, Y. X., and Zhou, Z. G. (2015). Effect of cropping system on radiation use efficiency in double-cropped wheat–cotton. *Field Crops Res.* 170, 21–31. doi: 10.1016/j.fcr.2014.09.013

Fan, S. L., Yu, S. X., Song, M. Z., and Yuan, R. H. (2006). Construction of molecular linkage map and QTL mapping for earliness in short-season cotton. *Cotton Sci.* 18, 135–139. doi: 10.3969/j.issn.1002-7807.2006.03.002

Fang, L., Wang, Q., Hu, Y., Jia, Y., Chen, J., Liu, B., et al. (2017). Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* 49, 1089–1098. doi: 10.1038/ng.3887

Fang, X., Liu, X., Wang, X., Wang, W., Liu, D., and Zhang, J. (2017). Fine-mapping qFS07.1 controlling fiber strength in upland cotton (*Gossypium hirsutum* L.). *Theor. Appl. Genet.* 130, 795–806. doi: 10.1007/s00122-017-2852-1

Feng, L., Dai, J. L., Tian, L. W., Zhang, H. J., Li, W. J., and Dong, H. Z. (2017). Review of the technology for high-yielding and efficient cotton cultivation in the northwest inland cotton-growing region of China. *Field Crop Res.* 208, 18–26. doi: 10.1016/j.fcr.2017.03.008

Huang, C., Nie, X. H., Shen, C., You, C. Y., Li, W., Zhao, W. X., et al. (2017). Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. *Plant Biotechnol. J.* 3, 1–13. doi: 10.1111/pbi.12722

Jia, X., Pang, C., Wei, H., Wang, H., Ma, Q., Yang, J., et al. (2016). High-density linkage map construction and QTL analysis for earliness-related traits in *Gossypium hirsutum*. *BMC Genom.* 17:909. doi: 10.1186/s12864-016-3269-y

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. doi: 10.1186/1746-4811-9-29

Lande, R., and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124, 743–756.

Li, C., Wang, X., Dong, N., Zhao, H., Xia, Z., Wang, R., et al. (2013). QTL analysis for early-maturing traits in cotton using two upland cotton (*Gossypium hirsutum* L.) crosses. *Breed Sci.* 63, 154–163. doi: 10.1270/jsbbs.63.154

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinform.* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). 1000 Genome project data processing subgroup: the sequence alignment/map format and SAM tools. *Bioinform.* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lu, H. Q., Dai, J. L., Li, W. J., Tang, W., Zhang, D. M., Eneji, A. E., et al. (2017). Yield and economic benefits of late planted short-season cotton versus full-season cotton relayed with garlic. *Field Crops Res.* 200, 80–87. doi: 10.1016/j.fcr.2016.10.006

Ma, L., Liu, M., Yan, Y., Qing, C., Zhang, X., Zhang, Y., et al. (2018). Genetic dissection of maize embryonic callus regenerative capacity using multi-locus genome-wide association studies. *Front. Plant Sci.* 9:561. doi: 10.3389/fpls.2018.00561

Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., et al. (2018). Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* 50, 803–813. doi: 10.1038/s41588-018-0119-7

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297–1303. doi: 10.1101/gr.107524.110

Misra, G., Badoni, S., Anacleto, R., Graner, A., Alexandrov, N., and Sreenivasulu, N. (2017). Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. *Sci. Rep.* 7:12478. doi: 10.1038/s41598-017-12778-6

Nie, X., Huang, C., You, C., Li, W., Zhao, W., Shen, C., et al. (2016). Genome-wide SSR-based association mapping for fiber quality in nation-wide upland cotton inbreed cultivars in China. *BMC Genom.* 17:352. doi: 10.1186/s12864-016-2662-x

Paterson, A. H., Brubaker, C. L., and Wendel, J. F. (1993). A rapid method for extraction of cotton (*Gossypium Spp.*) genomic DNA suitable for RFLP or PCR analysis. *Plant Mol. Biol. Rep.* 11, 122–127. doi: 10.1007/BF02670470

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2018). pKWmEB: in tegration of Kruskal–Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Rounsley, S. D., Ditta, G. S., and Yanofsky, M. F. (1995). Diverse roles for MADS box genes in *Arabidopsis* development. *Plant Cell* 7, 1259–1269. doi: 10.2307/3870100

Said, J. I., Song, M. Z., Wang, H. T., Lin, Z. X., Zhang, X. L., Fang, D. D., et al. (2015). A comparative meta-analysis of QTL between intraspecific *Gossypium hirsutum* and interspecific *G. hirsutum* × *G. barbadense populations. Mol. Genet. Genom.* 290, 1003–1025. doi: 10.1007/s00438-014-0963-9

Shang, L. G., Liang, Q. Z., Wang, Y. M., Wang, X. C., Wang, K. B., Abduweli, A., et al. (2015). Identification of stable QTLs controlling fiber traits properties in multi-environment using recombinant inbred lines in Upland cotton (*Gossypium* hirsutum L.). *Euphytica* 205, 877–888. doi: 10.1007/s10681-015-1434-z

Song, M. Z., Fan, S. L., Pang, C. Y., Wei, H. L., Liu, J., and Yu, S. X. (2015). Genetic analysis of fiber quality traits in short season cotton (*gossypium hirsutum*, l.). *Euphytica* 202, 97–108. doi: 10.1007/s10681-014-1226-x

Song, M. Z., Yu, S. X., Fan, S. L., Ruan, R., H., and Huang, Z. M. (2005). Genetic analysis of main agronomic traits in short season upland cotton (*G.hirsutum* l.). *Acta Gossypii Sinica*. 17, 94–98.

Su, J., Fan, S., Li, L., Wei, H., Wang, C., Wang, H., et al. (2016c). Detection of favorable QTL alleles and candidate genes for lint percentage by GWAS in Chinese upland cotton. *Front. Plant Sci.* 7:1576. doi: 10.3389/fpls.2016.01576

Su, J., Li, L., Pang, C., Wei, H., Wang, C., Song, M., et al. (2016a). Two genomic regions associated with fiber quality traits in chinese upland cotton under apparent breeding selection. *Sci. Rep.* 6:38496. doi: 10.1038/srep38496

Su, J., Li, L., Zhang, C., Wang, C., Gu, L., Wang, H., et al. (2018). Genome-wide association study identified genetic variations and candidate genes for plant architecture component traits in Chinese upland cotton. *Theor. Appl. Genet.* 131, 1299–1314. doi: 10.1007/s00122-018-3079-5

Su, J., Pang, C., Wei, H., Li, L., Liang, B., Wang, C., et al. (2016b). Identification of favorable SNP alleles and candidate genes for traits related to early maturity via GWAS in upland cotton. *BMC Genom.* 17:687. doi: 10.1186/s12864-016-2875-z

Sun, Z. W., Wang, X. F., Liu, Z. W., Gu, Q. S., Zhang, Y., Li, Z. K., et al. (2017). Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in *Gossypium hirsutum* L. *Plant Biotechnol. J.* 1, 1–15. doi: 10.1111/pbi.12693

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tamba, C. L., and Zhang, Y. M. (2018). A fast mrMLM algorithm for multi-locus genome-wide association studies. *bioRxiv*. doi: 10.1101/341784

Tan, Z. Y., Fang, X. M., Tang, S. Y., Zhang, J., Liu, D. J., Teng, Z. H., et al. (2015). Genetic map and QTL controlling fiber quality traits in Upland cotton (*Gossypium hirsutum* L.). *Euphytica* 203, 615–628. doi: 10.1007/s10681-014-1288-9

Tang, S. Y., Teng, Z. H., Zhai, T. F., Fang, X. M., Liu, F., Liu, D. X., et al. (2015). Construction of genetic map and QTL analysis of fiber quality traits for Upland cotton (*Gossypium hirsutum* L.). *Euphytica* 201, 195–213. doi: 10.1007/s10681-014-1189-y

Ulloa, M., and Meredith, W. R. Jr. (2000). Genetic linkage map and QTL analysis of agronomic and fiber quality traits in an intraspecific population. *J. Cotton Sci.* 4, 161–170.

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform*. doi: 10.1093/bib/bbw145

Wu, X., Li, Y., Shi, Y., Song, Y., Zhang, D., Li, C., et al. (2016). Joint-linkage mapping and GWAS reveal extensive genetic loci that regulate male inflorescence size in maize. *Plant Biotechnol. J.* 14, 1551–1562. doi: 10.1111/pbi.12519

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Yu, S. X., Song, M. Z., Fan, S. L., Wang, W., and Yuan, R. H. (2005). Biochemical genetics of short-season cotton cultivars that express early maturity without senescence. *J. Integr. Plant. Biol.* 47, 334–342. doi: 10.1111/j.1744-7909.2005.00029.x

Zeng, L., Meredith, W. R., Gutiérrez, O. A., and Boykin, D. L. (2009). Identification of associations between SSR markers and fiber traits in an exotic germplasm derived from multiple crosses among *Gossypium* tetraploid species. *Theor. Appl. Genet.* 119, 93–103. doi: 10.1007/s00122-009-1020-7

Zhang, J., Feng, J., Ni, Y., Wen, Y., Niu, Y., Tamba, C. L., et al. (2017).pLARmEB: integration of least angle regression with empirical Bayes for multi locus genome-wide association studies. *Heredity (Edinb)*. 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol*. 33, 531–537. doi: 10.1038/nbt.3207

Zhang, T. Z., Qian, N., Zhu, X. F., Chen, H., Wang, S., Mei, H. X., et al. (2013). Variations and transmission of qtl alleles for yield and fiber qualities in upland cotton cultivars developed in china. *PLoS ONE* 6:e57220. doi: 10.1371/journal.pone.0057220

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42:355. doi: 10.1038/ng.546

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 17 August 2018
doi: 10.3389/fpls.2018.01083

Single-Locus and Multi-Locus Genome-Wide Association Studies in the Genetic Dissection of Fiber Quality Traits in Upland Cotton (*Gossypium hirsutum* L.)

Chengqi Li*, Yuanzhi Fu, Runrun Sun, Yuanyuan Wang and Qinglian Wang

Collaborative Innovation Center of Modern Biological Breeding, School of Life Science and Technology, Henan Institute of Science and Technology, Xinxiang, China

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Jun Zhang,
Shandong Academy of Agricultural Sciences, China
Zhiying Ma,
Agricultural University of Hebei, China
Youlu Yuan,
Institute of Cotton Research (CAAS),
China

****Correspondence:***
Chengqi Li
lichq2010@126.com

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *16 March 2018*
Accepted: *04 July 2018*
Published: *17 August 2018*

Citation:
Li C, Fu Y, Sun R, Wang Y and Wang Q (2018) Single-Locus and Multi-Locus Genome-Wide Association Studies in the Genetic Dissection of Fiber Quality Traits in Upland Cotton (Gossypium hirsutum L.). Front. Plant Sci. 9:1083. doi: 10.3389/fpls.2018.01083

A major breeding target in Upland cotton (*Gossypium hirsutum* L.) is to improve the fiber quality. To address this issue, 169 diverse accessions, genotyped by 53,848 high-quality single-nucleotide polymorphisms (SNPs) and phenotyped in four environments, were used to conduct genome-wide association studies (GWASs) for fiber quality traits using three single-locus and three multi-locus models. As a result, 342 quantitative trait nucleotides (QTNs) controlling fiber quality traits were detected. Of the 342 QTNs, 84 were simultaneously detected in at least two environments or by at least two models, which include 29 for fiber length, 22 for fiber strength, 11 for fiber micronaire, 12 for fiber uniformity, and 10 for fiber elongation. Meanwhile, nine QTNs with 10% greater sizes (R^2) were simultaneously detected in at least two environments and between single- and multi-locus models, which include TM80185 (D13) for fiber length, TM1386 (A1) and TM14462 (A6) for fiber strength, TM18616 (A7), TM54735 (D3), and TM79518 (D12) for fiber micronaire, TM77489 (D12) and TM81448 (D13) for fiber uniformity, and TM47772 (D1) for fiber elongation. This indicates the possibility of marker-assisted selection in future breeding programs. Among 455 genes within the linkage disequilibrium regions of the nine QTNs, 113 are potential candidate genes and four are promising candidate genes. These findings reveal the genetic control underlying fiber quality traits and provide insights into possible genetic improvements in Upland cotton fiber quality.

Keywords: GWAS, multi-locus model, fiber quality, Upland cotton (*Gossypium hirsutum* L.), QTN, candidate gene

INTRODUCTION

Cotton produces a fine natural fiber that is an important raw material for the textile industry. In recent years, technology development in the textile industry has been more rapid than improvements in the quality of cotton fiber, resulting in an inability to meet the industry needs, which include stronger, thinner, and more regular cotton fibers. China is the largest cotton producing country in the world, with the yield of Chinese cotton cultivars being equal to or slightly higher than those developed in the USA and Australia. However, the fiber qualities of the Chinese cotton cultivars, especially fiber strength (FS), are not as good (Wang et al., 2009). Upland cotton (*Gossypium hirsutum* L.) (2n = 4x = 52), one of the 50 *Gossypium* species and the leading natural fiber crop, produces more than 95% of the total cotton because of its high yield and wide adaptability (Chen et al., 2007). Improving the fiber quality is a major breeding target in Upland cotton.

Traditional breeding methods play important roles in cotton breeding. Predecessors bred a number of high-quality resource materials by hybridization, backcrossing, and other means using high fiber quality genes from Sea Island cotton (*Gossypium barbadense*) (Liang, 1999; Zhang et al., 2012). However, there still exists a negative correlation between fiber quality and yield, and complex correlated relationships among fiber quality traits (Miller and Rawlings, 1967; Smith and Coyle, 1997), which leads to the consequences that yield and quality, and individual fiber quality index, could not be simultaneously improved using traditional breeding strategies. The application of molecular markers that are closely linked to or significantly associated with the target quantitative trait loci (QTLs), for marker-assisted selection (MAS), can transform traditional phenotypic selection into direct genotypic selection, thereby improving the selection efficiency (Lee, 1995; Mohan et al., 1997). Therefore, it is important to elucidate the molecular genetics of cotton fiber qualities using molecular marker technology.

Association mapping based on linkage disequilibrium (LD) is a powerful tool for dissecting the genetic bases of complex plant traits. In contrast to the traditional linkage mapping, association mapping can effectively associate genotypes with phenotypes in natural populations and simultaneously detect many natural allelic variations in a single study (Huang and Han, 2014). Its high resolution, cost efficiency, and non-essential pedigrees have allowed association mapping to be applied in the dissection of many important cotton phenotypes, such as yield and its components (Mei et al., 2013; Zhang et al., 2013; Jia et al., 2014; Qin et al., 2015), fiber quality (Abdurakhmonov et al., 2008, 2009; Zhang et al., 2013; Cai et al., 2014; Qin et al., 2015; Nie et al., 2016), early maturity (Li et al., 2016a), disease resistance (Mei et al., 2014; Zhao et al., 2014), salt resistance (Saeed et al., 2014; Du et al., 2016), plant architecture (Li et al., 2016b), and seed quality (Liu et al., 2015). All of those studies, however, were based on using a limited number of simple sequence repeat markers (SSRs). The genetic bases of the quantitative traits could not be fully revealed at the genome-wide level.

As there is wide application of high-density genotyping platforms, the development of numerous single nucleotide polymorphism markers (SNPs) makes it possible to dissect the genetic architecture of quantitative traits through the genome-wide association studies (GWASs). Presently, GWAS has been successfully employed for several major crops, such as rice (Spindel et al., 2016), maize (Xu et al., 2017), wheat (Zegeye et al., 2014), barley (Visioni et al., 2013), oat (Newell et al., 2011), rapeseed (Zhou et al., 2017), soybean (Zhang J. et al., 2015), peanut (Zhang et al., 2017), and sorghum (Morris et al., 2013). For cotton fiber quality, Su et al. (2016b) performed a GWAS of fiber quality traits using 355 Upland cotton accessions and 81,675 SNPs developed from specific-locus amplified fragment sequences. They detected 16, 10, and 7 SNPs significantly associated with fiber length (FL), FS, and fiber uniformity (FU), respectively. In the study by Islam et al. (2016), the fiber quality data and 6,071 SNPs generated through genotyping-by-sequencing and 223 SSRs of 547 recombinant inbred lines were used to conduct a GWAS. One QTL cluster associated with four fiber quality traits, which include short fiber content, FS, FL, and FU, on chromosome A7 was identified and validated. Additionally, using the first commercial high-density CottonSNP63K array, Gapare et al. (2017) identified 17 and 50 significant SNP associations for FL and fiber micronaire (FM), respectively. Sun et al. (2017) and Huang et al. (2017) detected 46 and 79 significant SNPs, respectively, associated with several fiber quality traits. The above studies allowed the unraveling of the genetic architecture of fiber quality traits in cotton at the genome-wide level. However, the GWAS performed was based on the single-locus models, such as the general linear model (GLM) and the mixed linear model (MLM) (Bradbury et al., 2007). Multiple tests require that the test number undergoes a Bonferroni correction. The typical Bonferroni correction is often too conservative, which results in many important loci associated with the target traits being eliminated because they do not satisfy the stringent criterion of the significance test.

The multi-locus models are better alternatives for GWASs because they do not require the Bonferroni correction, and thus more marker-trait associations may be identified. Recently, several new multi-locus GWAS models, such as multi-locus RMLM (mrMLM, Wang et al., 2016), fast multi-locus random-SNP-effect EMMA (FASTmrEMMA, Wen et al., 2017), and Iterative modified-Sure Independence Screening EM-Bayesian LASSO (ISIS EM-BLASSO, Tamba et al., 2017), were developed. In this study, several models, including the single-locus and multi-locus models, were simultaneously used for the GWAS of fiber quality traits in Upland cotton based on a recently developed CottonSNP80K array (Cai et al., 2017), and the candidate genes were further identified. The results provide an insight into the complicated genetic architecture of the fiber quality traits in Upland cotton and reveal the whole-genome quantitative trait nucleotides (QTNs) for MAS in future breeding programs.

MATERIALS AND METHODS

Plant Materials

A total of 169 Upland cotton accessions were examined in the present study, including 62 and 25 from ecological cotton-growing areas of the Yellow and Yangtze Rivers, respectively, in addition to 50 from Northwestern China, 22 from Northern China, and 10 from other countries (**Supplementary Table S1**). These accessions were elite cultivars originating in, or introduced

Abbreviations: FASTmrEMMA, fast multi-locus random-SNP-effect EMMA; FE, Fiber elongation; FL, Fiber length; FM, Fiber micronaire; FS, Fiber strength; FU, Fiber uniformity; GLM, general linear model; GWAS, Genome-wide association study; ISIS EM-BLASSO, Iterative modified-Sure Independence Screening EM-Bayesian LASSO; LD, Linkage disequilibrium; MAS, Marker-assisted selection; MLM, Mixed linear model; mrMLM, multi-locus RMLM; QTLs, Quantitative trait loci; QTN, quantitative trait nucleotide; SNP, Single-nucleotide polymorphism; SSR, Simple sequence repeat.

to, China. All accessions showed stable inheritances after many generations of self-pollination.

Experimental Design and Trait Investigation

All materials were planted in the two different ecological cotton-growing areas of China, the Yellow River (Xinxiang City, Henan Province) and Northwestern China (Shihezi City, Xinjiang Province) during 2012 and 2013. The experiment adopted a randomized complete block design with single row plots and two replications. In Xinxiang, 14–16 plants were arranged in each row, with a row length of 5 m and a row interval of 1.0 m. In Shihezi, 38–40 plants were arranged in each row, with a row length of 5 m and a row interval of 0.45 m. Local normal management was carried out for all activities. For descriptive purposes, the four environments, 2012 Xinxiang, 2013 Xinxiang, 2012 Shihezi, and 2013 Shihezi, are designated as E1, E2, E3, and E4, respectively.

Lint fiber samples of ∼15 g, taken from each row, were sent to the Fiber Quality Testing Center of the Institute of Cotton Research, Chinese Academy of Agricultural Sciences for the determination of fiber qualities (HVISPECTRUM, HVICC calibration level). Altogether, five fiber quality traits—FL (mm), FS (cN/Tex), FM, FU (%), and fiber elongation (FE, %), were investigated. To reduce environmental errors, the best linear unbiased predictors (BLUPs) for the five traits per genotype were estimated using the lme4 package (Bates et al., 2011). The BLUP values and single environments were used for the GWAS.

SNP Genotype Calling

Genomic DNA of each accession was extracted from young leaf tissues for genotyping using the DNAsecure Plant Kit (TIANGEN). A CottonSNP80K array containing 77,774 SNPs (Cai et al., 2017), which was recently developed based on the sequencing of "TM-1" (Zhang T. Z. et al., 2015) and the re-sequencing of 100 different cultivars in Upland cotton, with 5× coverage on an average (Fang et al., 2017), were applied to genotype the 169 accessions. The image files were saved and analyzed using the GenomeStudio Genotyping Module (v1.9.4, Illumina). All 77,774 SNPs corresponded to the three separate signal clusters, AA, AB, and BB. However, from an evolutionary point of view, the polyploid cotton originated from an interspecific hybridization event between A- and D-genome diploid species around 1–2 million years ago, and the two extant progenitor relatives diverged from a common ancestor around 5–10 million years ago (Wendel and Cronn, 2003). In addition, Upland cotton is a type of cross-pollinated allotetraploid crop with a 10–15% natural hybridization rate. Thus, some SNPs in Upland cotton could contain five genotypes (AAAA, AAAB, AABB, ABBB, and BBBB). When these genotyping signals gather $>$ 3 clusters, the automatic SNP calling can produce errors; therefore, we confirmed the genotypes of these loci using a manual adjustment method as described by Cai et al. (2017). Thus, a more accurate clustering file was produced to improve the genotyping efficiency levels for the samples.

Population Structure and LD Estimation

Only SNPs with minor allele frequencies $\geq$0.05 and integrities $\geq$50% were used for population structure and LD analyses. The population structure was assessed using ADMIXTURE software (Alexander et al., 2009). To explore the population structure of the tested accessions, the number of genetic clusters (k) was predefined as 2–10. This analysis provided the maximum likelihood estimates of the proportion of each sample derived from each of the k sub-populations, and the corresponding Q-matrix was obtained for the subsequent GWAS. To determine the mapping resolution for GWAS, an LD analysis was performed for Upland cotton accessions. Pair-wise LD values between markers were calculated as the squared correlation coefficient (r^2) of alleles using the GAPIT software (Lipka et al., 2012).

GWAS

The GWAS was performed using six models, including three single-locus models: GLM (Bradbury et al., 2007), MLM (Bradbury et al., 2007), and compressed mixed linear model [CMLM; (Zhang et al., 2010)], and three multi-locus models: mrMLM (Wang et al., 2016), FASTmrEMMA (Wen et al., 2017), and ISIS EM-BLASSO (Tamba et al., 2017). In short, the GLM corrects only the population structure; the MLM corrects both population structure and kinship relationship among individuals; and the CMLM is equivalent to the MLM when individuals are clustered into groups based on kinship and the ratio of polygenic to residual variances is fixed by genome scanning. The three multi-locus models include two steps. The first step is to select all the potentially associated SNPs. In the next step, all the selected SNPs are included into one model, then their effects are estimated by empirical Bayes, and finally all the non-zero effects are further evaluated using the likelihood ratio test. FASTmrEMMA whitens the covariance matrix of the polygenic matrix K and environmental noise. In ISIS EM-BLASSO, an iterative modified sure independence screening along with SCAD algorithm was used to select potentially associated SNPs. In the three single-locus GWASs, significant levels of marker-trait association were set at an adjusted *P*-value of 1/n, after the Bonferroni correction (Cai et al., 2017; Sun et al., 2017), where n was the total number of SNPs used in GWAS. The Manhattan plots were drawn using the R package qqman (Turner, 2014). In the three multi-locus GWASs, the critical *P*-values were set at 0.01, 0.005, and 0.01 for mrMLM, FASTmrEMMA, and ISIS EM-BLASSO, respectively, in the first step. In the second step, all the critical LOD scores for significance were set at 3.0. The SNPs that met the above standards were identified as significant trait-associated QTNs.

Identification of Candidate Genes

The R software package "LDheatmap" was used to determine the LD heatmaps surrounding the significant trait-associated QTNs. Based on the *G. hirsutum* "TM-1" genome (Zhang T. Z. et al., 2015), the genes within the LD decay distance on either side of the significant trait-associated SNPs were mined. To investigate the functions of these genes, RNA-seq datasets with two biological repetitions of 12 vegetative and reproductive tissues (root, stem, leaf, ovules from −3, −1, 0, 1, and 3 days post-anthesis, and fibers from 5, 10, 20, and 25 days post-anthesis) of *G. hirsutum* "TM-1," were downloaded from the NCBI SRA database under accession code PRJNA248163 (http://www.ncbi.nlm.nih.gov/sra/?term=PRJNA248163; Zhang T. Z. et al., 2015). Normalized fragments per kilobase of transcript per million fragments mapped (FPKM) values were calculated to indicate the expression levels of these genes. The average of the two biological replicates was recorded as the final FPKM value. A heatmap of the expression patterns—based on FPKM values—of genes was created using Mev 4.9 (Saeed et al., 2003). Further gene annotations were performed from several databases for non-redundant protein sequences (ftp://ftp.ncbi.nih.gov/blast/db/FASTA; Altschul et al., 1997), gene ontology (http://www.geneontology.org; Ashburner et al., 2000), Cluster of Orthologous Groups of proteins (http://www.ncbi.nlm.nih.gov/COG; Tatusov et al., 2000), and the Kyoto Encyclopedia of Genes and Genomes (ftp://ftp.genome.jp/pub/kegg/; Kanehisa et al., 2004).

RESULTS

Phenotypic Variations in Fiber Quality Traits

Phenotypic values for five fiber quality traits of the 169 accessions in four environments (**Supplementary Table S2**) were used for the variation analysis. The phenotypic evaluation revealed a broad variation range among accessions. Descriptive statistics of phenotypic variation for the five fiber quality traits are listed in **Table 1**. The mean FL were 27.90, 28.52, 29.23, and 29.08 mm, respectively, in the four experiments. The minimum FL was 22.43 mm in E2, and the maximum FL was 34.48 mm in E3. Analogously, the other four traits of FS, FM, FU, and FE, exhibited values in the range of 23.40–39.90 cN/Tex, 2.10–6.03, 78.10–88.90%, and 5.70–7.50%, with means of 29.03 cN/Tex, 4.53, 84.53, and 6.59%, respectively. The CV ranges for FL, FS, FM, FU, and FE in the four environments were 4.69–5.40%, 6.85–9.52%, 8.87–15.73%, 1.34–1.74%, and 0.91–3.88%, respectively, and the average CVs for the same were 4.96, 8.59, 11.18, 1.52, and 2.81%, respectively. These data indicated different degrees of diversity in fiber quality traits in the natural population. The frequency distributions of the phenotypes (**Figure 1**) showed that the fiber quality traits exhibited the genetic characteristics of quantitative traits with continuous distributions across different environments. Furthermore, some of the traits exhibited multimodal or partial distributions, suggesting that the main effect genes/QTNs related to the target traits could exist in cotton genome.

TABLE 1 | Descriptive statistics of phenotypic values of five fiber quality traits in four environments.

Trait[a]	Env[b]	Min	Max	Average	Std	CV (%)
FL (mm)	E1	23.18	31.32	27.90	1.36	4.86
	E2	22.43	33.06	28.52	1.39	4.87
	E3	24.20	34.48	29.23	1.58	5.40
	E4	24.91	34.40	29.08	1.36	4.69
FS (cN/Tex)	E1	23.80	37.80	28.16	2.68	9.52
	E2	23.50	38.70	30.14	2.82	9.35
	E3	23.40	35.20	28.23	1.93	6.85
	E4	24.30	39.90	29.58	2.56	8.66
FM	E1	3.67	6.00	5.05	0.45	8.87
	E2	3.38	5.84	4.96	0.46	9.19
	E3	2.10	6.03	3.93	0.62	15.73
	E4	2.59	5.21	4.17	0.46	10.92
FU (%)	E1	79.50	86.15	83.28	1.20	1.45
	E2	81.20	87.70	85.08	1.14	1.34
	E3	78.10	88.30	85.12	1.48	1.74
	E4	80.90	88.90	84.64	1.30	1.54
FE (%)	E1	6.00	7.35	6.57	0.23	3.56
	E2	6.50	6.90	6.71	0.06	0.91
	E3	5.80	7.50	6.80	0.26	3.88
	E4	5.70	6.80	6.29	0.18	2.90

[a] *FL, fiber length; FS, fiber strength; FM, fiber micronaire; FU, fiber uniformity; FE, fiber elongation.*

[b] *E1, E2, E3, and E4 indicate four environments: 2012 Xinxiang, 2013 Xinxiang, 2012 Shihezi, and 2013 Shihezi, respectively.*

Characteristics of Polymorphic SNPs

The genotypes of 169 accessions were examined using Illumina GenomeStudio software. Only the SNPs with minor allele frequencies ≥0.05, and integrities ≥50% in the population, were used for screening polymorphic loci. Thus, 53,848 high-quality SNPs were obtained out of 77,774. Their characteristics are summarized in **Table 2** and **Supplementary Figure S1**. These SNPs were not evenly distributed across the *G. hirsutum* genome, and there were 28,454 and 25,394 SNPs in the A and D subgenomes, respectively. The average marker density was approximately one SNP per 38.02 kb. In the A subgenome, chromosome A6 had the most markers (2,982), with a marker density of one SNP per 34.60 kb, and A4 had the least markers (1,050), with a marker density of one SNP per 59.92 kb. In the D subgenome, chromosome D6 had the most markers (3,128), with a marker density of one SNP per 20.55 kb, and D4 had the least markers (1,040), with a marker density of one SNP per 49.48 kb. The polymorphism information content values ranged from 0.255 to 0.309 among chromosomes, and the mean polymorphism information content values of the A and D subgenomes were 0.285 and 0.284, respectively.

Population Structure and LD

To estimate the number of sub-populations in the population of 169 Upland cotton accessions, a population structure analysis was performed using the 53,848 SNPs. The results indicated that the minimum number of cross-validation errors was $k = 6$, which was thus determined to be the optimum k; and the testing accessions could be separated into six sub-populations (**Figure 2A**). The varietal population in this study was considered to be not highly structured and could be used for further association mapping. Thus, the corresponding Q-matrix from $k = 6$ was obtained for the subsequent GWAS. An LD analysis showed that the average LD decay distance for each of the 26

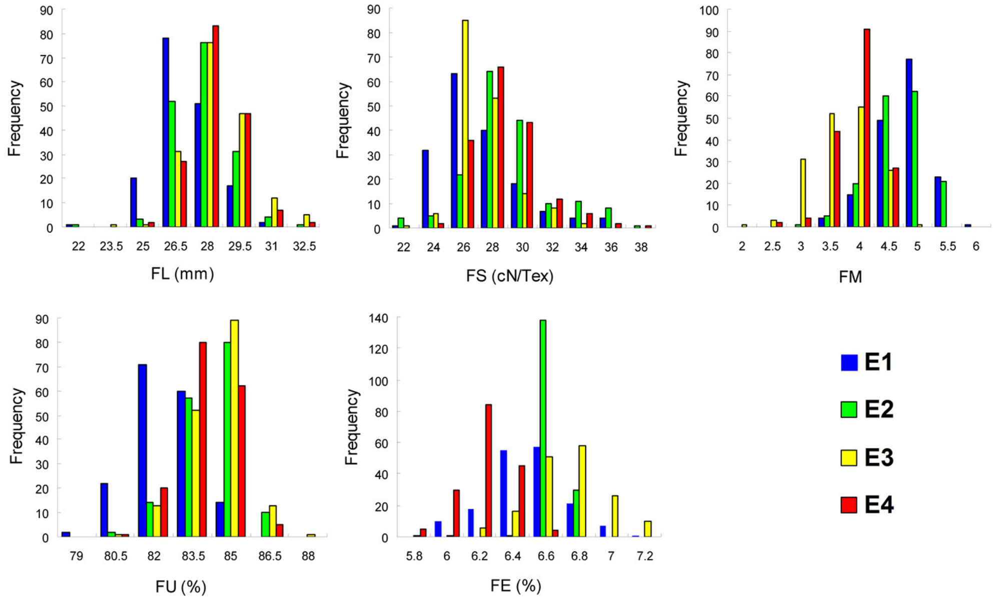


FIGURE 1 | Frequency of the five fiber quality traits in 169 Upland cotton accessions. FL, fiber length; FS, fiber strength; FM, fiber micronaire; FU, fiber uniformity; FE, fiber elongation; E1, E2, E3, and E4 indicate four environments: 2012 Xinxiang, 2013 Xinxiang, 2012 Shihezi, and 2013 Shihezi, respectively.

chromosomes ranged from 38.56 to 669.65 kb, and the average LD decay distance of all of the chromosomes (i.e., Upland cotton genome) was estimated to be 444.99 kb, with half of the maximum of mean r^2-values (**Figure 2B**).

GWAS for Fiber Quality Traits

Three single-locus GWAS models: GLM, MLM, and CMLM, and three multi-locus GWAS models: mrMLM, FASTmrEMMA, and ISIS EM-BLASSO, were used to identify the marker–trait associations. In single-locus GWAS, the SNPs with $-\log_{10}P \geq 4.73$ ($P = 1/53,848$) were regarded as significant trait-associated SNPs. In multi-locus GWAS, the SNPs with LOD scores greater than 3.0 were regarded as significant trait-associated SNPs. Based on these criteria, 342 QTNs for fiber quality traits were detected using the values of individual environments (including BLUP) and the six models (**Supplementary Table S3**). To obtain reliable results, only the QTNs simultaneously detected in at least two environments, or by at least two models (either single-locus or multi-locus), were displayed. Finally, 84 QTNs controlling fiber quality traits were obtained (**Table 3**).

Based on FL, 29 QTNs were detected. Five SNPs, including TM10103, TM10107, TM10110, TM10764, and TM39339, located on A5 and A11, were significantly associated with the E2, E3, and/or BLUP values by a single-locus GWAS, and this explained 11.76–16.67% of the phenotypic variations. 22 SNPs, including TM119, TM3930, and TM4397, located on A1, A2, A5, A6, A7, A8, A9, A10, A11, A12, D1, D5, D10, and D13, were significantly associated with the E1, E2, E3, E4, and/or BLUP values by a multi-locus GWAS, and this explained 3.14–23.57% of the phenotypic variations. Two SNPs, TM57840, and TM80185, respectively located on D5 and D13, were significantly associated with the E1, E2, E3, and/or BLUP values by both single-locus and multi-locus GWAS, which explained 10.35–14.46% of phenotypic variations in single-locus GWAS and 3.94–36.66% in multi-locus GWAS.

Based on FS, 22 QTNs were detected. Five SNPs, including TM10764, TM14418, TM14424, TM20073, and TM21123, located on A5, A6, and A7, were significantly associated with the E1, E2, E3, E4, and/or BLUP values by a single-locus GWAS, thus explaining 7.56–15.16% of the phenotypic variations. Additionally, 12 SNPs, including TM5639, TM10540, and TM29912, located on A2, A5, A8, A9, A12, D1, D5, D9, and D10, were significantly associated with the E1, E2, E3, E4, and/or BLUP values by a multi-locus GWAS, thus explaining 1.37–25.24% of the phenotypic variations. Five SNPs, including TM1386, TM5421, TM14462, TM21135, and TM79685, respectively located on A1, A2, A6, A7, and D12, were significantly associated with the E1, E2, E3, E4, and/or BLUP values by both single-locus and multi-locus GWASs, and this explained 8.81–11.64% of the phenotypic variations in the single-locus GWAS and 6.32–23.95% in the multi-locus GWAS.

Based on FM, 11 QTNs were detected. Two SNPs, TM10764 and TM18615, respectively located on A5 and A7, were significantly associated with the E1, E2 and/or BLUP values by a single-locus GWAS, and this explained 3.49–12.24% and 10.74–12.04% of the phenotypic variations. Five SNPs, TM22010, TM33781, TM42632, TM55481, and TM57773, located on A8, A10, A12, D4, and D5, respectively, were significantly associated with the E1, E2, E3, and/or BLUP values by a multi-locus GWAS, thus explaining 0.96–10.54% of the phenotypic variations. Four SNPs, TM18616, TM19501, TM54735, and TM79518, located on A7, D3, and D12, were significantly associated with the E1, E2, E3, and/or BLUP values by both single-locus and multi-locus

GWASs, thus explaining the phenotypic variations of 10.94–12.72% in the single-locus GWAS and 5.70–53.97% in the multi-locus GWAS.

Based on FU, 12 QTNs were detected. One SNP, TM41077, located on A12, was significantly associated with the E1 and BLUP values by a single-locus GWAS, and this explained 11.38–11.64% of the phenotypic variations. Eight SNPs, including TM18205, TM19379, and TM43826, located on A6, A7, A13, D2, D5, D8, and D10, were significantly associated with the E1, E2, E3, E4, and/or BLUP values by a multi-locus GWAS, thus explaining 2.13–24.32% of the phenotypic variations. Three SNPs, TM11317, TM77489, and TM81448, respectively located on A5, D12, and D13, were significantly associated with the E1, E4, and/or BLUP values by both single-locus and multi-locus GWASs, thus explaining the phenotypic variations of 10.28–14.53% in the single-locus GWAS and 5.29–26.18% in the multi-locus GWAS.

TABLE 2 | Summary of the SNPs in 26 chromosomes of *Gossypium hirsutum*.

Chr.	Chr. size (kb)	No. of SNPs	SNP density (kb/SNP)	Polymorphism information content value
A1	99884.70	2371	42.13	0.301
A2	83447.91	1392	59.95	0.283
A3	100263.00	1744	57.49	0.277
A4	62913.77	1050	59.92	0.284
A5	92047.02	2575	35.75	0.300
A6	103170.40	2982	34.60	0.294
A7	78251.02	2125	36.82	0.290
A8	103626.30	2870	36.11	0.281
A9	74999.93	2439	30.75	0.277
A10	100866.6	2037	49.52	0.274
A11	93316.19	1915	48.73	0.280
A12	87484.87	2051	42.65	0.283
A13	83159.57	2903	28.65	0.285
D1	61456.01	1860	33.04	0.284
D2	67284.55	2371	28.38	0.307
D3	46690.66	1394	33.49	0.276
D4	51454.13	1040	49.48	0.282
D5	61933.05	1595	38.83	0.286
D6	64294.64	3128	20.55	0.275
D7	55312.61	2708	20.43	0.300
D8	65894.14	2273	28.99	0.309
D9	50995.44	2227	22.90	0.255
D10	63374.67	1734	36.55	0.290
D11	66087.77	1408	46.94	0.274
D12	59109.84	1968	30.04	0.273
D13	60534.30	1688	35.86	0.280

Based on FE, 10 QTNs were detected. Nine SNPs, including TM13701, TM37254, and TM42798,r located on A6, A11, A12, A13, D1, D7, D10, and D11, were significantly associated with the E1, E2, E3, E4, and/or BLUP values by a multi-locus GWAS, thus explaining 3.59–34.06% of the phenotypic variations. One SNP, TM47772, located on D1, was significantly associated with the E1 and/or E3 values by both single-locus and multi-locus GWASs, thus explaining 14.55% of the phenotypic variations in the single-locus GWAS and 4.54–19.68% in the multi-locus GWAS.

Identification and Expression of Candidate Genes for Fiber Quality

Among the 84 QTNs, nine QTNs—TM80185 (D13) associated with FL, TM1386 (A1) and TM14462 (A6) associated with FS, TM18616 (A7), TM54735 (D3), and TM79518 (D12) associated with FM, TM77489 (D12) and TM81448 (D13) associated with FU, and TM47772 (D1) associated with FE, were simultaneously detected in at least two environments, and by both single-locus and multi-locus GWASs (**Supplementary Figures S2–S6**), indicating that they were more stable. Considering the LD decay distance of the Upland cotton population used in this study, the regions within 400-kb on either side of the nine QTNs were used for the further identification of candidate genes. The LD analysis showed that a high LD level existed among the SNPs within 400-kb upstream and downstream of the nine QTNs in

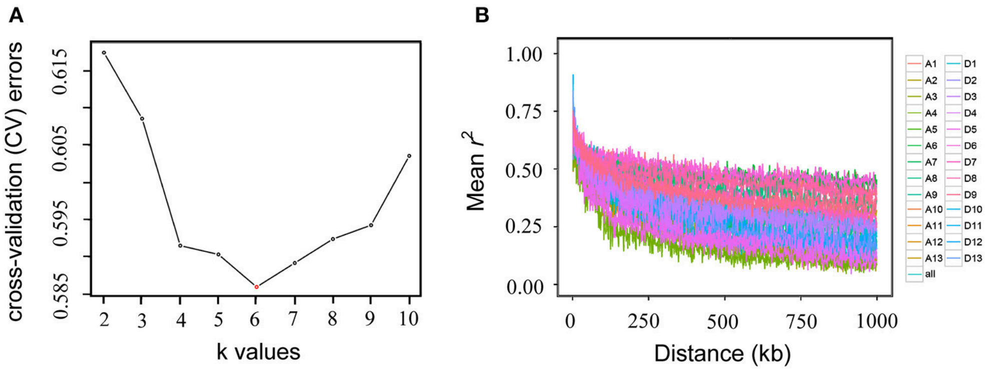


FIGURE 2 | Population structure **(A)** and linkage disequilibrium decay **(B)** of 169 Upland cotton accessions. The accessions were divided into six sub-populations (the minimum number of cross-validation errors occurred when $k = 6$). Genome-wide average linkage disequilibrium decay was estimated in each of the 26 chromosomes and in all chromosomes.

TABLE 3 | Significant fiber quality trait-associated QTNs simultaneously detected in at least two environments or by at least two models.

Trait[a]	SNP	Position (bp)	Alleles	Chr.	Single-locus GWAS			Env[d]	Multi-locus GWAS			Env[d]
					–log10*P*	R^2 (%)[b]	Model[c]		LOD	R^2 (%)[b]	Model[c]	
FL	TM10103	3462099	T/G	A5	5.10–5.41	12.02–13.71	G	E2, E3, Blup				
	TM10107	3488471	A/G	A5	4.82–6.71	14.02–16.67	G, M, C	E2, E3, Blup				
	TM10110	3505884	C/G	A5	5.09–5.42	11.76–13.47	G	E2, E3, Blup				
	TM10764	15474110	A/G	A5	4.85–4.85	14.69–14.69	M, C	Blup				
	TM39339	81827835	T/A	A11	5.16–5.22	12.19–13.28	G	E2, Blup				
	TM119	2771681	T/C	A1					3.36–4.06	7.81–10.32	MR, I	E1
	TM3930	3420685	T/C	A2					5.62–6.50	10.18–11.20	MR, I	E1
	TM4397	13758183	A/G	A2					6.62–8.35	7.49–10.46	F, MR, I	E2
	TM10319	5874999	A/C	A5					3.25–3.26	7.01–10.98	MR, I	E1
	TM10453	8801892	A/G	A5					4.69–6.05	5.66–10.20	MR, I	E1, E2
	TM10454	8829389	T/C	A5					6.70–7.87	9.33–9.56	I	E3, Blup
	TM10976	21365948	A/G	A5					4.55–4.71	17.78–23.57	MR, I	E4
	TM18271	99455219	T/C	A6					3,79–6.14	5.47–6.56	F, I	E3
	TM19208	14663920	T/G	A7					3.17–3.92	13.87–14.71	MR, I	E2
	TM27227	60265814	A/G	A8					3.73–3.78	11.70–12.84	I	E3, E4
	TM28899	77587957	A/G	A8					5.06–6.10	10.45–15.84	MR, I	E1
	TM31735	41160852	T/C	A9					4.06–6.05	3.33–5.15	F, I	E1, Blup
	TM33839	1954843	A/G	A10					3.38–4.83	5.42–11.24	MR, I	E1, Blup
	TM37371	13595750	A/G	A11					4.70–6.37	3.14–4.03	MR, I	E2
	TM42899	81101484	T/A	A12					4.69–5.08	5.07–5.35	MR, I	E1
	TM47849	1787530	T/C	D1					3.42–6.08	3.35–7.98	F	E1, Blup
	TM57343	16937262	A/G	D5					5.14–6.56	5.34–8.33	MR, I	E3
	TM58061	32206837	A/C	D5					3.78–4.92	4.43–6.03	MR, I	E2
	TM58758	59288520	T/C	D5					3.25–5.70	6.30–12.46	MR, I	E1, E3
	TM75008	58123221	T/C	D10					4.36–6.86	5.50–6.46	F, I	Blup
	TM75026	58453007	A/G	D10					3.33–4.19	4.17–5.60	F, MR, I	E3
	TM81924	55032877	A/C	D13					3.33–4.95	5.70–8.16	MR, I	E2
	TM57840	30021662	A/G	D5	4.88	10.35	G	E2	8.45–10.39	28.29–36.66	MR, I	E2
	TM80185	3106437	A/G	D13	4.96–5.42	13.08–14.46	G	E1, E3	3.28–6.88	3.94–8.88	F, I	E1, Blup
FS	TM10764	15474110	A/G	A5	4.94–6.26	10.48–15.16	G, M, C	E1, E2, E3, Blup				
	TM14418	30941574	T/C	A6	4.80–4.85	7.56–9.05	G	E1, E3				
	TM14424	31197620	T/C	A6	4.85–4.96	7.71–9.24	G	E1, E3				
	TM20073	28183664	T/G	A7	5.72–5.82	10.83–11.65	G	E3, Blup				
	TM21123	70595913	A/G	A7	4.83–5.30	8.61–8.67	G	E4, Blup				
	TM5639	80304252	T/C	A2					3.24–4.95	10.82–25.24	MR, I	E4
	TM10540	11387213	T/G	A5					3.19–7.74	5.03–15.89	MR, I	E2, Blup
	TM29912	101941614	T/A	A8					3.10–5.43	2.67–3.93	F, I	E3, Blup
	TM33273	65822047	A/C	A9					5.44–5.75	17.75–22.20	MR, I	E2
	TM42806	78617984	A/G	A12					3.23–3.52	11.50–12.06	MR, I	E2
	TM47849	1787530	T/C	D1					4.09–5.33	1.37–6.90	F, I	E2, E3, Blup
	TM57401	18161586	A/G	D5					5.48–8.15	8.96–10.25	MR, I	E1
	TM58758	59288520	T/C	D5					4.38–6.25	3.85–19.66	MR, I	E2, E3, Blup
	TM58839	61435904	T/G	D5					3.03–4.22	3.13–7.34	F, I	E1, Blup
	TM72234	38761458	A/G	D9					3.64–4.23	10.77–15.12	MR, I	E4
	TM74995	57945654	A/T	D10					4.21–4.23	14.26–19.43	MR, I	E4
	TM75026	58453007	A/G	D10					3.64–5.53	3.94–7.84	I	E3, Blup
	TM1386	41010954	T/C	A1	5.49–5.59	9.30–10.14	G	E1, Blup	5.21	23.95	I	E2

(Continued)

TABLE 3 | Continued

Trait[a]	SNP	Position (bp)	Alleles	Chr.	Single-locus GWAS			Env[d]	Multi-locus GWAS			Env[d]
					–log10*P*	R^2 (%)[b]	Model[c]		LOD	R^2 (%)[b]	Model[c]	
	TM5421	75968294	A/G	A2	4.87	10.20	G	E4	4.27–5.68	8.46–9.59	MR, F, I	E4
	TM14462	32121709	T/C	A6	4.83–5.68	10.27–11.64	G	E1, E2, Blup	5.03	6.32	I	E2
	TM21135	70682969	A/G	A7	5.02	10.09	G	E3	5.27–5.63	12.14–20.86	MR, I	E3
	TM79685	53877369	T/G	D12	4.82	8.81	G	E1	5.40	9.23	I	E1
FM	TM10764	15474110	A/G	A5	4.75–5.35	3.49–12.24	G, M, C	E2, Blup				
	TM18615	3643524	T/C	A7	5.04–5.80	10.74–12.04	G, M, C	E1				
	TM22010	5162186	A/T	A8					3.20–5.33	4.71–8.20	F, MR, I	E3
	TM33781	65693	A/G	A10					3.48–7.61	4.75–10.54	MR, I	E1, Blup
	TM42632	75299391	T/C	A12					3.02–3.39	3.12–3.94	F, I	E1
	TM55481	2866176	A/G	D4					3.00–5.66	0.96–3.34	MR, I	E1, E2, Blup
	TM57773	27544038	A/G	D5					3.30–4.56	2.93–3.96	I	E1, Blup
	TM18616	3646710	A/C	A7	5.11–5.52	10.94–12.16	G, M, C	E1	7.55–8.78	22.06–39.50	MR, I	E1, Blup
	TM19501	21060224	A/G	A7	4.84	11.05	G	E2	3.50	13.06	MR	E2
	TM54735	30908501	T/C	D3	4.76	11.33	G	Blup	4.00–9.64	5.70–16.76	F, MR, I	E3, Blup
	TM79518	51416454	T/G	D12	4.89–5.64	11.79–12.72	G, M, C	E3, Blup	5.44–8.17	27.75–53.97	MR, I	E3, Blup
FU	TM41077	26645691	G/C	A12	4.75–4.85	11.38–11.64	G	E1, Blup				
	TM18205	98260650	T/C	A6					3.36–4.77	7.64–10.52	MR, I	E1
	TM19379	18309921	T/C	A7					3.67–6.80	2.13–3.57	I	E3, Blup
	TM43826	15282624	A/G	A13					3.90–5.05	8.07–8.57	MR, I	Blup
	TM51438	21650323	A/G	D2					3.79–5.63	8.19–13.93	MR, I	Blup
	TM57831	29951748	A/G	D5					3.21–5.27	5.68–13.73	MR, I	Blup
	TM58758	59288520	T/C	D5					5.47–6.05	17.07–17.26	I	E2, E3
	TM67147	4674102	T/C	D8					3.07–3.13	9.82–15.67	MR, I	E1
	TM74995	57945654	A/T	D10					4.21–8.56	14.26–24.32	MR, I	E4, Blup
	TM11317	28285041	A/G	A5	5.09	12.82	G	Blup	4.70	7.79	F	Blup
	TM77489	3329594	T/C	D12	4.88	13.36	G	Blup	3.74–4.13	5.29–6.66	F	E1, Blup
	TM81448	45426771	C/G	D13	4.76–6.35	10.28–14.53	G, M, C	E4, Blup	7.85	26.18	MR	E4
FE	TM13701	2630501	T/C	A6					4.22–4.47	20.05–20.68	MR, I	Blup
	TM37254	7081938	A/G	A11					4.23–4.74	6.05–6.06	F, I	E2
	TM42798	78429684	C/G	A12					3.39–5.30	4.81–7.21	F	E1, E3
	TM43034	84964849	A/C	A12					3.66–5.80	15.22–15.30	MR, I	E1
	TM43327	3481958	A/G	A13					3.38–4.24	9.21–10.04	MR, I	E2
	TM48070	5563241	A/G	D1					3.25–4.10	3.59–5.91	MR, I	Blup
	TM63323	4045155	T/C	D7					4.35–5.89	32.47–34.06	MR, I	E4
	TM74999	57965498	A/G	D10					4.17–4.77	7.41–8.20	MR, F	E4
	TM77062	58739941	A/C	D11					4.13–4.34	17.94–21.38	MR, I	E2
	TM47772	723752	T/C	D1	5.68	14.55	G	E3	3.34–7.75	4.54–19.68	MR, F, I	E1, E3

[a] *FL, fiber length; FS, fiber strength; FM, fiber micronaire; FU, fiber uniformity; FE, fiber elongation.*

[b] *R^2 (%) means phenotypic variation explained by marker.*

[c] *G, M, C, MR, F, and I represent GLM, MLM, CMLM, mrMLM, FASTmrEMMA, and ISIS EM-BLASSO, respectively.*

[d] *E1, E2, E3, E4, and Blup indicate 2012 Xinxiang, 2013 Xinxiang, 2012 Shihezi, 2013 Shihezi, and best linear unbiased predictor, respectively.*

D13 (**Figure 3A**) for FL, A1 (**Figure 3B**) and A6 (**Figure 3C**) for FS, A7 (**Figure 3D**), D3 (**Figure 3E**), and D12 (**Figure 3F**) for FM, D12 (**Figure 3G**) and D13 (**Figure 3H**) for FU, and D1 (**Figure 3I**) for FE. Multiple LD blocks were included in almost all of the LD regions except those in A6 (**Figure 3C**). As a result, 455 genes were around the above nine QTNs. The normalized FPKM values of 455 genes, representing their expression levels, are displayed in **Supplementary Table S4**. To investigate which genes were responsible for fiber quality, only those genes that presented greater expression levels in ovules and/or fiber during their developmental stages, while being less expressed in root, stem, and leaf, were used for further functional analyses. Thus, 113 genes, marked in bold in **Supplementary Table S4**, were obtained. A heatmap of the expression patterns of these genes with hierarchical clustering based on FPKM values is shown in **Figure 4**. Considering that the five fiber quality traits are directly

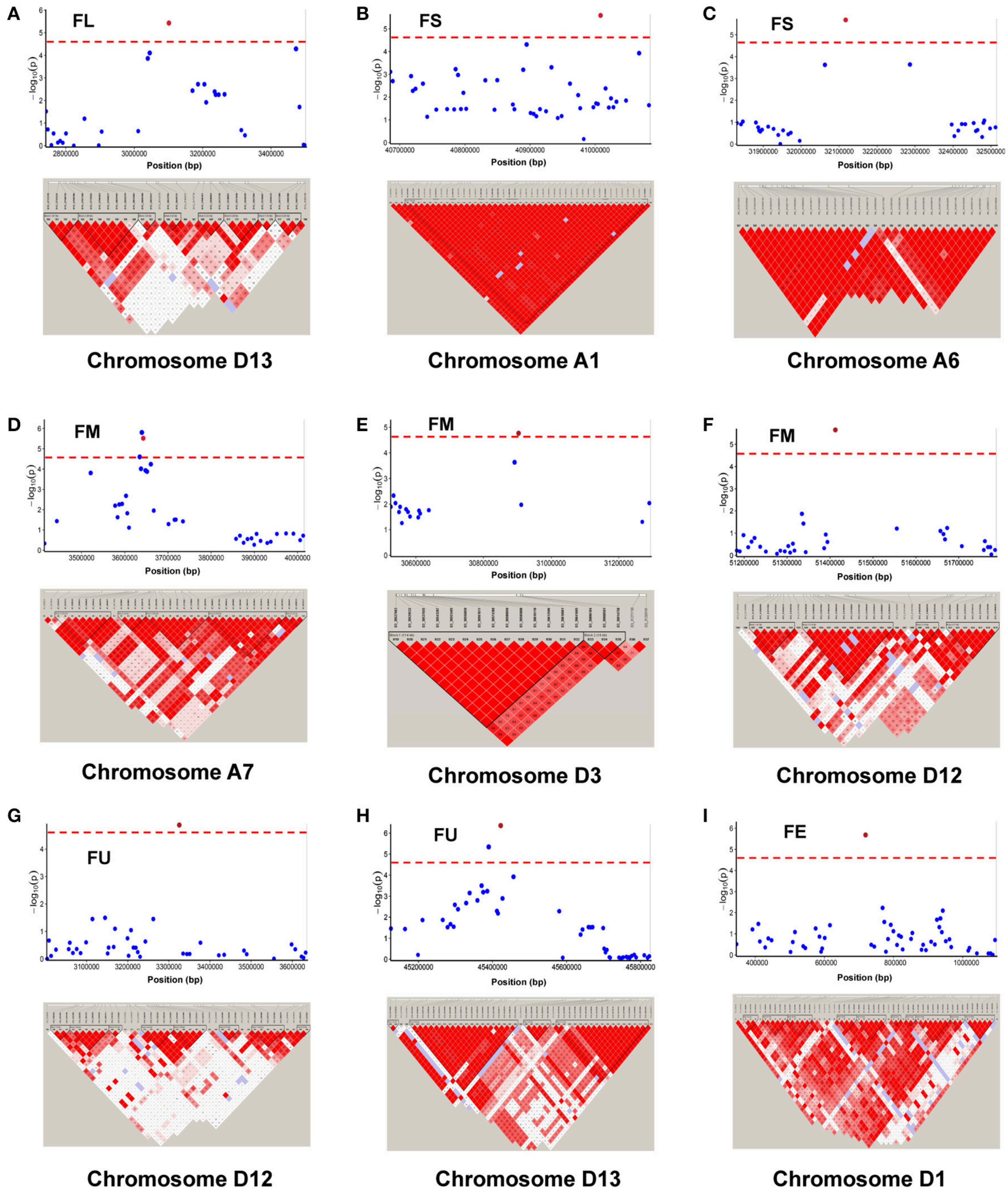


FIGURE 3 | Genomic location of nine QTNs simultaneously detected in at least two environments, by both single-locus GWAS and multi-locus GWAS, and LD heatmaps surrounding nine QTNs for **(A)** fiber length (FL) on chromosome D13, **(B,C)** fiber strength (FS) on chromosomes A1 and A6, **(D–F)** fiber micronaire (FM) on chromosomes A7, D3, and D12, **(G,H)** fiber uniformity (FU) on chromosome D12 and D13, and **(I)** fiber elongation (FE, %) on chromosome D1.

related to fiber development and are significantly positively correlated with each other, these genes were merged into a group for a systematic summary according to the functional annotation from the non-redundant protein, gene ontology, Cluster of Orthologous Groups of proteins, and the Kyoto Encyclopedia of Genes and Genomes analyses (**Supplementary Table S5**). These 113 genes could be classified into 10 categories (**Figure 5**), which include 9 in "Cellular component/cell division" (A), 19 in "Substance transport and metabolism" (B), 19 in "RNA Transcription" (C), 11 in "Translation, ribosomal structure and biogenesis" (D), 6 in "Defense/resistance-responsive" (E), 3 in "Post-translational modification, protein turnover, chaperones"

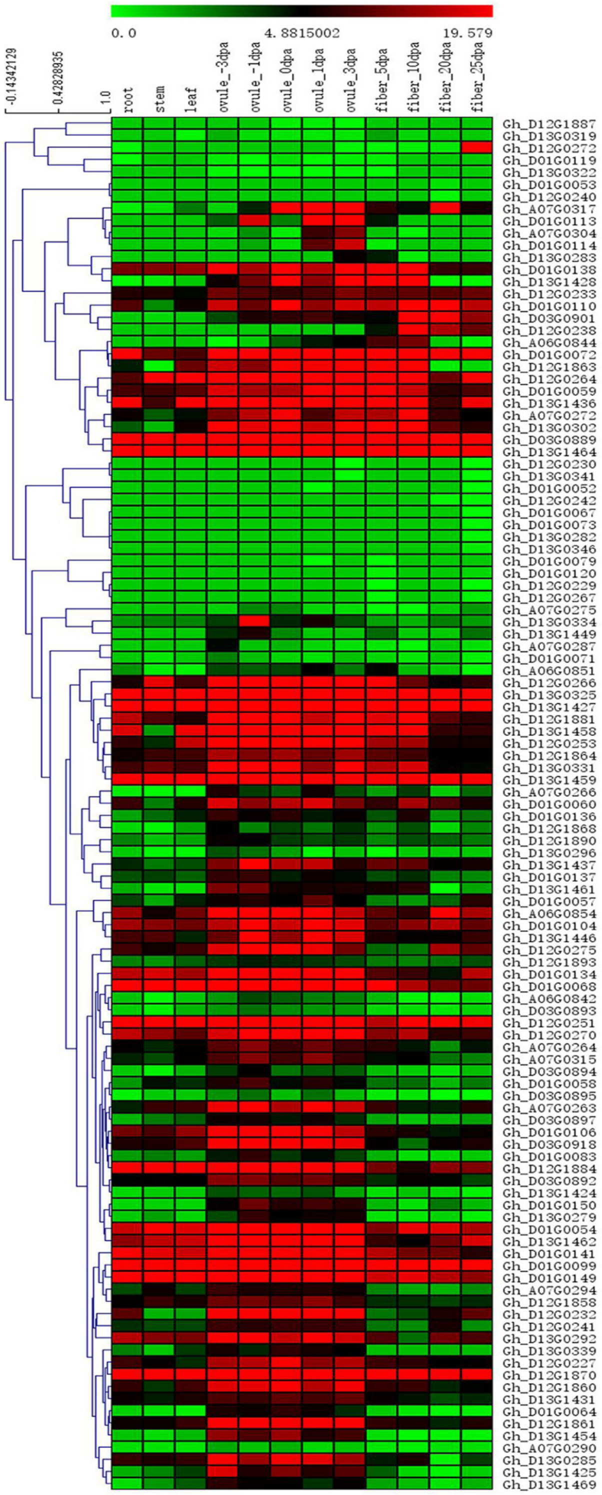


FIGURE 4 | Heatmap of expression patterns of 113 genes with hierarchical clustering based on FPKM values. These genes presented higher expression levels in ovules and/or fiber during their developmental stages, while being less expressed in root, stem, and leaf. The values in the horizontal color bar are automatically generated in Mev 4.9 according to the FPKM values; red indicates high expression, and green indicates low expression.

(F), 2 in "Energy production and conversion" (G), 19 in "Putative and uncharacterized proteins" (H), 23 in "General function prediction only" (I), and 2 in "Function unknown" (J). Several promising candidate genes were found through further bioinformatics analyses. *Gh_D13G1461* is homologous to Arabidopsis *AT1G50660*, which is the predicted protein sequence for the *BRANCHLESS TRICHOMES* gene, a key positive regulator of trichome branching (Marks et al., 2009; Kasili et al., 2015). *Gh_D12G0232* is homologous to Arabidopsis *AT2G03500*, which encodes a nuclear localized member of the *MYB* family of transcriptional regulators. The *MYB* transcription factor plays a role in cotton fiber and trichome development (Machado et al., 2009). Cellulose is the main component of cotton fiber. *Gh_D01G0052* and *Gh_D12G0240* are both homologous with Arabidopsis *AT1G09790*, which is annotated as a COBRA-like protein 6 precursor. In *Arabidopsis thaliana*, the COBRA is involved in determining the orientation of cell expansion, playing an important role in cellulose deposition (Roudier et al., 2005). Thus, the four genes might be promising candidate genes for improving the fiber quality.

DISCUSSION

Large Numbers of High-Quality SNPs Ensure Effective GWAS in Cotton

Association mapping is a powerful tool in dissecting the genetic basis of plant complex traits. Prior to the availability of next-generation sequencing techniques; however, SSR markers were mainly used to detect molecular markers associated with the target traits. Due to a limited number of markers, the genetic basis of the quantitative traits could not be fully revealed at the genome-wide level. With the wide application of high-density genotyping platforms, the development of numerous SNPs makes it possible to perform GWASs of the genetic bases of complex traits. In cotton, the SNPs developed from next-generation sequencing methods, such as specific-locus amplified fragment sequencing and genotyping-by-sequencing, were used to perform GWASs for lint percentage (Su et al., 2016a), fiber quality (Islam et al., 2016; Su et al., 2016b), early maturity (Su et al., 2016c), and *Verticillium* wilt resistance (Li T. et al., 2017). Furthermore, the first commercial high-density CottonSNP63K array, developed from 13 different discovery sets that represent a diverse range of *G. hirsutum* germplasm, as well as five other species, provided a new resource for the genetic dissection of cotton's quantitative traits (Hulse-Kemp et al., 2015). Presently, based on the CottonSNP63K array, the GWASs have been performed to unravel the agronomically and economically important traits in cotton, including yield components, fiber quality, growth period, plant height, and stomatal conductance (Gapare et al., 2017; Huang et al., 2017; Sun et al., 2017). Compared with CottonSNP63K, the recently developed CottonSNP80K array is more useful for dissecting the genetic architecture of important traits in Upland cotton because the SNP loci in the array benefited from the whole-genome sequencing of *G. hirsutum* acc. TM-1

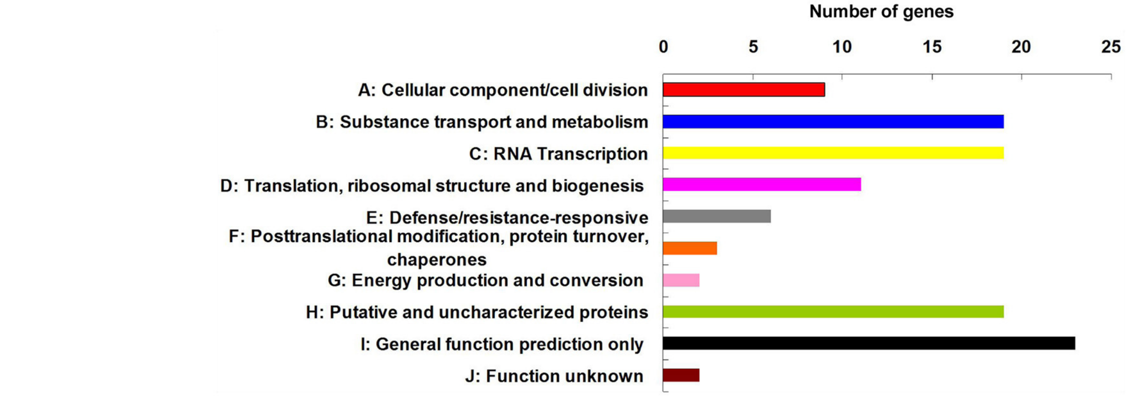


FIGURE 5 | Functional classification of 113 candidate genes, which presented higher expression levels in ovules and/or fiber during the stages of their development, while being less expressed in root, stem, and leaf.

(Zhang T. Z. et al., 2015) and 1,372,195 intraspecific non-unique SNPs identified by the re-sequencing of *G. hirsutum* accessions (Fang et al., 2017). In addition, each SNP marker in the CottonSNP80K array is addressable, which avoids the disturbances caused by homeologous/paralogous genes. The diverse application tests indicate that CottonSNP80K played important roles in germplasm genotyping, varietal verification, functional genomics studies, and molecular breeding in cotton (Cai et al., 2017). In this study, 53,848 high-quality SNPs out of 77,774 from the CottonSNP80K array, accounting for 69.24% of all loci, were screened in our experimental accessions. The large number of high-quality SNPs will be very conducive to unravel the genetic architecture of the target traits through GWASs.

Combining Single- and Multi-Locus GWASs Can Improve the Power and Robustness of GWAS

With the development of molecular quantitative genetics, a large number of association mapping methods have emerged for the genetic dissection of complex traits in plants (Feng et al., 2016). However, the methods used in most of the previous studies are single-locus analysis approaches based on a fixed-SNP-effect mixed linear model under a polygenic background and population structure controls. These methods require a Bonferroni correction for multiple tests. To control the experimental error at a genome-wide level of 0.05, the significance level for each test should be adjusted by 0.05/n (n is the total number of SNPs). The use of stringent probability thresholds reduces the risk of accepting false positives but does not reduce the risk of rejecting true positives caused by setting the very high thresholds. Multi-locus models, such as Bayesian LASSO (Yi and Xu, 2008), penalized Logistic regression (Hoggart et al., 2008), adaptive mixed LASSO (Wang et al., 2010), and EBAYES LASSO (Wen et al., 2015), can improve the efficiency and accuracy of QTL detection in GWAS. An obvious advantage of these models is that no Bonferroni correction is required because of the multi-locus nature. In particular, several recently developed multi-locus models, including mrMLM (Wang et al., 2016), FASTmrEMMA (Wen et al., 2017), and LASSO (ISIS EM-BLASSO) (Tamba et al., 2017), have been demonstrated as having the highest power and accuracy levels for QTL detection when compared with some former methods. As the inheritance of quantitative traits is complex and the number of markers is several times larger than the sample sizes, it is necessary to simultaneously use multiple methods for GWAS. Several examples can be found in previous studies. Li H. G. et al. (2017) performed a GWAS to reveal the genetic control underlying the branch angle in rapeseed by simultaneously using a single-locus model, MLM, and a multi-locus model, mrMLM. As a result, more than 55% of the loci identified using mrMLM overlapped part or most of the region of those obtained using MLM. Misra et al. (2017) determined the genetic basis of cooked grain length and width in rice using four GWAS methods—EMMAX, mrMLM, FASTmrEMMA, and ISIS EM-BLASSO. Thus, employing integrated single-locus and multi-locus GWAS models led to the verification of the significance of the underlying target regions, GWi7.1 and GWi7.2, and simultaneously identified the novel candidate genes. In this study, using three single-locus and three multi-locus models, 342 significant QTNs were identified. More loci were identified using multi-locus models than using single-locus models, and 15 loci were simultaneously identified in both single-locus and multi-locus models (**Supplementary Table S3**). These findings demonstrated the reliability of association analysis consequences and the practicality of combining single-locus and multi-locus GWASs to improve the power and robustness of association analyses.

Stable QTNs for Fiber Quality Traits Detected in Our GWAS

The marker loci/QTLs that are detected across multiple populations, environments and/or mapping methods, are highly stable and can enhance the efficiency and accuracy of the MAS (Su et al., 2010; Li et al., 2013). In cotton, using linkage mapping,

TABLE 4 | 12 QTNs controlling fiber quality traits identified in both this and previous studies.

Trait[a]	GWAS in this study			Previous studies			
	Maker associated	Chr.	Position (bp)	Marker linkaged/associated[b]	Chr.	Position (bp)	References
FL	TM58426	D5	52167190	BNL4047 (AM)	D5	51715146∼51715301	Sethi et al., 2017
	TM72875	D9	47994726	DPL0395 (LM), MGHES-55 (AM)	D9	48340706∼48340931, 48074891∼48075112	Sun et al., 2012; Iqbal and Rahman, 2017
FS	TM5639	A2	80304252	HAU880 (LM)	A2	80045222∼80045391	Wang et al., 2013
	TM21292	A7	72067994	i18340Gh, i44206Gh, i39753Gh, i02033Gh, i02034Gh, i02035Gh, i02037Gh, i49171Gh, i37604Gh (AM)	A7	71993462∼72249786	Sun et al., 2017
	TM43422	A13	5198708	i30934Gh (AM)	A13	5168143	Sun et al., 2017
	TM63860	D7	14495698	BNL3854 (LM)	D7	14236226∼14236344	An et al., 2010
	TM74995	D10	57945654	TM74991 (LM)	D10	57899125	Tan et al., 2018
FM	TM52959	D2	60834004	NAU2353 (LM)	D2	60579477∼60579638	Sun et al., 2012
FU	TM72633	D9	44334923	MGHES-6 (AM)	D9	44634167∼44634349	Iqbal and Rahman, 2017
	TM74995	D10	57945654	TM74991 (LM)	D10	57899125	Tan et al., 2018
FE	TM3939	A2	3531460	BNL1434 (AM)	A2	3419328∼3419575	Kantartzi and Stewart, 2008; Sethi et al., 2017
	TM56516	D4	47872954	i12839Gh (AM)	D4	47872770	Sun et al., 2017
	TM72628	D9	44115527	BNL1030 (AM)	D9	43992085∼43992321	Kantartzi and Stewart, 2008
	TM74999	D10	57965498	TM74991 (LM)	D10	57899125	Tan et al., 2018
	TM80198	D13	3477308	NAU2730 (LM)	D13	3582661∼3582860	Sun et al., 2012

[a] *FL, fiber length; FS, fiber strength; FM, fiber micronaire; FU, fiber uniformity; FE, fiber elongation.*

[b] *AM and LM mean association mapping and linkage mapping, respectively.*

Jia et al. (2011) located five QTLs for boll weight and lint percentage that were stably expressed in several environments by two mapping methods. Li et al. (2012) identified two QTLs for the node of the first fruiting branch and its height by two mapping methods. Sun et al. (2012) identified two QTLs for FS, which were simultaneously detected in four environments. Cai et al. (2014) performed association mapping of fiber quality traits and identified 70 significantly associated marker loci, of which 36 and four coincided with previously reported QTLs identified using linkage and association mapping populations, respectively. Here, 342 QTNs significantly associated with the fiber quality traits were detected using the values of individual environments (including BLUPs) and the six models. However, to obtain reliable results, only the QTNs simultaneously detected in at least two environments or by at least two models were displayed, and thus, 84 QTNs controlling the fiber quality traits were obtained. Of them, 29 were for FL, 22 were for FS, 11 were for FM, 12 were for FU, and 10 were for FE. These QTNs are highly stable and can potentially be used in the MAS of target traits. Additionally, nine QTNs, TM80185 (D13) for FL, TM1386 (A1) and TM14462 (A6) for FS, TM18616 (A7), TM54735 (D3), and TM79518 (D12) for FM, TM77489 (D12) and TM81448 (D13) for FU, and TM47772 (D1) for FE, were simultaneously detected in at least two environments, and by both single-locus and multi-locus GWASs. These nine QTNs also exhibited high phenotypic contributions of more than 10% in either a single-locus or multi-locus GWAS. Therefore, they could be given priority for MAS in future breeding programs.

Comparison of Our GWAS With the Results in Previous Studies

Presently, several QTLs/markers related to cotton fiber qualities have been identified using linkage mapping and association mapping in previous studies (Shen et al., 2005; Abdurakhmonov et al., 2008, 2009; Kantartzi and Stewart, 2008; An et al., 2010; Sun et al., 2012, 2017; Wang et al., 2013; Zhang et al., 2013; Cai et al., 2014; Qin et al., 2015; Islam et al., 2016; Li C. et al., 2016; Nie et al., 2016; Su et al., 2016b; Gapare et al., 2017; Huang et al., 2017; Iqbal and Rahman, 2017; Ma et al., 2017; Sethi et al., 2017; Tan et al., 2018). We compared the 342 QTNs detected in our GWAS (**Supplementary Table S3**) with SNPs and SSRs linked to/associated with QTLs for the same traits identified in previous studies by electronic PCR (e-PCR) based on their physical locations on the genome sequence (Zhang T. Z. et al., 2015). The markers linked to/associated with QTLs for the same traits that were located within the same region of ∼400 kb, were regarded as the same loci. Thus, 12 QTNs detected in our GWAS corresponded to previously reported SNPs and SSRs detected based on linkage and/or association mapping (**Table 4**). Specifically, two QTNs for FL, TM58426 (D5) and TM72875 (D9), corresponded to BNL4047 (Sethi et al., 2017) and DPL0395 (Sun et al., 2012)/MGHES-55 (Iqbal and Rahman, 2017), respectively; five QTNs for FS, TM5639 (A2), TM21292 (A7), TM43422 (A13), TM63860 (D7), and TM74995 (D10), corresponded to HAU880 (Wang et al., 2013), i18340Gh/i44206Gh/i39753Gh/i02033Gh/i02034Gh/i02035Gh/i02037Gh/i49171Gh/i37604Gh (Sun et al., 2017), i30934Gh (Sun et al., 2017), BNL3854 (An et al., 2010), and TM74991

(Tan et al., 2018), respectively; one QTN for FM, TM52959 (D2), corresponded to NAU2353 (Sun et al., 2012); two QTNs for FU, TM72633 (D9) and TM74995 (D10), corresponded to MGHES-6 (Iqbal and Rahman, 2017) and TM74991 (Tan et al., 2018), respectively; five QTNs for FE, TM3939 (A2), TM56516 (D4), TM72628 (D9), TM74999 (D10), and TM80198 (D13), corresponded to BNL1434 (Kantartzi and Stewart, 2008; Sethi et al., 2017), i12839Gh (Sun et al., 2017), BNL1030 (Kantartzi and Stewart, 2008), TM74991 (Tan et al., 2018), and NAU2730 (Sun et al., 2012), respectively. The 15 QTNs controlling the fiber quality, which were simultaneously detected in different populations with different genetic backgrounds, can potentially be used in the MAS of target traits.

Candidate Genes for Fiber Quality Traits

The identification of stable marker loci/QTLs could provide useful information for MAS. Candidate gene analyses are necessary for further gene cloning and functional verifications. Some candidate genes related to cotton fiber quality have already been identified using the GWAS approach. Islam et al. (2016) identified candidate genes related to fiber quality by gene expression and amino acid substitution analysis and suggested that the *Gh_A07G2049* (*GhRBB1_A07*) gene is a candidate for superior fiber quality in Upland cotton. Sun et al. (2017) identified 19 promising candidate genes related to FL and FS, of which, *Gh_A07G1758* could play a key role in the formation of cotton fiber, while *Gh_D03G0294* and *Gh_D05G1451* could play different roles during fiber development. In the study of Su et al. (2016b), three potential candidate genes, *CotAD_22823*, *CotAD_22824,* and *CotAD_22825*, for FL were identified, and the two peak SNPs (rsDt7:25931998 and rsDt7:25932026) associated with FL were positioned within one of the introns of *CotAD_22823*. In this study, 455 candidate genes surrounding the nine QTNs, which were simultaneously detected in at least two environments, were identified by both single-locus and multi-locus GWASs. Of the 455 candidate genes, 113 were highly expressed in ovules and/or fiber during their development, while being less expressed in root, stem, and leaf, suggesting that these genes might potentially affect the formation and development of cotton fiber, and thus contribute to fiber quality. These genes were categorized based on their functional characteristics from several databases. We cannot accurately determine which genes are directly related to fiber quality based on the data of this study. However, the results will provide useful information for future works. Cotton fiber development shares many similarities with the trichomes of Arabidopsis leaves in cellular and genetic features (Serna and Martin, 2006). Further, bioinformatics analyses indicated that the four genes, *Gh_D13G1461*, *Gh_D12G0232*, *Gh_D01G0052,* and *Gh_D12G0240*, may be promising candidate genes for improving the fiber quality. However, the formation of cotton fiber is a complicated physiological and biochemical process that might involve a large number of structural, regulatory, and biochemical pathway-related genes. Therefore, the functions of many genes in cotton remain to be elucidated.

CONCLUSION

This research reported the GWAS of fiber quality traits in Upland cotton based on a recently developed CottonSNP80K array. A total of 342 QTNs controlling the fiber quality traits were detected via three single-locus and three multi-locus models. Of these QTNs, 84 were simultaneously detected in at least two environments or by at least two models. Further, nine QTNs were simultaneously detected in at least two environments, and by both single- and multi-locus models. 12 QTNs corresponded to previously reported SNPs and SSRs. In total, 455 candidate genes were identified within 400-kb upstream and downstream of the above nine QTNs based on the genome sequence of Upland cotton. Among these genes, 113 might potentially affect the formation and development of cotton fiber and four might be promising candidate genes for improving fiber quality.

AUTHOR CONTRIBUTIONS

CL designed the experiment and wrote the manuscript. QW provided the experimental materials. YF, RS, and YW performed the experiments. All authors commented on the manuscript.

FUNDING

This research was supported by the National Natural Science Foundation of China (31671743), the Innovative Talent Support Program of Science and Technology of Henan Institute of Higher Learning (16HASTIT014), and the Technology Demonstration and Industrialization of Seed Industry Facing the Five Central Asian countries (161100510100).

ACKNOWLEDGMENTS

The authors would like to thank Prof. Jim M. Dunwell of the School of Agriculture, Policy and Development at the University of Reading, United Kingdom, for helping with English language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01083/full#supplementary-material

Supplementary Table S1 | Names and ecological sources of the 169 Upland cotton accessions.

Supplementary Table S2 | Phenotypic values for the fiber quality traits of the 169 accessions in four environments.

Supplementary Table S3 | All 342 QTNs for fiber quality traits detected using the values of individual environments (including BLUP) and the six models.

Supplementary Table S4 | Normalized fragments per kilobase of transcript per million fragments mapped values of the 455 candidate genes.

Supplementary Table S5 | A systematic summary of the 113 candidate genes.

Supplementary Figure S1 | Single nucleotide polymorphism distributions on the 26 chromosomes of Upland cotton.

Supplementary Figure S2 | QTN, TM80185 (D13), associated with FL, was simultaneously detected in at least two environments, by both single-locus and multi-locus GWASs.

Supplementary Figure S3 | QTNs, TM1386 (A1), and TM14462 (A6), associated with FS, were simultaneously detected in at least two environments, by both single-locus and multi-locus GWASs.

Supplementary Figure S4 | QTNs, TM18616 (A7), TM54735 (D3), and TM79518 (D12), associated with FM, were simultaneously detected in at least two environments, by both single-locus and multi-locus GWASs.

Supplementary Figure S5 | QTNs, TM77489 (D12), and TM81448 (D13), associated with FU, were simultaneously detected in at least two environments, by both single-locus and multi-locus GWASs.

Supplementary Figure S6 | QTN, TM47772 (D1), associated with FE, was simultaneously detected in at least two environments, by both single-locus and multi-locus GWASs.

REFERENCES

Abdurakhmonov, I. Y., Kohel, R. J., Yu, J. Z., Pepper, A. E., Abdullaev, A. A., Kushanov, F. N., et al. (2008). Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. *Genomics* 92, 478–487. doi: 10.1016/j.ygeno.2008.07.013

Abdurakhmonov, I. Y., Saha, S., Jenkins, J. N., Buriev, Z. T., Shermatov, S. E., Scheffler, B. E., et al. (2009). Linkage disequilibrium-based association mapping of fiber quality traits in *G. hirsutum* L. variety germplasm. *Genetics* 136, 401–417. doi: 10.1007/s10709-008-9337-8

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

An, C., Jenkins, J. N., Wu, J., Guo, Y., and McCarty, J. C. (2010). Use of fiber and fuzz mutants to detect QTL for yield components, seed, and fiber traits of Upland cotton. *Euphytica* 172, 21–34. doi: 10.1007/s10681-009-0009-2

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bates, D., Maechler, M., and Bolker, B. (2011). *Lme4: Linear Mixed Effects Models Using S4 Classes.* Available online at: http://cran.r-project.,org/web/packages/lme4/index.html (Accessed 1 September 2011).

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. doi: 10.1093/bioinformatics/btm308

Cai, C., Zhu, G., Zhang, T., and Guo, W. (2017). High-density 80K SNP array is a powerful tool for genotyping *G. hirsutum* accessions and genome analysis. *BMC Genomics* 18:654. doi: 10.1186/s12864-017-4062-2

Cai, C. P., Ye, W. X., Zhang, T. Z., and Guo, W. Z. (2014). Association analysis of fiber quality traits and exploration of elite alleles in Upland cotton cultivars/accessions (*Gossypium hirsutum* L.). *J. Integr. Plant Biol.* 56, 51–62. doi: 10.1111/jipb.12124

Chen, Z. J., Scheffler, B. E., Dennis, E., Triplett, B. A., Zhang, T. Z., Guo, W. Z., et al. (2007). Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* 145, 1303–1310. doi: 10.1104/pp.107.107672

Du, L., Cai, C., Wu, S., Zhang, F., Hou, S., and Guo, W. (2016). Evaluation and exploration of favorable QTL alleles for salt stress related traits in cotton cultivars (*G. hirsutum L.). PLoS ONE* 11:e0151076. doi: 10.1371/journal.pone.0151076

Fang, L., Gong, H., Hu, Y., Liu, C., Zhou, B., Huang, T., et al. (2017). Genomic insights into divergence and dual domestication of cultivated allotetraploid cottons. *Genome Biol.* 18:33. doi: 10.1186/s13059-017-1167-5

Feng, J. Y., Wen, Y. J., Zhang, J., and Zhang, Y. M. (2016). Advances on methodologies for genome-wide association studies in plants. *Acta Agron. Sin.* 42, 945–956. doi: 10.3724/SP.J.1006.2016.00945

Gapare, W., Conaty, W., Zhu, Q. H., Liu, S., Stiller, W., Llewellyn, D., et al. (2017). Genome-wide association study of yield components and fiber quality traits in a cotton germplasm diversity panel. *Euphytica* 213:66. doi: 10.1007/s10681-017-1855-y

Hoggart, C. J., Whittaker, J. C., Iorio, M. D., and Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* 4:e1000130. doi: 10.1371/journal.pgen.1000130

Huang, C., Nie, X. H., Shen, C., You, C. Y., Li, W., Zhao, W. X., et al. (2017). Population structure and genetic basis of the agronomic traits of Upland cotton in China revealed by a genome-wide association study using high-density SNPs. *Plant Biotechnol. J.* 15:1374. doi: 10.1111/pbi.12722

Huang, X. H., and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Ann. Rev. Plant Biol.* 65, 531–551. doi: 10.1146/annurev-arplant-050213-035715

Hulse-Kemp, A. M., Lemm, J., Plieske, J., Ashrafi, H., Buyyarapu, R., Fang, D. D., et al. (2015). Development of a 63k SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium* spp. *G3-Genes Genom. Genet.* 5, 1187–1209. doi: 10.1534/g3.115.018416

Iqbal, M. A., and Rahman, M. (2017). Identification of marker-trait associations for lint traits in cotton. *Front. Plant Sci.* 8:86. doi: 10.3389/fpls.2017.00086

Islam, M. S., Thyssen, G. N., Jenkins, J. N., Zeng, L. H., Delhom, C. D., McCarty, J. C., et al. (2016). A MAGIC population-based genome-wide association study reveals functional association of *GhRBB1_A07* gene with superior fiber quality in cotton. *BMC Genomics* 17:903. doi: 10.1186/s12864-016-3249-2

Jia, F., Sun, F. D., Li, J. W., Liu, A. Y., Shi, Y. Z., Gong, J. W., et al. (2011). Identification of QTL for boll weight and lint percentage of Upland cotton (*Gossypium hirsutum* L.) RIL population in multiple environments. *Mol. Plant Breed.* 9, 318–326.

Jia, Y. H., Sun, X. W., Sun, J. L., Pan, Z. E., Wang, X. W., He, S. P., et al. (2014). Association mapping for epistasis and environmental interaction of yield traits in 323 cotton cultivars under 9 different environments. *PLoS ONE* 9:e95882. doi: 10.1371/journal.pone.0095882

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, 277–280. doi: 10.1093/nar/gkh063

Kantartzi, S. K., and Stewart, J. M. (2008). Association analysis of fibre traits in *Gossypium arboreum* accessions. *Plant Breed.* 127, 173–179. doi: 10.1111/j.1439-0523.2008.01490.x

Kasili, R., Huang, C. C., Walker, J. D., Simmons, L. A., Zhou, J., Faulk, C., et al. (2015). *BRANCHLESS TRICHOMES* links cell shape and cell cycle control in *Arabidopsis* trichomes. *Development* 138, 2379–2388. doi: 10.1242/dev.058982

Lee, M. (1995). DNA marker and plant breeding programs. *Adv. Agron.* 55, 265–344. doi: 10.1016/S0065-2113(08)60542-8

Li, C., Dong, Y., Zhao, T., Li, L., Li, C., Yu, E., et al. (2016). Genome-wide SNP linkage mapping and QTL analysis for fiber quality and yield traits in the Upland cotton recombinant inbred lines population. *Front. Plant Sci.* 7:1356. doi: 10.3389/fpls.2016.01356

Li, C. Q., Ai, N. J., Zhu, Y. J., Wang, Y. Q., Chen, X. D., Li, F., et al. (2016b). Association mapping and favourable allele exploration for plant architecture traits in Upland cotton (*Gossypium hirsutum* L.) accessions. *J. Agr. Sci.-Cambridge* 154, 567–583. doi: 10.1017/S0021859615000428

Li, C. Q., Wang, C. B., Dong, N., Wang, X. Y., Zhao, H. H., Richard, C., et al. (2012). QTL detection for node of first fruiting branch and its height in Upland cotton (*Gossypium hirsutum* L.). *Euphytica* 188, 441–451. doi: 10.1007/s10681-012-0720-2

Li, C. Q., Wang, X. Y., Dong, N., Zhao, H. H., Xia, Z., Wang, R., et al. (2013). QTL analysis for early-maturing traits in cotton using two Upland cotton (*Gossypium hirsutum* L.) crosses. *Breed. Sci.* 63, 154–163. doi: 10.1270/jsbbs.63.154

Li, C. Q., Xu, X. J., Dong, N., Ai, N. J., and Wang, Q. L. (2016a). Association mapping identifies markers related to major early-maturating traits in upland cotton (*Gossypium hirsutum* L.). *Plant Breed.* 135, 483–491. doi: 10.1111/pbr.12380

Li, H. G., Zhang, L. P., Hu, J. H., Zhang, F. G., Chen, B. Y., Xu, K., et al. (2017). Genome-wide association mapping reveals the genetic control underlying branch angle in rapeseed (*Brassica napus* L.) *Front. Plant Sci.* 8:1054. doi: 10.3389/fpls.2017.01054

Li, T., Ma, X., Li, N., Zhou, L., Liu, Z., Han, H., et al. (2017). Genome-wide association study discovered candidate genes of *Verticillium* wilt resistance in Upland cotton (*Gossypium hirsutum* L.). *Plant Biotechnol. J.* 15, 1520–1532. doi: 10.1111/pbi.12734

Liang, Z. L. (1999). *The Genetics and Breeding of Interspecific Hybridization in Cotton*. Beijing: Science Press.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399. doi: 10.1093/bioinformatics/bts444

Liu, G. Z., Mei, H. X., Wang, S., Li, X. H., Zhu, X. F., and Zhang, T. Z. (2015). Association mapping of seed oil and protein contents in Upland cotton. *Euphytica* 205, 637–645. doi: 10.1007/s10681-015-1450-z

Ma, L., Zhao, Y., Wang, Y., Shang, L., and Hua, J. (2017). QTLs analysis and validation for fiber quality traits using maternal backcross population in Upland cotton. *Front. Plant Sci.* 8:2168. doi: 10.3389/fpls.2017.02168

Machado, A., Wu, Y., Yang, Y., Llewellyn, D. J., and Dennis, E. S. (2009). The *MYB* transcription factor *GhMYB25* regulates early fiber and trichome development. *Plant J.* 59, 52–62. doi: 10.1111/j.1365-313X.2009.03847.x

Marks, M. D., Wenger, J. P., Gilding, E., Jilk, R., and Dixon, R. A. (2009). Transcriptome analysis of *Arabidopsis* wild-type and *gl3-sst sim* trichomes identifies four additional genes required for trichome development. *Mol. Plant* 2, 803–822. doi: 10.1093/mp/ssp037

Mei, H. X., Ai, N. J., Zhang, X., Ning, Z. Y., and Zhang, T. Z. (2014). QTLs conferring FOV 7 resistance detected by linkage and association mapping in Upland cotton. *Euphytica* 197, 237–249. doi: 10.1007/s10681-014-1063-y

Mei, H. X., Zhu, X. F., and Zhang, T. Z. (2013). Favorable QTL alleles for yield and its components identified by association mapping in Chinese Upland cotton cultivars. *PLoS ONE* 8:e82193. doi: 10.1371/journal.pone.0082193

Miller, P. A., and Rawlings, J. O. (1967). Selection for increased lint yield and correlated responses in Upland cotton *Gossypium hirsutum* L. *Crop Sci.* 7, 637–640. doi: 10.2135/cropsci1967.0011183X000700060024x

Misra, G., Badoni, S., Anacleto, R., Graner, A., Alexandrov, N., and Sreenivasulu, N. (2017). Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. *Sci. Rep.* 7:12478. doi: 10.1038/s41598-017-12778-6

Mohan, M.,Nair, S., Bhagwat, A., Krishna, T. G., Yano, M., Bhatia, C. R., et al. (1997). Genome mapping, molecular markers and marker assisted selection in crop plants. *Mol. Breed.* 3, 87–103. doi: 10.1023/A:1009651919792

Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., et al. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U.S.A.* 110, 453–458. doi: 10.1073/pnas.1215985110

Newell, M. A., Cook, D., Tinker, N. A., and Jannink, J. L. (2011). Population structure and linkage disequilibrium in oat (*Avena sativa* L.): implications for genome-wide association studies. *Theor. Appl. Genet.* 122, 623–632. doi: 10.1007/s00122-010-1474-7

Nie, X., Huang, C., You, C., Li, W., Zhao, W., Shen, C., et al. (2016). Genome-wide SSR-based association mapping for fiber quality in nation-wide Upland cotton inbreed cultivars in China. *BMC Genomics* 17:352. doi: 10.1186/s12864-016-2662-x

Qin, H. D., Chen, M., Yi, X. D., Bie, S., Zhang, C., Zhang, Y. C., et al. (2015). Identification of associated SSR markers for yield component and fiber quality traits based on frame map and Upland cotton collections. *PLoS ONE* 10:e0118073. doi: 10.1371/journal.pone.0118073

Roudier, F., Fernandez, A. G., Fujita, M., Himmelspach, R., Borner, G. H. H., Schindelman, G., et al. (2005). COBRA, an Arabidopsis extracellular glycosyl-phosphatidyl inositol-anchored protein, specifically controls highly anisotropic expansion through its involvement in cellulose microfibril orientation. *Plant Cell* 17:1749. doi: 10.1105/tpc.105.031732

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., et al. (2003). Tm4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34:374.

Saeed, M., Guo, W. Z., and Zhang, T. Z. (2014). Association mapping for salinity tolerance in cotton (*Gossypium hirsutum* L.) germplasm from US and diverse regions of China. *Aust. J. Crop Sci.* 8, 338–346.

Serna, L., and Martin, C. (2006). Trichomes: different regulatory networks lead to convergent structures. *Trend Plant Sci.* 11, 274–280. doi: 10.1016/j.tplants.2006.04.008

Sethi, K., Siwach, P., and Verma, S. K. (2017). Linkage disequilibrium and association mapping of fibre quality traits in elite Asiatic cotton (*Gossypium arboreum*) germplasm populations. *Czech. J. Genet. Plant Breed.* 53, 159–167. doi: 10.17221/142/2016-CJGPB

Shen, X., Guo, W., Zhu, Xi., Yuan, Y., Yu, J.Z., Kohel, R.J., et al. (2005). Molecular mapping of QTLs for fiber qualities in three diverse lines in Upland cotton using SSR markers. *Mol. Breed.* 15, 169–181. doi: 10.1007/s11032-004-4731-0

Smith, C. W., and Coyle, G. G. (1997). Association of fiber quality parameters and within-boll yield components in Upland cotton. *Crop Sci.* 37, 1775–1779. doi: 10.2135/cropsci1997.0011183X003700060019x

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J. L., et al. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116, 395–408. doi: 10.1038/hdy.2015.113

Su, C. F., Lu, W. G., Zhao, T. J., and Gai, J. Y. (2010). Verification and fine-mapping of QTL conferring days to flowering in soybean using residual heterozygous lines. *Chin. Sci. Bull.* 55, 499–508. doi: 10.1007/s11434-010-0032-7

Su, J., Fan, S., Li, L., Wei, H., Wang, C., Wang, H., et al. (2016a). Detection of favorable QTL alleles and candidate genes for lint percentage by GWAS in Chinese Upland cotton. *Front. Plant Sci.* 7:1576. doi: 10.3389/fpls.2016.01576

Su, J., Li, L., Pang, C., Wei, H., Wang, C., Song, M., et al. (2016b). Two genomic regions associated with fiber quality traits in Chinese Upland cotton under apparent breeding selection. *Sci. Rep.* 6:38496. doi: 10.1038/srep38496

Su, J., Pang, C., Wei, H., Li, L., Liang, B., Wang, C., et al. (2016c). Identification of favorable SNP alleles and candidate genes for traits related to early maturity via GWAS in Upland cotton. *BMC Genomics* 17:687. doi: 10.1186/s12864-016-2875-z

Sun, F. D., Zhang, J. H., Wang, S. F., Gong, W. K., Shi, Y. Z., Liu, A. Y., et al. (2012). QTL mapping for fiber quality traits across multiple generations and environments in Upland cotton. *Mol. Breed.* 30, 569–582. doi: 10.1007/s11032-011-9645-z

Sun, Z., Wang, X., Liu, Z., Gu, Q., Zhang, Y., Li, Z., et al. (2017). Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in *Gossypium hirsutum* L. *Plant Biotechnol. J.* 15, 982–996. doi: 10.1111/pbi.12693

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tan, Z., Zhang, Z., Sun, X., Li, Q., Sun, Y., Yang, P., et al. (2018). Genetic map construction and fiber quality QTL mapping using the CottonSNP80K array in Upland cotton. *Front. Plant Sci.* 9:225. doi: 10.3389/fpls.2018.00225

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 133–136 doi: 10.1093/nar/28.1.33

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Biorxiv*. [preprint] doi: 10.1101/005165

Visioni, A., Tondelli, A., Francia, E., Pswarayi, A., Malosetti, M., Russell, J., et al. (2013). Genome-wide association mapping of frost tolerance in barley (*Hordeum vulgare* L.). *BMC Genomics* 14:424. doi: 10.1186/1471-2164-14-424

Wang, D., Eskridge, K. M., and Crossa, J. (2010). Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. *J. Agric. Biol. Environ. Stat.* 16, 170–184. doi: 10.1007/s13253-010-0046-2

Wang, F. R., Xu, Z. Z., Sun, R., Gong, Y. C., Liu, G. D., Zhang, J. X., et al. (2013). Genetic dissection of the introgressive genomic components from *Gossypium barbadense* L. that contribute to improved fiber quality in *Gossypium hirsutum* L. *Mol. Breed.* 32, 547–562. doi: 10.1007/s11032-013-9888-y

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wang, Y. Q., Yang, W. H., Xu, H. X., Zhou, D. Y., Feng, X. A., Kuang, M., et al. (2009). The main problems and recommendations in Chinese cotton production. *Chin. Agr. Sci. Bull.* 25, 86–90.

Wen, J., Zhao, X., Wu, G., Xiang, D., Liu, Q., Bu, S. H., et al. (2015). Genetic dissection of heterosis using epistatic association mapping in a partial ncii mating design. *Sci. Rep.* 5:18376. doi: 10.1038/srep18376

Wen, Y. J., Zhang, H. W., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief Bioinform.* 18:906. doi: 10.1093/bib/bbx028

Wendel, J. F., and Cronn, R. C. (2003). Polyploidy and the evolutionary history of cotton. *Adv. Agron.* 78, 139–186. doi: 10.1016/S0065-2113(02)78004-8

Xu, Y., Xu, C., and Xu, S. (2017). Prediction and association mapping of agronomic traits in maize using multiple omic data. *Heredity* 119, 174–184. doi: 10.1038/hdy.2017.27

Yi, N., and Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* 179, 1045–1055. doi: 10.1534/genetics.107.085589

Zegeye, H., Rasheed, A., Makdis, F., Badebo, A., and Ogbonnaya, F. C. (2014). Genome-wide association mapping for seedling and adult plant resistance to stripe rust in synthetic hexaploid wheat. *PLoS ONE* 9:e105593. doi: 10.1371/journal.pone.0105593

Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., et al. (2015). Genome wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics* 16:217. doi: 10.1186/s12864-015-1441-4

Zhang, J. F., Shi, Y. Z., Liang, Y., Jia, Y. J., Zhang, B. C., Li, J. W., et al. (2012). Evaluation of yield and fiber quality traits of chromosome segment substitution lines population (BC_5F_3 and $BC_5F_{3:4}$) in cotton. *J. Plant Resour. Environ.* 13, 773–781.

Zhang, T. Z., Hu, Y., Jiang, W. K., Fang, L., Guan, X. Y., Chen, J. D., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotech.* 33, 531–537. doi: 10.1038/nbt.3207

Zhang, T. Z., Qian, N., Zhu, X. F., Chen, H., Wang, S., Mei, H. X., et al. (2013). Variations and transmission of QTL alleles for yield and fiber qualities in Upland cotton cultivars developed in China. *PLoS ONE* 8:e57220. doi: 10.1371/journal.pone.0057220

Zhang, X., Zhang, J., He, X., Wang, Y., Ma, X., and Yin, D., et al (2017). Genome-wide association study of major agronomic traits related to domestication in peanut. *Front. Plant Sci.* 8:1611. doi: 10.3389/fpls.2017.01611

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhao, Y., Wang, H., Chen, W., and Li, Y. H. (2014). Genetic structure, linkage disequilibrium and association mapping of verticillium wilt resistance in elite cotton (*Gossypium hirsutum* L.) germplasm population. *PLoS ONE* 9:e86308. doi: 10.1371/journal.pone.0086308

Zhou, Q., Zhou, C., Zheng, W., Mason, A., Fan, S., Wu, C., et al. (2017). Genome-wide SNP markers based on SLAF-seq uncover breeding traces in rapeseed (*Brassica napus* L.). *Front. Plant Sci.* 8:648. doi: 10.3389/fpls.2017.00648

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 20 August 2018
doi: 10.3389/fpls.2018.01184

Genome-Wide Association Studies for Dynamic Plant Height and Number of Nodes on the Main Stem in Summer Sowing Soybeans

*Fangguo Chang, Chengyu Guo, Fengluan Sun, Jishun Zhang, Zili Wang, Jiejie Kong, Qingyuan He, Ripa A. Sharmin and Tuanjie Zhao**

National Center for Soybean Improvement, Key Laboratory of Biology and Genetics and Breeding for Soybean, Ministry of Agriculture, State Key Laboratory for Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China

Plant height (PH) and the number of nodes on the main stem (NN) serve as major plant architecture traits affecting soybean seed yield. Although many quantitative trait loci for the two traits have been reported, their genetic controls at different developmental stages in soybeans remain unclear. Here, 368 soybean breeding lines were genotyped using 62,423 single nucleotide polymorphism (SNP) markers and phenotyped for the two traits at three different developmental stages over two locations in order to identify their quantitative trait nucleotides (QTNs) using compressed mixed linear model (CMLM) and multi-locus random-SNP-effect mixed linear model (mrMLM) approaches. As a result, 11 and 13 QTNs were found by CMLM to be associated with PH and NN, respectively. Among these QTNs, 8, 3, and 4 for PH and 6, 6, and 8 for NN were found at the three stages, and 3 and 6 were repeatedly detected for PH and NN. In addition, 34 and 30 QTNs were found by mrMLM to be associated with PH and NN, respectively. Among these QTNs, 11, 13, and 16 for PH and 11, 15, and 8 for NN were found at the three stages. A majority of these QTNs overlapped with the previously reported loci. Moreover, one QTN within the known *E2* locus for flowering time was detected for the two traits at all three stages, and another that overlapped with the *Dt1* locus for stem growth habit was also identified for the two traits at the mature stage. This may explain the highly significant correlation between the two traits. Our findings provide evidence for mixed major plus polygenes inheritance for dynamic traits and an extended understanding of their genetic architecture for molecular dissection and breeding utilization in soybeans.

Keywords: soybean, genome-wide association study, quantitative trait nucleotide, plant height, number of nodes on the main stem, dynamic development

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
HaiYan Lü,
Henan Agricultural University, China
Jia Wen,
University of North Carolina
at Charlotte, United States

****Correspondence:***
Tuanjie Zhao
tjzhao@njau.edu.cn

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *15 May 2018*
Accepted: *24 July 2018*
Published: *20 August 2018*

Citation:
Chang F, Guo C, Sun F, Zhang J, Wang Z, Kong J, He Q, Sharmin RA and Zhao T (2018) Genome-Wide Association Studies for Dynamic Plant Height and Number of Nodes on the Main Stem in Summer Sowing Soybeans. Front. Plant Sci. 9:1184. doi: 10.3389/fpls.2018.01184

INTRODUCTION

Soybean (*Glycine max* L. Merr.) is an important source of plant protein and oil for human consumption. Improving seed yield is the major target for soybean breeders. Plant architecture can strongly affect the suitability and productivity of seed yield in agricultural crops (Li R. et al., 2014). Plant height (PH) and the number of nodes on the main stem (NN) as key plant type traits have obvious effects on the seed yield of soybean because they are related to some important characteristics such as lodging and adaptability (Chapman et al., 2003; Liu et al., 2011). PH and NN are highly correlated with some other soybean agronomic traits such as days to flowering (DF) and days to maturity (DM), which are thought to be mainly adaptive traits in response to the photoperiod, allowing each cultivar to adapt to limited geographic regions (Zhang et al., 2004, 2015; Panthee et al., 2007).

Plant height and NN are complex traits governed by many quantitative trait loci (QTL) in soybeans (Lee et al., 1996; Zhang et al., 2004; Liu et al., 2011, 2013; Yao et al., 2015; Cao et al., 2017). To date, more than 200 and 30 QTLs for PH and NN have been reported on SoyBase[1] via linkage mapping. Recently, a study showed that highly significant correlations were observed among yield-related traits such as PH and NN. In addition, 23 novel QTLs and 8 QTL hotspots were identified for yield and quality-related traits by QTL analysis in a soybean RILs population. In particular, most loci associated with these traits were co-located in the same genomic region on three chromosomes (Chr04, Chr06, and Chr19), which was consistent with the results of phenotypic correlation analysis (Liu et al., 2017). Fang et al. (2017) identified 245 loci for 84 agronomic traits via genome-wide association studies (GWAS) in 809 soybean accessions and further dissected the genetic networks underlying the phenotypic correlations of traits. Of these traits, PH and NN exhibited a significant positive correlation. Some major genes were also cloned to reveal the molecular mechanism of PH and NN. Two known loci, *Dt1* for stem growth habit and *E2* for DF, were involved in regulating PH and NN and other agronomic traits in soybeans (Kato et al., 2015; Zhang et al., 2015). *Dt1* plays a primary role in determinate stem varieties and has an epistasis effect on the *Dt2* locus, another stem growth habit locus involved in the development of PH in soybeans (Bernard, 1972; Liu et al., 2016). The *E*2 locus encodes a homolog of *GIGANTEA*, which regulates the expression of *CO* and *FT* in *Arabidopsis* and controls soybean flowering through regulating *GmFT2a* (Watanabe et al., 2011). On the other hand, a target trait such as PH or NN performs dynamically when plants grow gradually. However, the phenotypes of PH and NN were mostly investigated at the mature stage. Sun et al. (2006) reported that different QTL architectures have been found for PH at the different developmental stages through linkage mapping. Although several studies of the developmental behavior of quantitative traits have been reported in soybeans (Vodkin et al., 2004; Li W. et al., 2007; Xin et al., 2008; Teng et al., 2009), the genetic architecture of dynamic development behavior of complex traits remains to be further explored.

With the wide application of next-generation sequencing techniques, high-throughput single nucleotide polymorphism (SNPs) have been discovered and utilized to construct high-resolution genetic maps and to conduct GWAS (Hyten et al., 2008; Michael and VanBuren, 2015; Song et al., 2016). GWAS is a powerful approach because it takes full advantage of all recombination events that occur in the evolutionary process of a natural population. It has been successfully used to explore the genetic basis for a broad range of complex traits in many plant species such as *Arabidopsis* (Atwell et al., 2010; Horton et al., 2012), rice (Huang et al., 2010; Yang et al., 2014), maize (Kump et al., 2011; Li H. et al., 2013), and soybean (Hwang et al., 2014; Sonah et al., 2015; Zhou et al., 2015; Zhang et al., 2016; Chang and Hartman, 2017).

The mixed linear models (MLMs) have been widely used for GWAS. The compressed MLM (CMLM) was also utilized to reduce computing time and to improve statistical power for quantitative trait nucleotide (QTN) detection (Zhang et al., 2010). Nevertheless, the current GWAS methods such as MLM and CMLM are mainly based on the single-locus genome-wide scan, which often requires correction for multiple tests. The typical Bonferroni correction is so conservative that some small-effect loci may not reach the significance threshold. With the rapid development of statistical methods, several multi-locus GWAS approaches have been developed to improve the power of QTN detection (Cho et al., 2010; Segura et al., 2012; Moser et al., 2015). The obvious advantage of these methods is no Bonferroni correction due to the nature of multi-locus methods. Recently, Wang et al. (2016) proposed a new multi-locus random-SNP-effect mixed linear model (mrMLM) method to improve the power and accuracy of GWAS. Differing from the other multi-locus methods, the mrMLM is a two-stage method. At the first stage, the SNP effect is viewed as being random, and all the potentially associated markers are selected by a random-SNP-effect MLM with a modified Bonferroni correction for significance test. At the second stage, all the selected markers are placed into one model and all the non-zero effects are further detected by a likelihood ratio test for QTN identification.

Summer-planting soybean is a major soybean crop grown in the region between the Yangtze River and the Huai River in the southern region of middle China, an important soybean production area. Although the genetic architecture of some agronomic traits such as PH was reported in our previous study (Cao et al., 2017) in the summer planting soybean, the genetic bases of dynamic PH and NN for them remain largely unknown. The aim of this study was to dissect the genetic basis of PH and NN at three different developmental stages in 368 summer planting soybean genotypes using the GWAS strategy. Our findings will provide useful genetic information for soybean molecular breeding.

MATERIALS AND METHODS

Plant Materials, Field Trials and Phenotypic Evaluation

A soybean breeding line (SBL) population containing 368 accessions was established to service the local soybean breeding. All these pure lines were obtained from the National Center for soybean improvement, Nanjing Agricultural University, Nanjing, China. All experimental materials were planted at Jiangpu (JP) (32°12′N and 118°37′E) and Fengyang (FY) (32°47′N and 117°19′E) Station in Jiangsu and Anhui province, respectively, on June 20, 2011. At each location, the experimental design was a randomized complete block with 50 cm × 50 cm hill plots and three replications. The phenotypes for PH and NN were

[1] www.soybase.org

measured at the three different developmental stages over two locations: Stage 1 (35 days after the emergence of seedlings), Stage 2 (50 days after the emergence of seedlings) and Stage 3 (90 days after the emergence of seedlings). All the phenotypes were named PH1, PH2 and PH3 for PH, and NN1, NN2 and NN3 for NN. PH and NN were the averages of three measurements per plot.

Statistical Analysis

The analysis of variance (ANOVA) was performed for all traits using the PROC GLM procedure of SAS version 9.3 (SAS Institute, Inc., Cary, NC, United States). The model for the phenotype of a trait was $y_{\mathrm{ijk}} = \mu + G_{\mathrm{i}} + E_{\mathrm{j}} + GE_{\mathrm{ij}} + R_{\mathrm{k(j)}} + e_{\mathrm{ijk}}$ where μ is the total mean, G_i is the effect of the i^{th} genotype, E_j is the effect of the j^{th} environment, GE_{ij} is the interaction effect between the i^{th} genotype and the j^{th} environment, $R_{\mathrm{k(j)}}$ is the effect of the k^{th} block within the j^{th} environment, and e_{ijk} is a random error following $N(0, \sigma_{\mathrm{e}}^2)$. Descriptions of all traits were determined by the mean of each trait over two locations. The broad-sense heritability (h^2) was calculated as: $h^2 = \sigma_{\mathrm{g}}^2 / \left(\sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{ge}}^2/n + \sigma_{\mathrm{e}}^2/nr\right)$ for combined environments and $h^2 = \sigma_{\mathrm{g}}^2 / \left(\sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{e}}^2\right)$ for an individual environment, where σ_{g}^2 is the genotypic variance, σ_{ge}^2 is the genotype by environment interaction variance, σ_{e}^2 is the error variance, n is the number of environments, and r is the number of replications. Variance components and correlation coefficients were estimated by the PROC VARCOMP and CORR procedure of SAS, respectively. To minimize the effects of environmental variation, the best linear unbiased predictors (BLUPs) of individual lines for each trait were calculated using the R package lme4 (Bates et al., 2015).

Genotyping, SNPs Polymorphism and Haplotype Block Estimation

High-throughput SNPs were generated by RAD-seq. The quality control of sequencing data and methods of calling variations are described in our previous study (Li et al., 2016). A total of 62423 SNPs with a minor allele frequency (MAF) $\geq$ 5% were used for further analysis in the present study.

The MAF of the SNPs was calculated using VCFtools software (Danecek et al., 2011). Haplotype blocks were estimated using pLINK V1.90 software (Purcell et al., 2007) with the command option –blocks, following the default algorithm as described by Gabriel et al. (2002). The visualization of haplotype blocks was carried out with the R package LDheatmap (Shin et al., 2006). The estimated parameters for SNPs polymorphism were displayed using circos (Krzywinski et al., 2009).

Linkage Disequilibrium Estimation

Linkage disequilibrium (LD) between pairwise SNPs was calculated as the squared correlation coefficient (r^2) of alleles using the linkage disequilibrium tools option of RTM-GWAS V1.1 software (He et al., 2017). The r^2 value was calculated for all pairwise SNPs with a 100 kb summary bin setting within the 5 Mb distance and then averaged across the whole genome. Because of the substantial difference in recombination rate between euchromatic and heterochromatic regions, the r^2 value was calculated separately for the two chromosomal regions. The physical length of the euchromatic and heterochromatic regions for each chromosome was defined as in the *G. max* 1.0 reference genome. The LD decay rate was measured as the chromosomal distance at which the average pairwise r^2 dropped to half its maximum value (Huang et al., 2010). Only r^2 for SNPs with pairwise distances less than 5 Mb in either the euchromatic or heterochromatic region was used to draw the average LD decay figure by R script.

Population Structure and Principal Component Analysis

Filtering SNPs used the –indep-pairwise command option of pLINK. The pruned data were then used to estimate population structure using ADMIXTURE V1.3.0 software (Alexander et al., 2009). In the ADMIXTURE setting, the number of clusters (K) was set from 1 to 10 initially; then, each Q and the relevant P-value was estimated. The most likely number of subpopulations was determined using the method described in Evanno et al. (2005). A principal component analysis (PCA) of whole-genome SNPs was performed using EIGENSOFT V5.0.2 software (Price et al., 2006) smartpca program, and the first two eigenvectors were plotted in two dimensions. The neighbor-joining tree was constructed using TASSEL V5.2 software (Bradbury et al., 2007).

Genome-Wide Association Studies

After excluding SNPs with an MAF $<$ 5%, 62423 SNPs were retained for 368 soybean accessions. To minimize false positives and increase statistical power, the population structure (Q) and kinship (K) matrix were estimated for the population. For the MLM, both the regular MLM and compressed CMLM involve the Q and K matrices as a fixed effect and random effect, where the Q matrix was replaced by the principal components (PCs) in CMLM. CMLM was implemented by the R package GAPIT (Genome Association Prediction Tool) V2 (Tang et al., 2016). Another R package, mrMLM V2.1, representing the mrMLM method was adopted (Wang et al., 2016). Thus, GWAS was conducted by combining the CMLM and mrMLM methods in this study. The critical threshold of significance for SNP-trait association was set at a P-value = 1.0×10^{-4} in CMLM according to the empirical value and at a LOD value of 3 in mrMLM. A QTN was defined as a haplotype block possessing SNPs identified as significantly associated with a trait (Schneider et al., 2016). The QTNs were named following the nomenclature described by McCouch et al. (1997). In addition, the abbreviation was used for the loci associated with the traits at the different stages. Thus, *qPH(NN)(1,2,3)-10-1* indicated a locus located on chromosome 10 associated with both PH and NN at all three stages.

Prediction of Candidate Genes

Genes annotated in *G. max* Williams 82 reference gene model 1.0 were used as the source of candidate genes. The prediction of candidate genes mainly referred to the genes with a known function in soybeans related to the trait or the orthologs in *Arabidopsis*.

TABLE 1 | Descriptive statistics for plant height (cm) and number of nodes on the main stem at three stages over two environments in the SBL population.

Trait	Stage	Mean	*SD*	Minimum	Maximum	Skewness	Kurtosis	CV (%)	F_G	F_E	$F_{R(E)}$	$F_{G \times E}$	h^2 (%)
PH	1	50.48	9.95	27.97	90.22	0.80	1.06	19.70	9.99***	84.43***	24.33***	1.53***	84.97
	2	65.96	10.41	39.53	109.00	0.53	1.17	15.78	6.90***	271.97***	35.33***	1.63***	76.57
	3	74.39	12.45	41.28	109.06	0.13	−0.12	16.74	8.14***	664.00***	36.17***	1.85***	77.70
NN	1	10.21	1.32	7.67	15.94	0.84	0.74	12.88	7.39***	52.21***	11.19***	1.28**	83.10
	2	12.86	1.29	8.97	16.89	0.21	0.25	10.06	5.59***	717.53***	19.38***	2.27***	60.00
	3	14.58	1.59	9.58	20.39	0.16	0.74	10.91	5.41***	2323.61***	17.10***	1.76***	68.07

*F_G, F_E, $F_{R(E)}$ and $F_{G \times E}$ represent the F-values for genotypic, environmental, block effects and genotype × environment interaction, respectively. Broad-sense heritability: $h^2 = \sigma_g^2 / \left(\sigma_g^2 + \sigma_{ge}^2/n + \sigma_e^2/nr\right)$, where σ_g^2 is the genotypic variance, σ_{ge}^2 is the genotype by environment interaction variance, σ_e^2 is the error variance, n is the number of environments, r is the number of replications. **$P < 0.001$, ***$P < 0.0001$.*

RESULTS

Trait Performance of the Tested Population

The phenotypic characteristics of PH and NN for the 368 soybean lines are shown in **Table 1**. Averaged over two environments, PH and NN showed a large variation at the three different stages with range values of 27.97–109.06 (cm) and 7.67-20.39, respectively. The absolute values of kurtosis and skewness were approximately 1 for both PH and NN (**Supplementary Figure S1**). Significant positive correlations were observed for PH and NN among the three stages and between PH and NN (**Supplementary Figure S2**) ($r > 0.60$, $P < 0.0001$). PH1 and PH3 were moderately correlated with NN3 and NN1 at $r = 0.49$ and 0.48, $P < 0.0001$, respectively. Analysis of variance (ANOVA) indicated that there were significant differences in the effects of genotypes, environments, and their interactions for the traits at all stages. Additionally, a relatively high heritability (≥60%) was estimated for PH and NN at all stages, indicating that the genetic effects play a primary role in the performance of PH and NN.

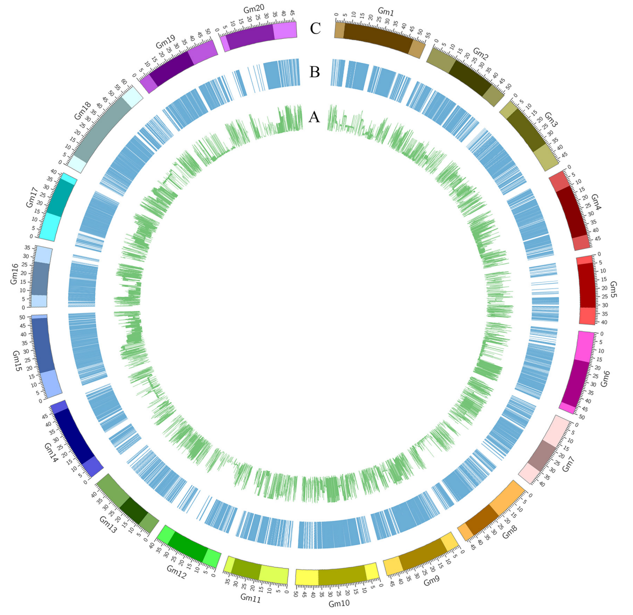


FIGURE 1 | Characterization of the SNPs in the soybean genome. **(A)** Minor allele frequency of SNPs across the whole genome. **(B)** Distribution of LD blocks (>50 kb) in the whole genome. **(C)** Chromosomal region with pericentromeric regions in a darker color and whole chromosome in a lighter color (distance unit is Mb).

Characterization of the SNPs, Population Structure, LD, and LD Haplotype Block Estimation

A total of 62423 SNPs with an MAF ≥ 0.05 were used for further analyses, with an average marker density of 1 SNP every 16.42 kb genome-wide, varying across chromosomes from 29.39 kb per SNP on chromosome 5 to 10.51 kb per SNP on chromosome 15. The MAF and haplotype block (>50 kb) for the population characteristics are presented in **Figure 1**. The most likely *K*-value was $K = 3$ based on the analysis of population structure (**Figure 2A**), which suggested that the overall population could be divided into three subpopulations. This result was also supported by the phylogenetic analysis (**Figures 2B,C**) and PCA (**Figure 2D**). The LD decay rate of the population was estimated at 400 kb in euchromatin, where r^2 dropped to half of its maximum value ($r^2 = 0.23$) (**Figure 3**). In heterochromatin, however, r^2 did not drop to half of its maximum value until 3.5 Mb. The haplotype analysis showed that 62423 SNPs were grouped into 5697 haplotype blocks. The size of the blocks ranged from 6 bp to 200 kb across the whole genome. The distribution of haplotype blocks is shown in **Supplementary Figure S3**.

GWAS for the Traits via CMLM

Genome-wide association studies was conducted using the BLUPs of individual performance over two environments. A total of 11 loci for PH and 13 loci for NN were identified in the CMLM association panel at the suggestive significance level ($P = 1 \times 10^{-4}$). There were 8, 3, and 4 loci for PH and 6, 6, and 8 loci for NN at the three developmental stages. Three and six loci were detected for PH and NN at more than two stages, respectively (**Table 2** and **Figure 4**). Notably, two major loci associated with PH and NN were identified under the Bonferroni correction for multiple tests (0.05/*N*) (bold in **Table 2**). *qPH(NN)(1,2,3)-10-1* was identified on chromosome 10 at all stages, while *qPH(NN)3-19-1* was identified on chromosome 19 only at the last stage.

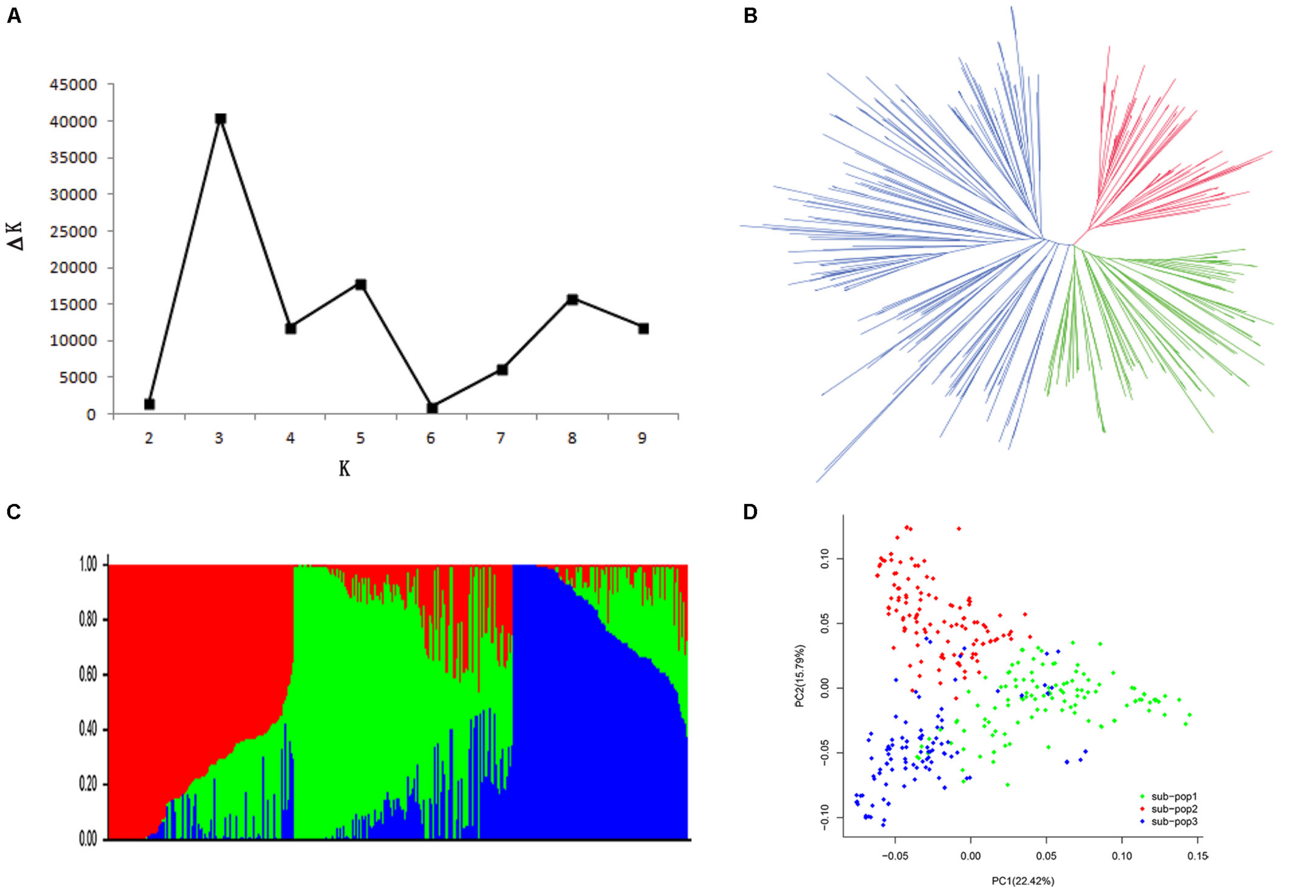


FIGURE 2 | Population structure of 368 soybean accessions. **(A)** Calculation of the true *K* of the SBL population according to Evanno et al. (2005). **(B)** A neighbor-joining tree of the tested accessions that could be divided into three subpopulations. **(C)** Population structure was estimated by ADMIXTURE. Three colors represent three subpopulations, respectively. Each vertical column represents one individual and each colored segment in each column represents the percentage of the individual in the population. **(D)** PCA plot of the 368 accessions; two-dimensional scales were used to reveal population stratification.

The haplotype analysis showed that the peak SNPs for these two major loci were located within the two haplotype blocks named H2842 and H5441. Three main haplotypes (total frequency > 80%), H2842-1, H2842-2, and H2842-3, were identified for H2842, whose frequencies were 19.02, 58.42, and 10.87%, respectively, and two main haplotypes, H5441-1 and H5441-2, were identified for H5441, with frequencies of 46.47% and 33.70%, respectively. We further analyzed the effect of these haplotypes for PH and NN. The results showed that there were significant differences for both PH and NN among the main haplotypes of H2842 or H5441 (**Supplementary Figure S4**). The LD and haplotype blocks for these two major loci are presented in **Figure 5**.

GWAS for Traits via mrMLM

To validate the reliability of the loci determined by the CMLM method and identify more loci associated with PH and NN, a multi-locus random effect MLM method was used to conduct GWAS.

Thirty-four loci for PH and 30 loci for NN were identified using the MRMLM method. Among them, 11, 13, and 16 loci for PH and 11, 15, and 8 loci for NN were detected at the three developmental stages (**Tables 3**, **4** and **Figure 6**). Due to the differential genetic control of PH and NN at the different stages, further comparative analysis showed that *qPH1(2)-10-1*, *qPH1(2)-13-5* and *qPH2(3)-7-1*, *qPH2(3)-9-3*, *qPH2(3)-10-1* and *qPH2(3)-13-1* were commonly detected for PH at the first and last two stages, respectively (bold in **Table 3**). *qNN1(2)-10-1*, *qNN1(2)-14-1* and *qNN2(3)-10-1*, *qNN2(3)-13-3* were commonly detected for NN at the first and last stages, respectively (bold in **Table 4**). It was clear that there were many differential loci for both PH and NN among the different stages. Similar to the CMLM results, *qPH (NN)(1,2,3)-10-1* as a major locus was identified and shared for PH and NN at three stages. The same peak SNP (Gm10_44346474, MAF = 0.41) could explain 13, 8, and 10% of the phenotypic variation for PH and 7, 4, and 11% of the phenotypic variation for NN at the three stages, respectively. *qPH (NN)3-19-1(3)*, another major locus, was detected for both PH and NN at the last stage. The same peak SNP (Gm19_44938780, MAF = 0.35) could explain 8 and 9% of the phenotypic variation for PH and NN at the last stage, respectively. In addition to these two major loci, *qPH1-4-1*, *qPH1-6-2,* and

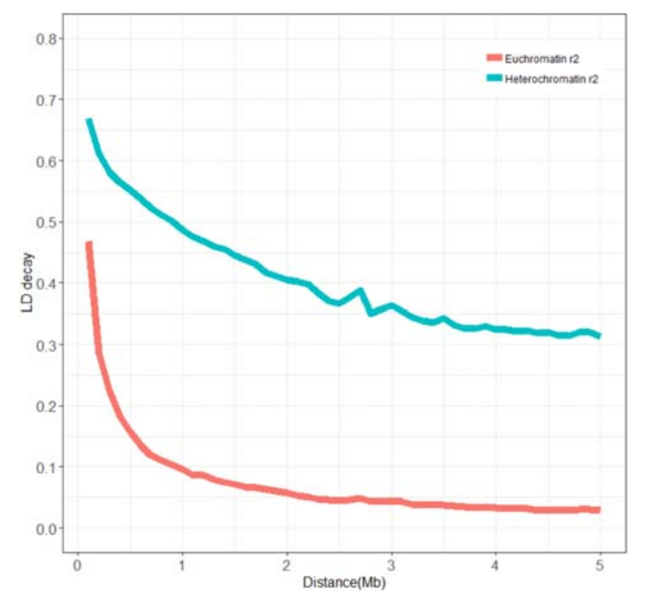


FIGURE 3 | Average LD decay rates in euchromatic and heterochromatic regions of the whole genome. The mean LD decay rate was estimated as the squared correlation coefficient (r^2) using all pairs of SNPs located within 5 Mb of physical distance in euchromatic (red) and heterochromatic (green) regions in the SBL population.

qPH1-14-1 for PH and *qNN1-4-1*, *qNN2-13-3,* and *qNN1-14-1* for NN were validated by the MRMLM method. Moreover, four loci, *qPH1-4-1*, *qPH1-6-2*, *qPH2-13-1,* and *qPH1-14-1* for PH and *qNN1-4-1*, *qNN1-6-1*, *qNN3-13-2,* and *qNN1(2)-14-1* for NN, were co-located in the same genomic regions. Some novel loci were identified for PH and NN via the MRMLM method compared with the CMLM method and are shown in **Tables 3**, **4**.

Prediction of Candidate Genes

To predict candidate genes for loci significantly associated with both PH and NN, we selected the putative genes tagged by the most significant SNPs. Two major loci coassociated with PH and NN pinpointed the known genes. *E2*(*Glyma10g36600*), encoding a homolog of GIGANTEA (*GI*) protein, is one of the key genes regulating soybean flowering and maturity by regulating *GmFT2a* (Watanabe et al., 2011). It was found 370 kb upstream of the peak SNP (Gm10_44346474, MAF = 0.41) of *qPH1-10-1* (*qNN1-10-1*) on Gm10, which was associated with both PH and NN at all three stages (**Tables 3**, **4**). *Dt1* (*Glyma19g37890*) is a homolog of *Arabidopsis TERMINAL FLOWER1* (*TFL1*) and plays a primary role in the soybean stem growth habit (Bernard, 1972). It was found 41 kb upstream of the peak SNP (Gm19_44938780, MAF = 0.35) of *qPH3-19-1* (*qNN3-19-3*) on Gm19, which was associated with both PH and NN at the last stage. In addition to these two major loci, we predicted the candidate genes for two other loci associated with both PH and NN. A putative gene, *Glyma04g40640,* was found 35 kb from the peak SNP (Gm04_46671367, MAF = 0.11) of *qPH1-4-1* (*qNN1-4-1*) on Gm04, which was associated with both PH and NN at the first stage. It was homologous to *Arabidopsis APRR5,* which is involved in various circadian-associated biological events such as flowering time in long-day photoperiod conditions and red light sensitivity of seedlings during early photomorphogenesis (Kamioka et al., 2016). The putative gene *Glyma13g06811* was identified 125 kb from the peak SNP (Gm13_6859748, MAF = 0.46) of *qPH2-13-1*(*qNN3-13-2*) on Gm13, which was associated with both PH and NN at the second and third stages, respectively. It is homologous to *Arabidopsis AGAMOUS-LIKE 8* (*AGL8*), which is a MADs-box transcription factor involved in various biological events such as flower and fruit development and the maintenance of inflorescence meristem characteristics (Ma et al., 1991).

DISCUSSION

Linkage Disequilibrium and the Statistical Method Are of Great Significance in GWAS

Genome-wide association studies are a powerful tool to elucidate the genetic architecture for complex quantitative traits in crops (Morris et al., 2013; Su et al., 2016; Sun et al., 2017). The mapping resolution and statistical power are the main considerations in GWAS. LD is one of the factors limiting the mapping resolution of GWAS. The recombination rate is one of the major factors affecting LD extension and is different between euchromatic and heterochromatic regions (Zhang et al., 2015). Previous studies suggested that there was a very diverse range of LD values in different crops and different chromosomal regions in a specific crop (Li Y.H. et al., 2014; Sonah et al., 2015; Zhang et al., 2015). Similar to previous studies, a large difference in LD decay rate was also observed between the euchromatin (400 kb) and heterochromatin regions (3.5 Mb) of soybeans in this study. A longer LD has been observed in self-pollinated crops such as soybean compared to cross-pollinated crops such as maize (Li et al., 2016). A previous study also reported a longer LD for the chromosomes involved in the domestication process. The QTLs for domestication traits such as seed weight and flowering were mapped in these regions (Li Y.H. et al., 2013).

Another concern with GWAS is the statistical method. In the present study, the CMLM method was used for the single-locus GWAS. Consistent with the previous study, we also found that the CMLM method could take less computing time than regular MLM and reduce false positive results simultaneously (Zhang et al., 2015). However, the single-locus GWAS methods often need correction for multiple tests. For instance, the typical Bonferroni correction corrects an $\alpha = 0.05$ to $\alpha = 0.05/m$, where m is the number of statistical tests performed. For a GWAS with 500,000 markers, the statistical significance threshold for an association would be corrected to $1e^{-7}$, such that no or a few loci could reach the significance threshold

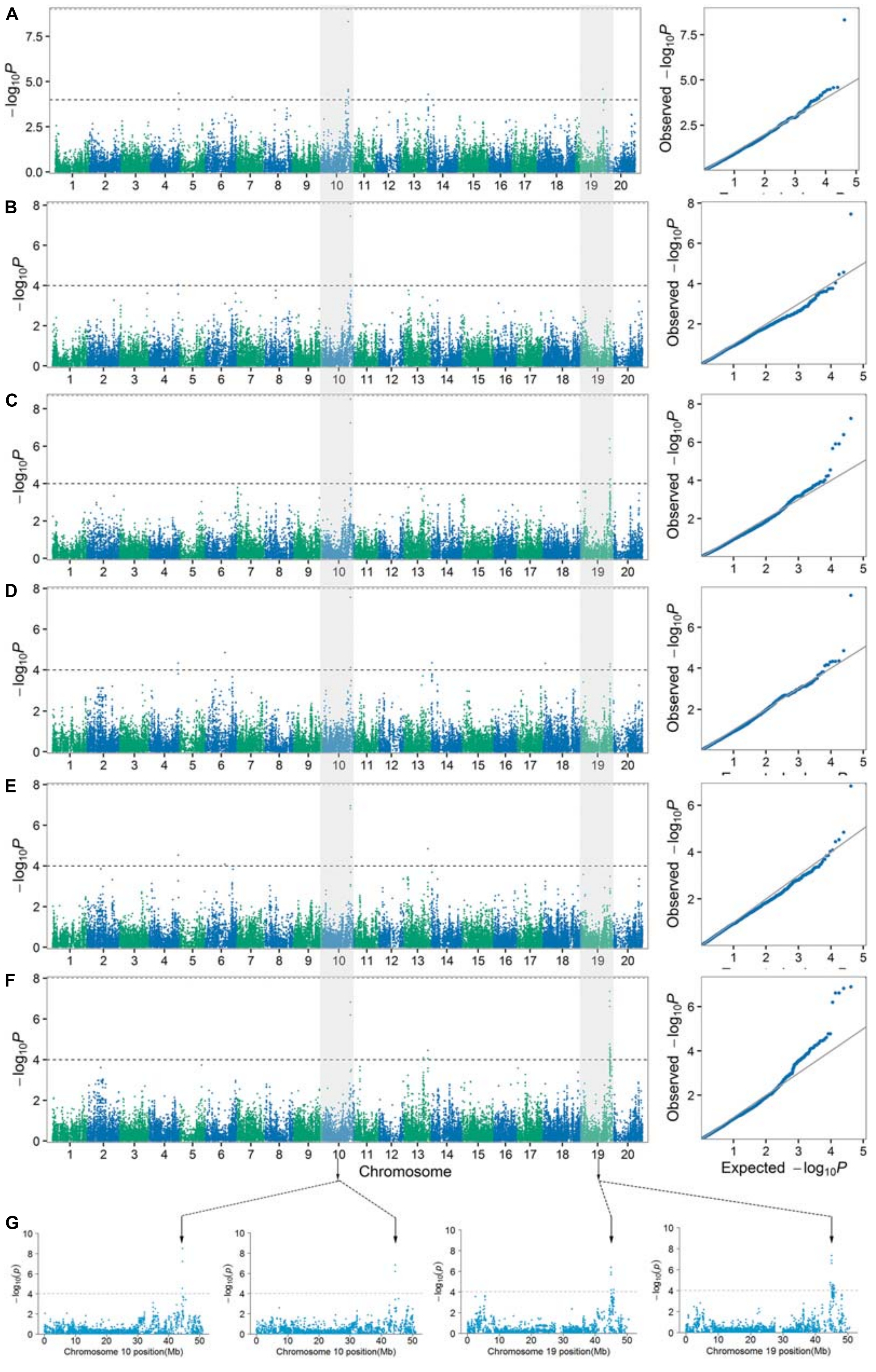


FIGURE 4 | Manhattan plots and quantile–quantile plots for PH1 **(A)**, PH2 **(B)**, PH3 **(C)**, NN1 **(D)**, NN2 **(E)**, and NN3 **(F)** over three stages in the SBL population. The major loci for PH3 (left) and NN3 (right) in their Manhattan plots were located on chromosomes 10 and 19, respectively **(G)**.

TABLE 2 | Quantitative trait nucleotides (QTNs) associated with PH and NN via CMLM in the SBL population.

Trait	QTN	Chr	SNP	Position (bp)	Allele	MAF	$-\text{Log}_{10}$ *P*[a]	R^{2}[b]	R^{2}[c]	Effect	Known QTL$_{S}$[d]
PH	*qPH1-4-1*	4	Gm04_46680158	46680158	T/C	0.10	4.35	0.35	0.38	−3.71	5-4
	qPH1-6-1	6	Gm06_43710729	43710729	G/A	0.27	4.15	0.35	0.38	−3.08	3-2,6-3
	qPH1-10-1	10	Gm10_44226599	44226599	A/C	0.13	4.48	0.35	0.38	3.03	18-2,19-2
	qPH1-10-2[e]	10	Gm10_44346474	44346474	A/T	0.41	9.26	0.35	0.42	3.57	*E2*
	qPH1-10-3	10	Gm10_44550928	44550928	C/T	0.33	4.45	0.35	0.38	−2.36	18-2,19-2
	qPH1-10-4	10	Gm10_45479097	45479097	C/T	0.11	4.13	0.35	0.38	3.31	
	qPH1-14-1	14	Gm14_926343	926343	G/A	0.21	4.29	0.35	0.38	2.40	34-6
	qPH1-19-1	19	Gm19_44544574	44544574	A/G	0.36	4.58	0.35	0.38	−2.53	1-1
	qPH2-4-1	4	Gm04_46680158	46680158	T/C	0.10	4.04	0.29	0.32	−3.36	5-4
	qPH2-10-2	10	Gm10_44346474	44346474	A/T	0.41	8.07	0.29	0.36	3.13	*E2*
	qPH2-10-3	10	Gm10_44550928	44550928	C/T	0.33	4.45	0.29	0.32	−2.24	18-2,19-2
	qPH3-10-2	10	Gm10_44346474	44346474	A/T	0.41	8.52	0.25	0.33	3.80	*E2*
	qPH3-19-2	19	Gm19_44938780	44938780	C/T	0.35	6.39	0.25	0.31	3.74	*Dt1*
	qPH3-19-3	19	Gm19_45721414	45721414	G/T	0.33	4.25	0.25	0.29	−2.62	3-1,4-2,6-1
	qPH3-20-1	20	Gm20_41211643	41211643	G/T	0.12	4.02	0.25	0.29	3.03	28-1
NN	*qNN1-4-1*	4	Gm04_46680158	46680158	T/C	0.10	4.33	0.34	0.37	−0.46	
	qNN1-6-1	6	Gm06_31222865	31222865	C/T	0.10	4.85	0.34	0.38	−0.52	4-2
	qNN1-10-1	10	Gm10_44346474	44346474	A/T	0.41	7.99	0.34	0.40	0.41	*E2*
	qNN1-14-1	14	Gm14_926343	926343	G/A	0.21	4.35	0.34	0.37	0.30	
	qNN1-18-1	18	Gm18_3716679	3716679	T/C	0.21	4.32	0.34	0.37	0.39	
	qNN1-19-1	19	Gm19_45415096	45415096	C/A	0.20	4.30	0.34	0.37	−0.34	
	qNN2-4-1	4	Gm04_46680158	46680158	T/C	0.10	4.53	0.29	0.33	−0.27	
	qNN2-6-1	6	Gm06_31222865	31222865	C/T	0.10	4.08	0.29	0.32	−0.28	4-2
	qNN2-10-1	10	Gm10_44346474	44346474	A/T	0.41	6.95	0.29	0.35	0.22	*E2*
	qNN2-10-2	10	Gm10_45479097	45479097	C/T	0.11	4.44	0.29	0.33	0.25	
	qNN2-13-1	13	Gm13_38510582	38510582	T/G	0.13	4.84	0.29	0.33	0.21	
	qNN2-14-1	14	Gm14_926343	926343	G/A	0.21	4.02	0.29	0.32	0.17	
	qNN3-10-1	10	Gm10_44346474	44346474	A/T	0.41	6.83	0.25	0.31	0.11	*E2*
	qNN3-13-2	13	Gm13_31053641	31053641	T/C	0.25	4.09	0.25	0.29	−0.09	1-8
	qNN3-13-1	13	Gm13_38510582	38510582	T/G	0.13	4.46	0.25	0.29	0.10	
	qNN3-19-2	19	Gm19_44558007	44558007	C/A	0.34	4.77	0.25	0.29	−0.09	
	qNN3-19-3	19	Gm19_44938780	44938780	C/T	0.35	7.36	0.25	0.32	0.13	*Dt1*
	qNN3-19-4	19	Gm19_45295148	45295148	C/G	0.39	4.61	0.25	0.29	0.09	
	qNN3-19-1	19	Gm19_45384848	45384848	A/G	0.40	4.10	0.25	0.29	−0.08	
	qNN3-19-5	19	Gm19_45727395	45727395	G/A	0.39	4.48	0.25	0.29	−0.09	

[a]*Negative* $\log_{10}$*-transformed P-value of the suggestive.* [b]*The contribution rate of the model without SNP.* [c]*The contribution rate of the model with SNP.* [d]*Based on the QTL list on SoyBase (www.soybase.org).* [e]*The QTNs shown in bold indicate that they reach the significance threshold based on the Bonferroni correction.*

after the correction. Such a situation is not always suited to the nature of complex traits. A previous study showed that no significantly associated locus for soybean seed weight was detected, possibly for this reason (Fang et al., 2017). Wang et al. (2016) also reported that some small-effect loci were not significantly associated with the traits in the single-locus approach under the Bonferroni correction but significantly associated with that in the mrMLM method. Actually, the small-effect loci should not be neglected in the genetic system of complex traits. Fortunately, several multi-locus GWAS methods have already been reported in previous studies (Segura et al., 2012; Wang et al., 2016), where no Bonferroni correction is needed because of the multi-locus nature. Thus, the multi-locus GWAS method may play an increasingly important role in dissecting the genetic architecture of complex traits in the post-association time.

Combination of CMLM and mrMLM GWAS Methods to Identify the Major and Minor Loci for PH and NN in the SBL Population

Plant height and NN are quantitative traits controlled by numerous loci in soybeans[2]. In the present study, 11 loci for PH and 13 loci for NN were detected via the CMLM method at the suggestive threshold level, only two of which were identified after the Bonferroni correction. To confirm the loci determined by the CMLM method and identify additional loci for PH and NN, mrMLM as a multi-locus method was used for GWAS in this study. As expected, we found 32 and 28 loci, except for two major loci, for PH and NN via the

[2] www.soybase.org

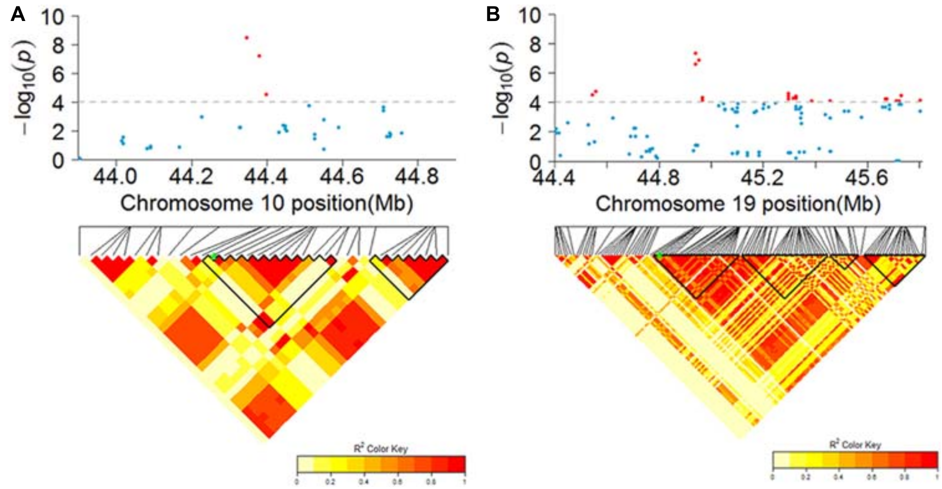


FIGURE 5 | Candidate regions of the genome showing significant association signals near identified major loci for PH and NN. The top of each panel shows the Manhattan plot indicating the level of SNP association with PH **(A)** or NN **(B)**. Gray horizontal dashed lines indicate the genome-wide suggestive threshold. The bottom of each panel shows the local LD of the chromosomal regions containing the peak SNP (SNP with the lowest *P*-value), whose position is indicated by a green asterisk. Nearby haplotype blocks are outlined in black triangles.

mrMLM method. Some novel loci in comparison with that obtained from the CMLM were located around the previously reported QTLs. For PH, the loci *qPH1-18-1* and *qPH1-18-2* detected by the mrMLM method have been previously identified by Sun et al. (2006), which were also located in the same genomic regions of the known QTLs. However, only six loci in addition to the two major loci, *qPH1-4-1*, *qPH1-6-1,* and *qPH1-14-1* for PH and *qNN1-4-1*, *qNN1-14-1,* and *qNN2-13-1* for NN, identified at a suggestive threshold level ($P = 10^{-4}$) in CMLM were validated in mrMLM association panel. One potential reason was that the mrMLM method improved the power and accuracy for QTN detection due to the nature of the statistical model. Thus, many novel loci were detected by the mrMLM method. Another possible reason was that the relatively stringent threshold still applied to the CMLM method. More loci might be commonly detected in two association panels if a lower threshold was adopted in the CMLM method, but the false positive results might increase under such conditions. Undoubtedly, mrMLM can identify not only the major loci but also the minor loci for quantitative traits compared with the CMLM method.

Genetic System of Dynamic PH and NN in Summer Planting Soybeans

Many agronomic and yield-related traits such as PH and NN were highly correlated in soybeans. Previously identified loci for these traits usually co-located in the same genomic regions (Malik et al., 2007; Liu et al., 2011, 2017; Fang et al., 2017). Similarly, significant positive correlations were observed between PH and NN at different stages in this study. Furthermore, six loci, including major and minor loci, were shared for PH and NN, suggesting that both PH and NN have a similar genetic system controlled by major and minor loci. Previous studies showed that the *E2* and *Dt1* loci have an effect on many agronomic and yield-related traits in soybean (Liu et al., 2011; Kato et al., 2015; Zhang et al., 2015). We also found that PH and NN shared these two loci and further confirmed that the genetic pattern of the *E2* locus was different from that of the *Dt1* locus. The former was detected for both PH and NN at all stages while the latter was only detected at the last stage.

Plant height and NN in soybeans are dynamic traits, as the phenotype changes constantly during the plant lifecycle. However, studies of the genetic basis of PH and NN have mainly measured the final PH and number of nodes on the main stem, especially when the phenotype was investigated at the mature stage. We accessed PH and NN in the summer-planting accessions at three different stages, which included one vegetative stage and two reproductive stages when plants were in late vegetative growth, as well as the flowering and mature stages, to reveal the genetic control underlying the dynamic PH and NN based on the GWAS strategy in this study. A previous report showed that the haplotype was more appropriate than the single SNP to uncover the genetic variation and improve the efficiency of breeding for target traits due to the existence of multiple alleles (Hao et al., 2012). We identified the *E2* and *Dt1* genes in two haplotype blocks. The analysis of haplotypes revealed that the main haplotypes of these two haplotype blocks were related to PH and NN over stages. Our results suggested that *E2* and *Dt1,* as the major loci, play different roles in regulating the development of PH and NN at different stages. Nine and 20 novel loci were identified for PH and NN at different stages via a new multi-locus GWAS method, respectively. Moreover, the differential loci were identified for both PH and NN among the different stages. These common and specific loci for PH and NN at different stages unveil the genetic architecture underlying the dynamic PH and NN. Although most of them have not been confirmed yet, candidate genes were predicted for several loci

TABLE 3 | QTNs associated with PH via mrMLM in the SBL population.

QTN	Chr	SNP	Allele	MAF	Position (bp)	Effect[a]	LOD score[b]	R^2	Known QTL[c]
qPH1-4-1[d]	4	Gm04_46680158	T/C	0.10	46680158	−4.42	9.43	0.11	5-4
qPH1-5-1	5	Gm05_36588358	G/A	0.23	36588358	2.11	4.84	0.05	24-1
qPH1-6-1	6	Gm06_33125042	T/C	0.07	33125042	2.89	4.48	0.03	30-1
qPH1-6-2	6	Gm06_43710729	G/A	0.27	43710729	−2.52	5.64	0.07	3-2,6-3
qPH1-6-3	6	Gm06_48441344	A/G	0.19	48441344	1.76	6.01	0.03	38-1
qPH1-9-1	9	Gm09_38680990	C/T	0.45	38680990	1.20	3.73	0.02	
qPH1-10-1[e]	10	Gm10_44346474	A/T	0.41	44346474	3.04	14.99	0.13	*E2*
qPH1-13-5	13	Gm13_41776586	C/T	0.05	41776586	3.04	4.11	0.03	39-1
qPH1-14-1	14	Gm14_1243675	A/G	0.21	1243675	1.92	4.59	0.04	34-6
qPH1-18-1	18	Gm18_7790416	T/C	0.34	7790416	−0.91	3.61	0.01	23-6
qPH1-18-2	18	Gm18_51243803	A/G	0.22	51243803	−1.66	3.26	0.03	26-12
qPH2-6-4	6	Gm06_13482496	G/A	0.36	13482496	−1.37	3.40	0.03	
qPH2-7-1	7	Gm07_37502305	T/C	0.35	37502305	1.59	4.15	0.04	37-6
qPH2-8-1	8	Gm08_41977134	G/A	0.42	41977134	1.30	3.01	0.03	
qPH2-9-2	9	Gm09_7987471	C/T	0.23	7987471	1.60	5.16	0.03	mqPH-009
qPH2-9-3	9	Gm09_41284917	G/A	0.05	41284917	2.42	3.55	0.02	
qPH2-10-1	10	Gm10_44346474	A/T	0.41	44346474	2.21	7.67	0.08	*E2*
qPH2-13-1	13	Gm13_6859748	A/G	0.46	6859748	1.37	4.05	0.03	20-5
qPH2-13-2	13	Gm13_8246449	T/C	0.48	8246449	2.04	8.31	0.07	20-5
qPH2-13-3	13	Gm13_10898897	T/C	0.50	10898897	−1.57	5.56	0.04	26-11
qPH2-13-4	13	Gm13_28457573	T/G	0.06	28457573	−6.23	9.92	0.13	5-8,15-1
qPH2-13-5	13	Gm13_41776586	C/T	0.05	41776586	3.49	5.58	0.04	39-1
qPH2-14-2	14	Gm14_3056277	T/C	0.19	3056277	2.04	6.30	0.04	34-6
qPH2-20-1	20	Gm20_41211643	G/T	0.12	41211643	2.12	4.68	0.03	28-1
qPH3-2-1	2	Gm02_2328263	C/A	0.20	2328263	−3.03	9.85	0.06	9-3
qPH3-2-2	2	Gm02_46786754	C/T	0.47	46786754	−1.67	3.98	0.03	6-12
qPH3-3-1	3	Gm03_1920671	A/G	0.26	1920671	1.70	4.19	0.02	
qPH3-3-2	3	Gm03_6757344	T/C	0.26	6757344	1.63	6.27	0.02	
qPH3-6-5	6	Gm06_31222865	C/T	0.10	31222865	2.25	4.44	0.02	30-1
qPH3-7-2	7	Gm07_984279	C/G	0.09	984279	2.56	3.46	0.02	
qPH3-7-1	7	Gm07_37502305	T/C	0.35	37502305	1.88	3.21	0.04	37-6
qPH3-8-2	8	Gm08_18166829	A/C	0.46	18166829	1.46	5.53	0.02	
qPH3-9-3	9	Gm09_41284917	G/A	0.05	41284917	2.93	3.96	0.02	
qPH3-10-2	10	Gm10_34971279	C/G	0.13	34971279	2.99	8.06	0.04	
qPH3-10-1	10	Gm10_44346474	A/T	0.41	44346474	3.02	15.21	0.10	*E2*
qPH3-13-1	13	Gm13_6859748	A/G	0.46	6859748	1.93	6.83	0.04	20-5
qPH3-13-6	13	Gm13_27311661	G/C	0.29	27311661	2.89	9.19	0.08	5-8,15-1
qPH3-14-3	14	Gm14_10617104	A/G	0.39	10617104	1.32	4.43	0.02	34-6
qPH3-19-1	19	Gm19_44938780	C/T	0.35	44938780	−2.81	14.27	0.08	*Dt1*
qPH3-20-2	20	Gm20_27547938	G/A	0.33	27547938	3.00	12.48	0.09	16-1

[a] *Quantitative trait nucleotide effect.* [b] *LOD value, the significant threshold for the transformed P-value.* [c] *The QTLs located in the same region of reported QTLs.* [d] *The QTN identified for PH at the same stage by both methods are underlined in Table.* [e] *The QTNs identified for PH at least two stages are displayed in bold in Table.*

associated with both PH and NN, some of which are located in the same regions as known QTLs for PH and NN in SoyBase. Further studies should confirm these loci and identify candidate genes for them.

CONCLUSION

More loci, including 34 loci for PH and 30 loci for NN, were identified by the mrMLM method than by the single-locus CMLM method. A few loci were commonly identified for PH and NN via the two methods at the different developmental stages. One stable locus that overlapped with the *E2* gene was identified for PH and NN at all three stages, while another major locus, referred to as the *Dt1* gene, was determined at the last stage by both methods. Most loci were mainly detected at only one or two of the examined developmental stages. The dynamic PH and NN was controlled by a set of specific loci and a few common loci in summer planting soybeans.

TABLE 4 | QTNs associated with NN via mrMLM in the SBL population.

QTN	Chr	SNP	Allele	MAF	Position (bp)	QTN effect	LOD score	R^2	Known QTL
qNN1-2-1	2	Gm02_5017572	G/A	0.48	5017572	0.25	7.62	0.07	
qNN1-4-1[a]	4	Gm04_46680158	T/C	0.10	46680158	−0.37	7.84	0.05	
qNN1-5-1	5	Gm05_27229418	A/G	0.32	27229418	0.22	3.62	0.04	
qNN1-6-1	6	Gm06_43710729	G/A	0.27	43710729	−0.26	3.90	0.05	1-3,1-4,2-1
qNN1-10-1[b]	10	Gm10_44346474	A/T	0.41	44346474	0.27	8.56	0.07	*E2*
qNN1-10-2	10	Gm10_45479097	C/T	0.11	45479097	−0.42	6.69	0.07	
qNN1-14-1	14	Gm14_1243675	A/G	0.21	1243675	0.20	4.48	0.03	
qNN1-15-1	15	Gm15_7957092	T/G	0.31	7957092	−0.25	5.99	0.06	
qNN1-17-1	17	Gm17_38445432	T/G	0.23	38445432	0.18	3.93	0.02	
qNN1-17-2	17	Gm17_38559614	T/C	0.07	38559614	−0.28	3.24	0.02	
qNN1-19-1	19	Gm19_2487047	C/T	0.11	2487047	−0.27	3.60	0.03	
qNN2-4-2	4	Gm04_9024569	T/C	0.39	9024569	0.16	8.26	0.07	5-1
qNN2-6-2	6	Gm06_14119587	C/T	0.06	14119587	0.18	3.28	0.02	7-2
qNN2-6-3	6	Gm06_17601986	C/A	0.18	17601986	−0.18	6.57	0.06	7-2
qNN2-6-4	6	Gm06_41843359	C/A	0.45	41843359	−0.09	3.02	0.02	1-3,1-4,2-1
qNN2-8-1	8	Gm08_18313651	T/A	0.17	18313651	0.13	4.01	0.03	
qNN2-10-3	10	Gm10_4827909	G/A	0.08	4827909	−0.16	4.01	0.03	
qNN2-10-1	10	Gm10_44378814	T/C	0.41	44378814	0.12	4.39	0.04	*E2*
qNN2-12-1	12	Gm12_8229718	C/T	0.23	8229718	0.11	3.08	0.02	
qNN2-13-1	13	Gm13_7126248	C/A	0.42	7126248	0.13	6.71	0.05	1-6
qNN2-13-3	13	Gm13_38510582	T/G	0.13	38510582	0.14	3.73	0.02	
qNN2-14-1	14	Gm14_1243675	A/G	0.21	1243675	0.12	3.99	0.03	
qNN2-14-2	14	Gm14_1557061	C/T	0.20	1557061	0.12	3.06	0.03	
qNN2-18-1	18	Gm18_8519411	C/T	0.34	8519411	0.16	6.78	0.07	
qNN2-19-2	19	Gm19_45415096	C/A	0.20	45415096	0.15	4.77	0.04	
qNN2-20-1	20	Gm20_40247542	G/A	0.32	40247542	0.09	3.24	0.02	
qNN3-6-5	6	Gm06_50072605	A/G	0.46	50072605	0.05	5.68	0.04	1-3
qNN3-10-1	10	Gm10_44346474	A/T	0.41	44346474	0.09	11.42	0.11	*E2*
qNN3-11-1	11	Gm11_8556480	G/A	0.22	8556480	−0.08	6.94	0.06	3-3,3-4,3-5,3-6
qNN3-13-2	13	Gm13_6859748	A/G	0.46	6859748	0.05	5.09	0.03	1-6
qNN3-13-3	13	Gm13_38510582	T/G	0.13	38510582	0.08	4.30	0.03	
qNN3-14-3	14	Gm14_11281151	T/C	0.40	11281151	0.08	7.22	0.08	
qNN3-16-1	16	Gm16_27302091	A/T	0.18	27302091	0.06	4.28	0.03	
qNN3-19-3	19	Gm19_44938780	C/T	0.35	44938780	−0.09	9.66	0.09	*Dt1*

[a] *The QTNs identified for NN at the same stage by both methods are underlined in Table.* [b] *The QTNs identified for NN at least two stages are displayed in bold in Table.*

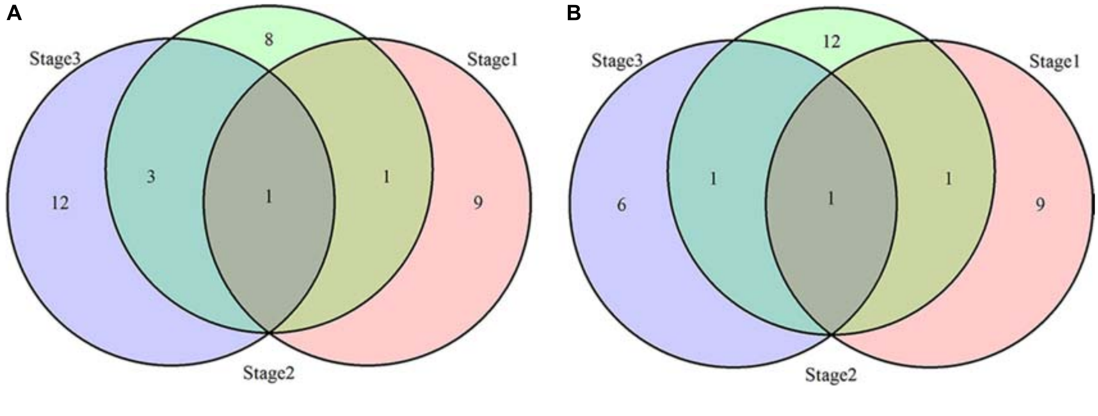


FIGURE 6 | Venn diagrams for loci associated with PH **(A)** and NN **(B)** over three stages.

AUTHOR CONTRIBUTIONS

TZ conceived and designed the experiments. CG, FS, JZ, JK, and QH performed the experiments. FC and ZW analyzed the data. FC and RS drafted the manuscript. TZ revised the paper.

FUNDING

This work was supported by grants from the National Key R&D Program for Crop Breeding (2016YFD0100304), the National Natural Science Foundation of China (Grant Nos. 31571691 and 31271750), the MOE Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT_17R55), the Fundamental Research Funds for the Central Universities (KYT201801), the Jiangsu Collaborative Innovation Center for Modern Crop Production (JCIC-MCP) Program, and Postgraduate Research and Practice Innovation Program of Jiangsu Province (KYCX18_0655).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01184/full#supplementary-material

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Atwell, S., Huang, Y. S., Vilhjalmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465, 627–631. doi: 10.1038/nature08800

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bernard, R. L. (1972). Two genes affecting stem termination in soybeans. *Crop Sci.* 12, 235–239. doi: 10.2135/cropsci1972.0011183X001200020028x

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. doi: 10.1093/bioinformatics/btm308

Cao, Y., Li, S., He, X., Chang, F., Kong, J., Gai, J., et al. (2017). Mapping QTLs for plant height and flowering time in a Chinese summer planting soybean RIL population. *Euphytica* 213:39. doi: 10.1007/s10681-016-1834-8

Chang, H. X., and Hartman, G. L. (2017). Characterization of insect resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Front. Plant Sci.* 8:670. doi: 10.3389/fpls.2017.00670

Chapman, A., Pantalone, V. R., Ustun, A., Allen, F. L., Landau-Ellis, D., Trigiano, R. N., et al. (2003). Quantitative trait loci for agronomic and seed quality traits in an F2) and F4:6 soybean population. *Euphytica* 129, 387–393. doi: 10.1023/A:1022282726117

Cho, S., Kim, K., Kim, Y. J., Lee, J. K., Cho, Y. S., Lee, J. Y., et al. (2010). Joint identification of multiple genetic variants via Elastic-Net variable selection in a genome-wide association analysis. *Ann. Hum. Genet.* 74, 416–428. doi: 10.1111/j.1469-1809.2010.00597.x

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 18:161. doi: 10.1186/s13059-017-1289-9

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *Science* 296, 2225–2229. doi: 10.1126/science.1069424

Hao, D., Cheng, H., Yin, Z., Cui, S., Zhang, D., Wang, H., et al. (2012). Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments. *Theor. Appl. Genet.* 124, 447–458. doi: 10.1007/s00122-011-1719-0

He, J., Meng, S., Zhao, T., Xing, G., Yang, S., Li, Y., et al. (2017). An innovative procedure of genome-wide association analysis fits studies on germplasm population and plant breeding. *Theor. Appl. Genet.* 130, 2327–2343. doi: 10.1007/s00122-017-2962-9

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., et al. (2012). Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* 44, 212–216. doi: 10.1038/ng.1042

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967. doi: 10.1038/ng.695

Hwang, E. Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., et al. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15:1. doi: 10.1186/1471-2164-15-1

Hyten, D. L., Song, Q., Choi, I. Y., Yoon, M. S., Specht, J. E., Matukumalli, L. K., et al. (2008). High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theor. Appl. Genet.* 116, 945–952. doi: 10.1007/s00122-008-0726-2

Kamioka, M., Takao, S., Suzuki, T., Taki, K., Higashiyama, T., Kinoshita, T., et al. (2016). Direct repression of evening genes by CIRCADIAN CLOCK-ASSOCIATED1 in the Arabidopsis circadian clock. *Plant Cell* 28, 696–711. doi: 10.1105/tpc.15.00737

Kato, S., Fujii, K., Yumoto, S., Ishimoto, M., Shiraiwa, T., Sayama, T., et al. (2015). Seed yield and its components of indeterminate and determinate lines in recombinant inbred lines of soybean. *Breed. Sci.* 65, 154–160. doi: 10.1270/jsbbs.65.154

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Kump, K. L., Bradbury, P. J., Wisser, R. J., Buckler, E. S., Belcher, A. R., Oropeza-Rosas, M. A., et al. (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* 43, 163–168. doi: 10.1038/ng.747

Lee, S. H., Bailey, M. A., Mian, M. A., Shipe, E. R., Ashley, D. A., Parrott, W. A., et al. (1996). Identification of quantitative trait loci for plant height, lodging, and maturity in a soybean population segregating for growth habit. *Theor. Appl. Genet.* 92, 516–523. doi: 10.1007/BF00224553

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45, 43–50. doi: 10.1038/ng.2484

Li, L., Guo, N., Niu, J., Wang, Z., Cui, X., Sun, J., et al. (2016). Loci and candidate gene identification for resistance to *Phytophthora sojae* via association analysis in soybean [*Glycine max* (L.) Merr.]. *Mol. Genet. Genomics* 291, 1095–1103. doi: 10.1007/s00438-015-1164-x

Li, R., Li, J., Li, S., Qin, G., Novak, O., Pencik, A., et al. (2014). ADP1 affects plant architecture by regulating local auxin biosynthesis. *PLoS Genet.* 10:e1003954. doi: 10.1371/journal.pgen.1003954

Li, W., Sun, D., Du, Y., Chen, Q., Zhang, Z., Qiu, L., et al. (2007). Quantitative trait loci underlying the development of seed composition in soybean (*Glycine max* L. Merr.). *Genome* 50, 1067–1077. doi: 10.1139/G07-080

Li, Y. H., Zhao, S. C., Ma, J. X., Li, D., Yan, L., Li, J., et al. (2013). Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. *BMC Genomics* 14:579. doi: 10.1186/1471-2164-14-579

Li, Y. H., Zhou, G., Ma, J., Jiang, W., Jin, L. G., Zhang, Z., et al. (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* 32, 1045–1052. doi: 10.1038/nbt.2979

Liu, N., Li, M., Hu, X., Ma, Q., Mu, Y., Tan, Z., et al. (2017). Construction of high-density genetic map and QTL mapping of yield-related and two quality traits in soybean RILs population by RAD-sequencing. *BMC Genomics* 18:466. doi: 10.1186/s12864-017-3854-8

Liu, W., Kim, M. Y., Van, K., Lee, Y. H., Li, H., Liu, X., et al. (2011). QTL identification of yield-related traits and their association with flowering and maturity in soybean. *J. Crop Sci. Biotechnol.* 14, 65–70. doi: 10.1007/s12892-010-0115-7

Liu, Y., Li, Y., Reif, J. C., Mette, M. F., Liu, Z., Liu, B., et al. (2013). Identification of quantitative trait loci underlying plant height and seed weight in soybean. *Plant Genome* 6, 841–856. doi: 10.3835/plantgenome2013.03.0006

Liu, Y., Zhang, D., Ping, J., Li, S., Chen, Z., and Ma, J. (2016). Innovation of a regulatory mechanism modulating semi-determinate stem growth through artificial selection in soybean. *PLoS Genet.* 12:e1005818. doi: 10.1371/journal.pgen.1005818

Ma, H., Yanofsky, M. F., and Meyerowitz, E. M. (1991). AGL1-AGL6, an Arabidopsis gene family with similarity to floral homeotic and transcription factor genes. *Gene Dev.* 5, 484–495. doi: 10.1101/gad.5.3.484

Malik, M. F. A., Ashraf, M., Qureshi, A. S., and Ghafoor, A. (2007). Assessment of genetic variability, correlation and path analysis for yield and its components in soybean. *Pak. J. Bot.* 39, 405–413.

McCouch, S. R., Cho, Y. G., Yano, M., Paul, E., Blinstrub, M., Morishima, H., et al. (1997). Report on QTL nomenclature. *Rice Genet. Newsl.* 14, 11–13. doi: 10.1007/s10142-013-0328-1

Michael, T. P., and VanBuren, R. (2015). Progress, challenges and the future of crop genomes. *Curr. Opin. Plant Biol.* 24, 71–81. doi: 10.1016/j.pbi.2015.02.002

Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., et al. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. U.S.A.* 110, 453–458. doi: 10.1073/pnas.1215985110

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015). Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* 11:e1004969. doi: 10.1371/journal.pgen.1004969

Panthee, D. R., Pantalone, V. R., Saxton, A. M., West, D. R., and Sams, C. E. (2007). Quantitative trait loci for agronomic traits in soybean. *Plant Breed.* 126, 51–57. doi: 10.1111/j.1439-0523.2006.01305.x

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. doi: 10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Schneider, R., Rolling, W., Song, Q., Cregan, P., Dorrance, A. E., and McHale, L. K. (2016). Genome-wide association mapping of partial resistance to *Phytophthora sojae* in soybean plant introductions from the Republic of Korea. *BMC Genomics* 17:607. doi: 10.1186/s12864-016-2918-5

Segura, V., Vilhjalmsson, B. J., Platt, A., Korte, A., Seren, U., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

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, 1–9. doi: 10.18637/jss.v016.c03

Sonah, H., O'Donoughue, L., Cober, E., Rajcan, I., and Belzile, F. (2015). Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol. J.* 13, 211–221. doi: 10.1111/pbi.12249

Song, Q., Jenkins, J., Jia, G., Hyten, D. L., Pantalone, V., Jackson, S. A., et al. (2016). Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. *BMC Genomics* 17:33. doi: 10.1186/s12864-015-2344-0

Su, J., Pang, C., Wei, H., Li, L., Liang, B., Wang, C., et al. (2016). Identification of favorable SNP alleles and candidate genes for traits related to early maturity via GWAS in upland cotton. *BMC Genomics* 17:687. doi: 10.1186/s12864-016-2875-z

Sun, C., Zhang, F., Yan, X., Zhang, X., Dong, Z., Cui, D., et al. (2017). Genome-wide association study for 13 agronomic traits reveals distribution of superior alleles in bread wheat from the Yellow and Huai Valley of China. *Plant Biotechnol. J.* 15, 953–969. doi: 10.1111/pbi.12690

Sun, D., Li, W., Zhang, Z., Chen, Q., Ning, H., Qiu, L., et al. (2006). Quantitative trait loci analysis for the developmental behavior of soybean (*Glycine max* L. Merr.). *Theor. Appl. Genet.* 112, 665–673. doi: 10.1007/s00122-005-0169-y

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., et al. (2016). GAPIT Version 2: an enhanced integrated tool for genomic association and prediction. *Plant Genome* 9, 1–9. doi: 10.3835/plantgenome2015.11.0120

Teng, W., Han, Y., Du, Y., Sun, D., Zhang, Z., Qiu, L., et al. (2009). QTL analyses of seed weight during the development of soybean (*Glycine max* L. Merr.). *J. Hered.* 102, 372–380. doi: 10.1038/hdy.2008.108

Vodkin, L. O., Khanna, A., Shealy, R., Clough, S. J., Gonzalez, D. O., Philip, R., et al. (2004). Microarrays for global expression constructed with a low redundancy set of 27,500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. *BMC Genomics* 5:73. doi: 10.1186/1471-2164-5-73

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Watanabe, S., Xia, Z., Hideshima, R., Tsubokura, Y., Sato, S., Yamanaka, N., et al. (2011). A map-based cloning strategy employing a residual heterozygous line reveals that the GIGANTEA gene is involved in soybean maturity and flowering. *Genetics* 188, 395–407. doi: 10.1534/genetics.110.125062

Xin, D. W., Qiu, H. M., Shan, D. P., Shan, C. Y., Liu, C. Y., Hu, G. H., et al. (2008). Analysis of quantitative trait loci underlying the period of reproductive growth stages in soybean (*Glycine max* [L.] Merr.). *Euphytica* 162, 155–165. doi: 10.1007/s10681-008-9652-2

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., et al. (2014). Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.* 5:5087. doi: 10.1038/ncomms6087

Yao, D., Liu, Z. Z., Zhang, J., Liu, S. Y., Qu, J., Guan, S. Y., et al. (2015). Analysis of quantitative trait loci for main plant traits in soybean. *Genet. Mol. Res.* 14, 6101–6109. doi: 10.4238/2015.June.8.8

Zhang, J., Song, Q., Cregan, P. B., and Jiang, G. L. (2016). Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theor. Appl. Genet.* 129, 117–130. doi: 10.1007/s00122-015-2614-x

Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., et al. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics* 16:217. doi: 10.1186/s12864-015-1441-4

Zhang, W. K., Wang, Y. J., Luo, G. Z., Zhang, J. S., He, C. Y., Wu, X. L., et al. (2004). QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. *Theor. Appl. Genet.* 108, 1131–1139. doi: 10.1007/s00122-003-1527-2

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhou, L., Wang, S. B., Jian, J., Geng, Q. C., Wen, J., Song, Q., et al. (2015). Identification of domestication-related loci associated with flowering time and seed size in soybean with the RAD-seq genotyping method. *Sci. Rep.* 5:9350. doi: 10.1038/srep09350

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 21 August 2018
doi: 10.3389/fpls.2018.01146

Genome-Wide Association Studies for Five Forage Quality-Related Traits in Sorghum (*Sorghum bicolor* L.)

Jieqin Li [1†], Weijie Tang [2,3†], Ya-Wen Zhang [4†], Kai-Ning Chen [2], Chenchen Wang [1], Yanlong Liu [1], Qiuwen Zhan [1], Chunming Wang [3], Shi-Bo Wang [2], Shang-Qian Xie [2] and Lihua Wang [1*]*

[1] College of Agriculture, Anhui Science and Technology University, Fengyang, China, [2] College of Horticulture, Institute of Tropical Agriculture and Forestry, Hainan University, Haikou, China, [3] National Key Laboratory of Crop Genetics and Germplasm Enhancement, Jiangsu Plant Gene Engineering Research Center, Nanjing Agricultural University, Nanjing, China, [4] College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China

OPEN ACCESS

Edited by:
Zhenyu Jia,
University of California, United States

Reviewed by:
Suhong Bu,
Fujian Agriculture and Forestry University, China
Meng Li,
Nanjing Agricultural University, China

***Correspondence:**
Shang-Qian Xie
sqianxie@hainu.edu.cn
Lihua Wang
wanglihuaerr@126.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *15 May 2018*
Accepted: *18 July 2018*
Published: *21 August 2018*

Citation:
Li J, Tang W, Zhang Y-W, Chen K-N, Wang C, Liu Y, Zhan Q, Wang C, Wang S-B, Xie S-Q and Wang L (2018) Genome-Wide Association Studies for Five Forage Quality-Related Traits in Sorghum (Sorghum bicolor L.). Front. Plant Sci. 9:1146. doi: 10.3389/fpls.2018.01146

Understanding the genetic function of the forage quality-related traits, including crude protein (CP), neutral detergent fiber (NDF), acid detergent fiber (ADF), hemicellulose (HC), and cellulose (CL) contents, is essential for the identification of forage quality genes and selection of effective molecular markers in sorghum. In this study, we genotyped 245 sorghum accessions by 85,585 single-nucleotide polymorphisms (SNPs) and obtained the phenotypic data from four environments. The SNPs and phenotypic data were applied to multi-locus genome-wide association studies (GWAS) with the mrMLM software. A total of 42 SNPs were identified to be associated with the five forage quality-related traits. Moreover, three and two quantitative trait nucleotides (QTNs) were simultaneously detected among them by three and two multi-locus methods, respectively. One QTN on chromosome 5 was found to be associated simultaneously with CP, NDF, and ADF. Furthermore, 3, 2, 2, 5, and 2 candidate genes were identified to be responsible for CP, NDF, ADF, HC, and CL contents, respectively. These results provided insightful information of the forage quality-related traits and would facilitate the genetic improvement of sorghum forage quality in the future.

Keywords: sorghum, GWAS, forage quality-related traits, mrMLM, QTNs

INTRODUCTION

Sorghum (*Sorghum bicolor* L.) is a popular crop worldwide, which is used a food source, animal fodder, and raw material for alcoholic beverages and biofuels in industries (Paterson et al., 2009). Most of the important agronomic traits are genetically controlled by quantitative trait loci (QTLs) (Zou et al., 2012; Boyles et al., 2017). For example, the forage quality is an important quantitative trait. Thus, understanding their genetic mechanism is essential for identifying the candidate genes and selecting effective molecular markers in sorghum breeding.

The forage digestibility and crude protein (CP) content are the main focus for forage sorghum breeding (Murray et al., 2008). Forage digestibility is mainly determined by the cellulose (CL), hemicellulose (HC), and lignin contents (Wang H. et al., 2016), which are important components of the neutral detergent fiber (NDF). On the other hand, acid detergent fiber (ADF) is a portion of sorghum fiber and is obtained from acid detergent-treated forage. The two types of fibers, NDF and ADF, are the two vital components of forage digestibility. Recently, the forage quality traits have been studied in sorghum and some related QTLs have been identified (Murray et al., 2008; Shiringani and Friedt, 2011; Li et al., 2015). However, these identified QTLs were observed to be less sensitive due to the limitation of linkage analysis based on bi-parental mapping populations.

Compared with the linkage analysis of bi-parental mapping populations, genome-wide association studies (GWAS), which is based on linkage disequilibrium (LD) and provided sufficient genetic background information, have become a powerful alternative for the investigation of quantitative traits. There are three main strategies for GWAS. Firstly, a generalized linear model (GLM) was proposed for the genetic analysis of the quantitative traits (Price et al., 2006), but it did not effectively control the polygenic background. Secondly, a mixed linear model (MLM) was elaborated to take into account the population structure and polygenic background using the pedigree relationship or marker information (Zhang et al., 2005; Yu et al., 2006). These methods involve a large calculation burden due to the tremendous number of existing markers. Therefore, a series of rapid detection methods were finally developed, such as EMMA (Kang et al., 2008), FaST-LMM (Lippert et al., 2011), GRAMMAR-Gamma (Svishcheva et al., 2012), ECMLM (Li et al., 2014), SUPER (Wang et al., 2014), BOLT-LMM (Loh et al., 2015), and FarmCPU (Liu et al., 2016). Although the above methods have been widely adopted, the complex traits controlled by multiple QTNs could not be effectively identified. To address this issue, Zhang's group has developed a series of multi-locus GWAS methods, including mrMLM (Wang S. B. et al., 2016), FASTmrMLM (Tamba et al., 2017), FASTmrEMMA (Wen et al., 2017), ISIS EM-BLASSO (Tamba et al., 2017), pLARmEB (Zhang et al., 2017), and pKWmEB (Ren et al., 2018).

In our study, we utilized the advantageous multi-locus GWAS to investigate the sorghum forage quality-related traits. We genotyped 245 sorghum accessions by using 85,585 single-nucleotide polymorphisms (SNPs) and phenotyped them in the four environments. The data were analyzed by the multi-locus GWAS software, mrMLM.

MATERIALS AND METHODS

Plant Materials

The 245 sorghum accessions (**Table S1**) included 238 mini-core collection sorghum and 7 breeding varieties. These accessions were planted in the Fengyang campus of Anhui Science and Technology University (Fengyang, China, 32°52′ N, 177°33′ E) and Tengqiao town of Hainan Province (Tengqiao, China, 18°24′ N, 109°45′ E) in 2015 and 2016. All the experiments in the four environments used a completely randomized block design with three replicates. The aboveground parts were harvested when 70% accessions were at the heading stage. The harvested plants were dried at 75°C for three days. The plant material was then milled using a grinder and filtered using a 0.5 mm sieve.

Phenotypic Trait Evaluation and Data Analysis

Seven hundred and thirty-five sorghum samples (3 replicates) were measured for CP, CL, HC, NDF, and ADF using the traditional chemical methods, and simultaneously scanned for near-infrared (NIR) spectra with an Antaris^{TM} II FT-NIR Analyzer (Thermo, USA). A model was established using TQ Analyst software based on the NIR spectra and the results of the chemical analysis. The samples were then scanned for NIR spectra, and their CP, CL, HC, NDF, and ADF were calculated using the model. The mean of the phenotypic data and the correlation coefficients were calculated using Microsoft Excel.

DNA Extraction and RAD Sequencing

Total DNA was extracted using the DNAsecure Plant Kit (Qiagen, Cat.No. DP320). All the samples were standardized to 50 ng/μL, and 10 μL of each sample was digested with the enzymes, PstI (CTGCAG) and MspI (CCGG), at 37°C for 2 h and then at 65°C for 20 min. The digested samples were ligated with the adapters from Illumina (San Diego, CA, USA). The ligated samples were then pooled using the same volume (10 μL) for PCR-amplification in a single tube. The fragment length was analyzed using a Bioanalyzer (Agilent), and the PCR products were quantified by a Qubit3.0 fluorometer (Invitrogen). The GBS library was run on an Illumina Hiseq2500 (San Diego, CA, USA).

RAD-seq Data and Population Structure Analysis

The sequencing reads of the 245 samples were extracted from the raw data of RAD-seq and filtered by using fastx_barcode_splitter and fastq_quality_filter with parameters (-q 20 -p 80 -Q 33) of fastx_toolkit-0.0.13.2 (http://hannonlab.cshl.edu/fastx_toolkit/). The high-quality sequencing data were aligned using BWA MEM (Li and Durbin, 2009). The software—samtools, mpileup, and bcftools (Li et al., 2009), were then used to call the SNPs from the alignment files of the 245 samples; these were kept as the genotype of the sorghum population. These genotypic data were used to calculate the population structure using the fastSTRUCTURE software (Raj et al., 2014).

Genome-Wide Association Studies

The GWAS for the five forage quality-related traits (CP, CL, HC, NDF, and ADF) was performed using six methods, including mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, pKWmEB, and ISIS EM-BLASSO in the mrMLM software. The main model used in this study in the mrMLM software is as follows :$y = W\alpha + X\beta + Zu + \varepsilon$, where y is an $n \times 1$ phenotypic vector of quantitative traits, and n is the number of accessions. $W = (\omega_1, \omega_2, \cdots, \omega_c)$ is an $n \times c$ matrix of covariates (fixed effects), including a column vector of 1; the population structure or principal components can be incorporated into W. Moving on, α is a $c \times 1$ vector of fixed effects, including the intercept, and X is an $n \times 1$ vector of marker genotypes. $\beta \sim N(0, \sigma_\beta^2)$is the random effect of putative QTN. Z is an $n \times m$ design matrix, and $u \sim MVN_m(0, \sigma_g^2 K)$ is an $m \times 1$ vector of polygenic effects. K is a known $n \times n$ relatedness matrix. $\varepsilon \sim MVN(0, \sigma_e^2 I_n)$ is an $n \times 1$ vector of residual errors, and σ_e^2 is residual variance. I_n is an $n \times n$ identity matrix, and MVN denotes multivariate normal distribution. An LOD score of 3 was used as the critical threshold for significant QTNs for all the six methods.

Identification of Candidate Genes

Genes that were hit directly by the associated QTNs within a 50-kb stretch were selected to choose the candidate genes as described in Upadhyaya et al. (2016). The physical locations of the QTNs were recorded according to the assembly genome (Sorghum_bicolor_NCBIv3) and the annotation GFF file (https://www.ncbi.nlm.nih.gov/genome/108). The detailed functions of the corresponding genes were annotated by performing BLASTP search at the NCBI website, and the candidate genes were assigned to different biological processes based on the function of their homologs in other species in literature or with the help of data in the Conserved Domains Database. The selected candidate genes were associated with the main QTNs of the five traits if they made a contribution (r^2) greater than 5%.

RESULTS

Phenotype Analysis

Extensive phenotypic variations of CP, CL, HC, NDF, and ADF were observed in the 245 sorghum samples in the four environments, including two locations in 2 years (Fengyang and Tengqiao in 2015 and 2016, **Table 1**). The variation range of the five traits was 1.5 to 3.5-fold: the phenotype values of the CP content were 3.80 to 13.24% with 2.5 to 3.5-fold variation. The NDF content varied from 0.38 to 0.75 g/g with 1.5 to 1.9-fold variation, while the ADF content varied from 0.18 to 0.52 g/g with a 1.8 to 2.2-fold variation. Lastly, the HC and CL contents varied from 0.14 to 0.42 g/g and 0.12 to 0.45 g/g with 1.6 to 2.2-fold and 1.8 to 2.8-fold variations, respectively.

The correlation coefficients between a pair of traits were assessed. It was revealed that there were significant and positive correlations between ADF, NDF, CL, and HC. However, they correlated significantly but negatively with the CP phenotype, except for HC in 2015fy, 2015hn, and 2016fy and NDF in the 2016fy environments (**Table S2**). These results indicated that the four traits of ADF, NDF, HC, and CL could be genetically linked or that some genes could play pleiotropic roles in controlling these phenotypes.

RAD-Seq Genotyping And Population Structure

A total of 85,585 SNPs were identified in the genotypes of the 245 accessions using RAD-seq (**Table 2**). Chromosome 1 had the most SNP markers (11,719), while chromosome 10 had the least (5,994). The highest SNP density was observed on chromosome 3 with 1.5 SNP markers per 10 kb, whereas the lowest density was on chromosome 7 with 0.9 SNP markers per 10 kb. The average density was 1.2 markers per 10 kb. Altogether, the genotyping results were of high quality in this research. The population structure was analyzed using the fastSTRUCTURE software. The results showed that the best value for the number of sub-populations was 5 (**Figure 1**), which was selected to perform further GWAS analysis.

TABLE 1 | The statistical description for CP, CL, HC, NDF, and ADF in 245 sorghum accessions in the four environments.

Trait-environment	Mean	Range	SD	CV (%)
CP-2015fy	5.9005	3.25–11.25	0.950	16.10
CP-2015hn	8.5280	4.82–12.86	0.892	10.46
CP-2016fy	6.1888	3.80–10.41	0.640	10.34
CP-2016hn	8.7070	5.24–13.24	0.953	10.95
CL-2015fy	0.3219	0.1993–0.4504	0.0463	14.38
CL-2015hn	0.2787	0.1632–0.3914	0.0374	13.42
CL-2016fy	0.3115	0.2125–0.3851	0.0285	9.15
CL-2016hn	0.2661	0.1210–0.3453	0.0351	13.19
HC-2015fy	0.2592	0.1951–0.4165	0.0271	10.46
HC-2015hn	0.2653	0.1501–0.3298	0.0277	10.44
HC-2016fy	0.2365	0.1816–0.2992	0.0219	9.26
HC-2016hn	0.2613	0.1437–0.3185	0.0238	9.11
NDF-2015fy	0.6463	0.4327–0.7513	0.0692	10.71
NDF-2015hn	0.5999	0.3839–0.7414	0.0576	9.60
NDF-2016fy	0.6115	0.4734–0.7198	0.0434	7.10
NDF-2016hn	0.6024	0.4431–0.7282	0.0716	11.89
ADF-2015fy	0.3870	0.2376–0.5208	0.5340	13.80
ADF-2015hn	0.3341	0.2148–0.4687	0.0433	12.96
ADF-2016fy	0.3750	0.2501–0.4504	0.0382	10.19
ADF-2016hn	0.3124	0.1817–0.4096	0.0355	11.36

The unit of CP is % and that of the other four traits is g/g. Two locations: Fengyang (fy) and Tengqiao (hn); 2 years: 2015 and 2016.

GWAS Using Six Multi-Locus Methods

Six methods in the mrMLM software were used for the detection of QTNs. A total of 42 significant QTNs were detected for the five forage quality-related traits (CP, CL, HC, NDF, and ADF) across the four environments using six methods (**Table 3**). There were 5, 3, 3, 24, and 7 QTNs that were associated with CP, CL, HC, NDF, and ADF, respectively. Each trait was controlled by multiple QTNs. The 5 SNPs associated with the CP content were identified on chromosomes 2, 5, 7, and 9. The 3 SNPs associated with the CL content were present on chromosomes 2, 5, and 8, while the 3 SNPs associated with the HC content were located on chromosomes 1 and 9. The 24 SNPs associated with NDF were present on chromosomes 1, 2, 6, 7, 8, 9, and 10. Lastly, the 7 SNPs associated with the ADF content were present on chromosomes 3, 4, 5, 8, and 10. Among these QTNs, there were 4 significant QTNs, each of which was responsive for more than one trait. The three traits of ADF, CL, and NDF were associated with one QTN on chromosome 5 (RSS50197); both CL and NDF were associated with two QTNs (RSS21890 and RSS76122); ADF and CL were associated with one QTN (RSS68908) on chromosome 8.

Among the above six methods, pLARmEB was the most powerful and accountable for the identification of the 24 QTNs that mainly contributed to the NDF content trait (17 QTNs); however, their contributions were less than what were detected by other methods, except for one major QTN (RSS17673), whose contribution was greater than 5% (**Table 3**). The other methods of PKWmEB, ISIS EM-BLASSO, FASTmrMLM, mrMLM, and FASTmrEMA identified 12, 8, 8, 1, and 1 QTNs, respectively. About 43% (13 of 30) of these SNPs included the major QTNs ($r^2 > 5\%$). Besides, 3 QTNs (RSS50197, RSS21890, and RSS1510) were detected simultaneously by 3 methods, and another 5 QTNs (RSS35476, RSS83457, RSS76122, RSS22092, and RSS17673) were identified simultaneously by 2 methods. The remaining QTNs were detected by a single method, but most of them were considered as reliable because of the high thresholds at which they were detected.

Identification of Candidate Genes

The assembled sorghum genome and the annotation file from NCBI were used to annotate the genes associated with the significant QTNs. There were 14 candidate genes for five forage quality-related traits. The NDF and CP content traits were associated with five and three candidate genes, respectively. The remaining 6 genes were related to the CL, HC, and ADF content traits with each trait being associated with two genes (**Table 4**).

For the CP content trait, one candidate gene that was associated with the major QTN (RSS17673) encoded a serine/threonine-protein kinase (Sobic.002G217100), which was consistent with a previous study that concluded that serine/threonine-protein kinases are involved in signal cascade for nitrogen metabolism in plants (Champigny, 1995). Besides, two candidate genes were identified for the CP content on chromosomes 2 and 5 with one gene encoding a cysteine proteinase and the other encoding an uncharacterized protein. In addition, one main QTN associated with the CL content trait on chromosome 2 was identified, and the associated candidate gene encoded a kinesin-like protein. The kinesin protein is reported to be involved in the deposition of CL during secondary growth of fiber cells in Arabidopsis (Kong et al., 2015). Furthermore, 5 main QTNs were detected in association with the NDF content; two of these (RSS21890 and RSS50197) were co-localized with those for the CL content trait. Therefore, the same two candidate genes were identified for the NDF and CL content (Sobic.005G215300 and Sobic.002G390800). For the ADF content trait, 2 main QTNs were detected on chromosomes 3 and 10, where both candidate genes encoded a bHLH transcription factor (Sobic.003G272200 and Sobic.010G172100).

TABLE 2 | Number of SNPs on the 10 chromosomes of sorghum.

Chromosome	Length (kb)	No. of SNPs	SNP density (SNP/10 kb)
1	80884.392	11,719	1.4
2	77742.459	11,040	1.4
3	74386.277	11,181	1.5
4	68658.214	8,900	1.3
5	71854.699	7,958	1.1
6	61277.060	8,266	1.3
7	65505.356	6,086	0.9
8	62686.529	6,731	1.1
9	59416.394	7,710	1.3
10	61233.695	5,994	1

DISCUSSION

Genome-wide association study is an important alternative for mapping quantitative traits. It has been applied rapidly and extensively in plant research. These methods have been widely adopted, but only a few QTNs for each complex trait have been identified. In this study, we implemented the latest multi-locus GWAS methods available in mrMLM (Wang S. B. et al., 2016; Tamba et al., 2017; Wen et al., 2017; Zhang et al., 2017; Ren et al., 2018), which can effectively overcome the above issue and actively detect the QTNs associated with the quantitative traits. Six methods in the mrMLM software were used to identify the QTNs of five forage quality-related traits in sorghum. Of these methods, pLARmEB detected the most significant QTNs, but most of them contributed insignificantly to heritability (**Table 3**). Most of the significant QTNs associated with the NDF content, detected using pLARmEB, were observed to be in the 2015hn (13 QTNs) and 2015fy (4 QTNs) environments (**Table 3**). This result might be associated with the range of values for this phenotypic trait (**Table 1**) and the difference of environments between Hainan Tengqiao (18°24′ N, 109°45′ E) and Anhui Fengyang (32°52′ N, 177°33′ E). The range of NDF-2015hn and NDF-2015fy was 0.36 and 0.32, which was higher than that in 2016hn (0.28) and 2016fy (0.25), respectively (**Table 1**). Similar conclusions can be drawn for other traits. It means that the greater the difference in phenotype, the more favorable it is for the detection of the associated QTNs. Hainan and Anhui are

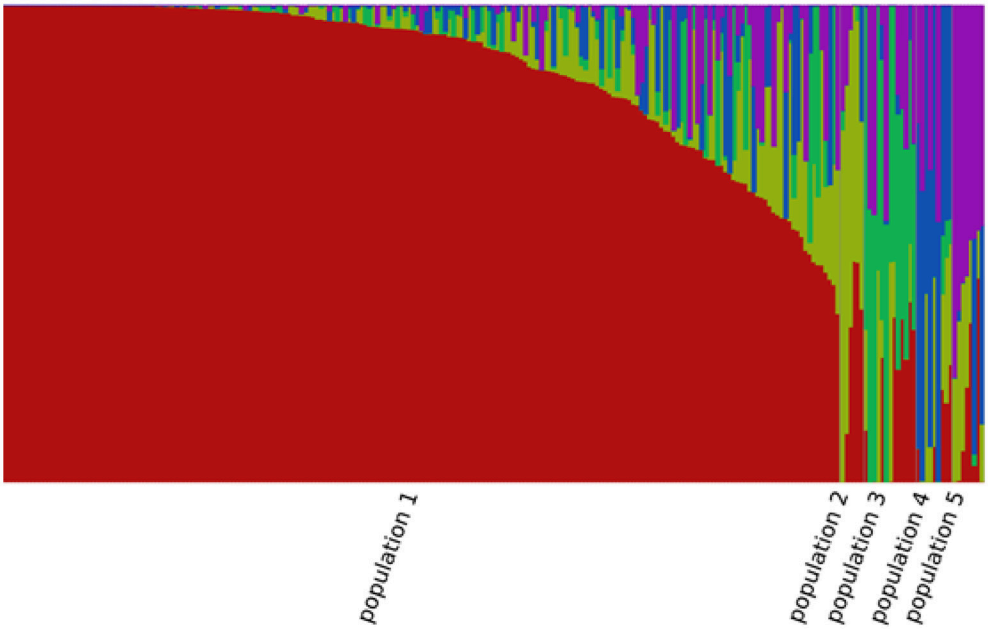


FIGURE 1 | Population structure of the 245 sorghum accessions.

located in the tropics and subtropics, respectively, where the environment is particularly different in different climatic zones. The previous study has revealed that the climatic conditions, including temperature, water availability, and soil, are important factors which affect the forage quality of sorghum (Hussin et al., 2007). In our study, the QTN RSS50197 associated with the ADF, CL, and NDF traits was uniquely detected in the same environment of 2015fy by using three GWAS methods. The above results revealed the influence of environment in QTN detection. However, the latest methods of multi-locus GWAS applied in our study are currently unable to detect the QTN-by-environment interaction. Thus, we hope that in the future new methods can be developed by the theoretical researchers.

According to the GWAS analysis, 5, 3, 3, 7, and 24 QTNs were identified for CP, CL, HC, ADF, and NDF content, respectively. Of the 5 candidate loci for the CP content, 2 were already identified in the previous studies. The locus on chromosome 9 was mapped in the same region by Murray et al. (2008) and Li et al. (2015) in sorghum as well. Of the 3 candidate loci for the CL content, 2 were identified in the same region on chromosomes 2 and 8 by Murray et al. (2008) and Shiringani and Friedt (2011). Similarly, of the 7 loci for the ADF content, 2 were mapped on chromosome 4, which was in agreement with the report of Shiringani and Friedt (2011). As for the 24 loci for the NDF content, the 2 loci on chromosome 6 and 1 loci on chromosome 8 were also identified by Shiringani and Friedt (2011). More importantly, several QTNs that were detected by the six methods in this study were novel identifications for forage quality-related traits in sorghum.

The QTLs for the NDF or ADF content co-localized with those for the CL or HC content, which has been reported previously in sorghum. Cardinal et al. (2003) reported colocalization of QTLs that are associated with the cell wall components, such as lignin, NDF, and ADF in stalks of maize. Murray et al. (2008) and Shiringani and Friedt (2011) also found colocalization of QTLs associated with the CL, HC, NDF, and ADF content traits in sorghum by QTL mapping. In this study, we detected 4 co-localized QTNs: 1 for three traits and 3 for two traits. All of these QTNs were associated with NDF or ADF and with CL or HC. NDF is mainly composed of CL, HC, and lignin, while ADF is composed of CL and lignin. The difference between NDF and ADF is whether they have HC as a component or not. Furthermore, we found that NDF and ADF significantly correlated with CL or HC. It is reasonable that these QTNs were co-localized.

Both NDF and ADF include CL and lignin. There are a series of reports about the biosynthesis and signaling pathways of CL and lignin in plants (Kim et al., 2013; McNamara et al., 2015; Yoon et al., 2015; Chezem and Clay, 2016). In this study, we identified 5 and 2 candidate genes for the NDF and ADF content traits, respectively. Of these candidate genes, 1 gene (Sobic.001G378300) encoded a sucrose synthase, which is an integral component of the CL synthesis mechanism. Gerber et al. (2014) reported that deficient sucrose synthase activity in developing wood does not specifically affect the CL biosynthesis but causes an overall decrease in the cell wall polymers. Furthermore, Poovaiah et al. (2014) reported that the lignin content increases in all the transgenic switchgrass lines, where sucrose synthase (PvSUS1) was overexpressed.

Lignin, CL, and HC are the main components of secondary cell walls (Zhong et al., 2011). Secondary cell wall biosynthesis

TABLE 3 | QTNs for CP, CL, HC, NDF, and ADF in the four environments using six multi-locus GWAS methods.

Trait	QTN	Chr	Pos (bp)	Environment	GWAS			
					Method	**Effect**	**LOD**	**r^2 (%)**
CP	RSS17673	2	60877961	2016hn	pKWmEB pLARmEB	−0.7575 −0.0014	5.11 5.11	10.89 2.87E-05
	RSS22092	2	74893615	2015fy	pKWmEB pLARmEB	−0.9453 −0.0014	4.17 3.31	9.37 1.28E-05
	RSS48493	5	65128246	2016hn	FASTmrEMMA	3.0389	4.56	5.82
	RSS62628	7	56446694	2016hn	EM_BLASSO	−0.8022	3.95	4.53
	RSS72060	9	564198	2015hn	pLARmEB	0.002	3.63	3.61E-05
CL	RSS21890	2	74389054	2015hn	FASTmrMLM EM_BLASSO pKWmEB	0.0259 0.026 0.026	3.99 4.49 4.75	10.58 10.62 14.85
	RSS50197	5	70141003	2015fy	FASTmrMLM pKWmEB	3.05E-05 3.45E-05	3.17 3.17	5.77E-06 6.08
	RSS68908	8	52710244	2016hn	pLARmEB	−0.0017	3.47	1.42E-02
HC	RSS1510	1	7334364	2015hn	FASTmrMLM EM_BLASSO pKWmEB	0.0273 0.0265 0.0265	5.32 5.36 5.36	7.70 7.27 10.37
	RSS3431	1	16641334	2016hn	EM_BLASSO	−2.06E-05	3.14	3.55E-06
	RSS76122	9	48549474	2016hn	FASTmrMLM pKWmEB	0.025 0.026	4.93 3.33	5.24 7.74
NDF	RSS973	1	5021420	2015fy	pLARmEB	−0.0011	9.61	9.00E-04
	RSS2859	1	12968078	2015hn	pLARmEB	0.0031	4.49	1.47E-02
	RSS3945	1	19028439	2015hn	pLARmEB	−0.001	5.26	6.00E-04
	RSS4300	1	21729035	2015fy	FASTmrMLM	−5.22E-05	3.91	5.72E-06
	RSS7796	1	66643031	2015hn	EM_BLASSO	0.0435	4.12	10.89
	RSS14238	2	9529690	2015hn	pLARmEB	0.0011	4.50	1.17E-03
	RSS21890	2	74389054	2015hn	pKWmEB	0.0419	4.80	15.72
	RSS44026	5	4665028	2015fy	FASTmrMLM	−1.00E-04	3.70	2.00E-04
	RSS50197	5	70141003	2015fy	pKWmEB	0.0428	4.58	7.07
	RSS54219	6	47408694	2015hn	pLARmEB	0.0014	3.37	1.58E-03
	RSS55031	6	49852623	2015hn	pLARmEB	−0.0013	4.35	1.24E-03
	RSS58234	6	58636709	2015fy	pLARmEB	0.002	3.80	1.89E-03
	RSS61032	7	8098766	2015hn	pLARmEB	0.0019	4.07	2.19E-03
	RSS65142	7	65417969	2015hn	pLARmEB	0.003	3.17	7.65E-03
	RSS65800	8	2362643	2015hn	pLARmEB	0.0012	3.64	1.15E-03
	RSS65801	8	2362646	2015hn	pLARmEB	0.0011	3.77	8.00E-04
	RSS66600	8	5538354	2015hn	pLARmEB	0.0017	3.28	2.12E-03
	RSS68217	8	48827265	2015hn	pLARmEB	0.0016	3.90	2.12E-03
	RSS70856	8	60604340	2015fy	pKWmEB	0.044	4.08	7.77
	RSS72128	9	886296	2015hn	pLARmEB	0.0013	3.46	5.00E-04
	RSS72803	9	2831721	2015fy	pLARmEB	0.002	3.61	4.72E-03
	RSS76122	9	48549474	2016hn	pKWmEB	0.0302	3.87	5.99
	RSS79370	9	58586399	2015fy	pLARmEB	−0.0037	3.76	1.33E-02
	RSS81889	10	12257429	2015hn	pLARmEB	−0.0012	3.15	2.21E-03
ADF	RSS29915	3	60813170	2015fy	mrMLM	0.0763	4.41	20.24
	RSS35476	4	6415156	2015hn	EM_BLASSO pKWmEB	−0.0159 −9.00E-04	3.27 3.38	2.98 4.43
	RSS40375	4	60973382	2015hn	EM_BLASSO	0.0144	3.51	3.09
	RSS50197	5	70141003	2015fy	FASTmrMLM EM_BLASSO pLARmEB	0.0292 0.0292 0.0032	4.42 4.42 3.64	3.97 3.97 4.91E-02
	RSS68908	8	52710244	2016hn	pLARmEB	−0.0018	3.44	1.41E-02
	RSS79627	10	386174	2015fy	pLARmEB	0.0047	3.10	0.28
	RSS83457	10	50561994	2015hn	FASTmrMLM pKWmEB	0.0178 0.0175	3.15 3.26	4.25 7.09

is positively regulated by NAD and MYB transcription factors (Zhong and Ye, 2014; Chezem and Clay, 2016). Moreover, studies have also identified several transcription factors (e.g., WRKY, ERF, and bHLH) that regulate the biosynthesis of secondary walls (Kim et al., 2013; Taylor-Teeples et al., 2015; Chezem and Clay, 2016). In this study, we identified a candidate gene encoding a bHLH transcription factor for CL and two bHLH genes for ADF. These transcription factors might also be involved in the regulation of CL or lignin biosynthesis. The function of the candidate genes identified in this work needs to be studied further by transformation experiments in the future.

TABLE 4 | Candidate genes for CP, CL, HC, NDF, and ADF traits.

Trait	Chr.	QTNs	Start	End	Gene	Function
CP	2	RSS17673	60921089	60923833	Sobic.002G217100	Serine/threonine-protein kinase
	2	RSS22092	74887469	74892115	Sobic.002G397001	Cysteine proteinase
	5	RSS48493	65140867	65142310	Sobic.005G171700	Uncharacterized protein
CL	2	RSS21890	74372243	74384041	Sobic.002G390800	Kinesin-like protein
	5	RSS50197	70130797	70134481	Sobic.005G215300	Laccase-15
HC	1	RSS1510	7364358	7367682	Sobic.001G095700	Transcription factor bHLH
	9	RSS76122	48547319	48550332	Sobic.009G132000	Uncharacterized protein
NDF	1	RSS7796	66642749	66650461	Sobic.001G378300	Sucrose synthase
	2	RSS21890	74372243	74384041	Sobic.002G390800	Kinesin-like protein
	5	RSS50197	70130797	70134481	Sobic.005G215300	Laccase-15
	8	RSS70856	60570342	60574129	Sobic.008G172200	Transcription factor TCP
	9	RSS76122	48547319	48550332	Sobic.009G132000	Uncharacterized protein
ADF	3	RSS29915	60842731	60844801	Sobic.003G272200	Transcription factor bHLH
	10	RSS83457	50554160	50558010	Sobic.010G172100	Transcription factor bHLH

AUTHOR CONTRIBUTIONS

JL, LW, and S-QX designed and conceived the experiments. Y-WZ, K-NC, and S-BW performed the computational analysis. WT and JL extracted the DNA and performed the experimental analysis. CCW and YL assisted with experiments in data collection and analysis. QZ and CW participated in the design and supervised the study. S-QX and JL discussed the results and interpretation of the final data. S-QX and JL drafted the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31671753, 31600667, 31760316), the key project of natural science research of Anhui provincial education department (KJ2016A177), the Key-construction Subject Plan of Anhui Province (WanJiaoMiKe[2014]28), the youth talent support program of Anhui Science and Technology University (XiaoRenFa[2015]69), Priming Scientific Research Foundation of Hainan University [KYQD(ZR)1721], and Science Foundation for The Youth Teachers of Hainan University in 2017 (hdkyxj201702).

ACKNOWLEDGMENTS

We would like to thank Hari D Upadhyaya (International Crops Research Institute for the Semi-Arid Tropics) for providing sorghum mini-core accessions and the two reviewers for their constructive comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01146/full#supplementary-material

REFERENCES

Boyles, R. E., Pfeiffer, B. K., Cooper, E. A., Rauh, B. L., Zielinski, K. J., Myers, M. T., et al. (2017). Genetic dissection of sorghum grain quality traits using diverse and segregating populations. *Theor. Appl. Genet*. 130, 697–716. doi: 10.1007/s00122-016-2844-6

Cardinal, A. J., Lee, M., and Moore, K. J. (2003). Genetic mapping and analysis of quantitative trait loci affecting fiber and lignin content in maize. *Theor. Appl. Genet*. 106, 866–874. doi: 10.1007/s00122-002-1136-5

Champigny, M. L. (1995). Integration of photosynthetic carbon and nitrogen metabolism in higher plants. *Photosynth. Res.* 46, 117–127. doi: 10.1007/BF00020422

Chezem, W. R., and Clay, N. K. (2016). Regulation of plant secondary metabolism and associated specialized cell development by MYBs and bHLHs. *Phytochemistry* 131, 26–43. doi: 10.1016/j.phytochem.2016.08.006

Gerber, L., Zhang, B., Roach, M., Rende, U., Gorzsas, A., Kumar, M., et al. (2014). Deficient sucrose synthase activity in developing wood does not specifically affect cellulose biosynthesis, but causes an overall decrease in cell wall polymers. *New Phytol.* 203, 1220–1230. doi: 10.1111/nph.12888

Hussin, A., Khan, S., Sulatani, M. I., and Mohammad, D. (2007). Locational variation in green fodder yield, dry matter yield, and forage quality of sorghum. *Pakistan J. Agric. Res.* 20, 1–2.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Kim, W. C., Ko, J. H., Kim, J. Y., Kim, J., Bae, H. J., and Han, K. H. (2013). MYB46 directly regulates the gene expression of secondary wall-associated cellulose synthases in Arabidopsis. *Plant J.* 73, 26–36. doi: 10.1111/j.1365-313x.2012.05124.x

Kong, Z., Ioki, M., Braybrook, S., Li, S., Ye, Z., Julie Lee, Y., et al. (2015). Kinesin-4 functions in vesicular transport on cortical microtubules and regulates cell wall mechanics during cell elongation in plants. *Mol. Plant* 8, 1011–1023. doi: 10.1016/j.molp.2015.01.004

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

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. doi: 10.1093/bioinformatics/btp352

Li, J. Q., Wang, L. H., Zhan, Q. W., Liu, Y. L., Zhang, Q., Li, J. F., et al. (2015). Mapping quantitative trait loci for five forage quality traits in a sorghum-sudangrass hybrid. *Genet. Mol. Res*. 14, 13266–13273. doi: 10.4238/2015.October.26.23

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y. M., Todhunter, R. J., et al. (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol*. 12:73. doi: 10.1186/s12915-014-0073-5

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8, 833–835. doi: 10.1038/nmeth.1681

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient Genome-wide association studies. *PLoS Genet*. 12:e1005767. doi: 10.1371/journal.pgen.1005767

Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet*. 47, 284–290. doi: 10.1038/ng.3190

McNamara, J. T., Morgan, J. L., and Zimmer, J. (2015). A molecular description of cellulose biosynthesis. *Annu. Rev. Biochem*. 84, 895–921. doi: 10.1146/annurev-biochem-060614-033930

Murray, S. C., Rooney, W. L., Mitchell, S. E., Sharma, A., Klein, P. E., Mullet, J. E., et al. (2008). Genetic improvement of sorghum as a biofuel feedstock: II. QTL for stem and leaf structural carbohydrates. *Crop Sci*. 48, 2180–2193. doi: 10.2135/cropsci2008.01.0068

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature* 457, 551–556. doi: 10.1038/nature07723

Poovaiah, C. R., Nageswara-Rao, M., Soneji, J. R., Baxter, H. L., Stewart, C. N., et al. (2014), Altered lignin biosynthesis using biotechnology to improve lignocellulosic biofuel feedstocks. *Plant Biotechnol. J*. 12, 1163–1173. doi: 10.1111/pbi.12225

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. doi: 10.1038/ng1847

Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE: variational Inference of population structure in large SNP data sets. *Genetics* 197, 573–589. doi: 10.1534/genetics.114.164350

Ren, W. L., Wen, Y. J., Dunwell, J. M., Zhang, Y. M. (2018). pKWmEB: integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Shiringani, A. L., and Friedt, W. (2011). QTL for fibre-related traits in grain x sweet sorghum as a tool for the enhancement of sorghum as a biomass crop. *Theor. Appl. Genet*. 123, 999–1011. doi: 10.1007/s00122-011-1642-4

Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M., and Aulchenko, Y. S. (2012). Rapid variance components-based method for whole-genome association analysis. *Nat. Genet*. 44, 1166–1170. doi: 10.1038/ng.2410

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol*. 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T. W., Gaudinier, A., et al. (2015). An Arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature* 517, 571–175. doi: 10.1038/nature14099

Upadhyaya, H. D., Wang, Y. H., Sastry, D. V., Dwivedi, S. L., Prasad, P. V., Burrell, A. M., et al. (2016). Association mapping of germinability and seedling vigor in sorghum under controlled low-temperature conditions. *Genome* 59, 137–145. doi: 10.1139/gen-2015-0122

Wang, H., Li, K., Hu, X., Liu, Z., Wu, Y., Huang, C. (2016). Genome-wide association analysis of forage quality in maize mature stalk. *BMC Plant Biol*. 16:227. doi: 10.1186/s12870-016-0919-9

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., and Zhang, Z. (2014). A SUPER powerful method for genome wide association study. *PLoS ONE* 9:e107684. doi: 10.1371/journal.pone.0107684

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep*. 6:19444. doi: 10.1038/srep19444

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinformatics* 19, 700–712. doi: 10.1093/bib/bbx028

Yoon, J., Choi, H., and An, G. (2015). Roles of lignin biosynthesis and regulatory genes in plant development. *J. Integr. Plant Biol*. 57, 902–912. doi: 10.1111/jipb.12422

Yu, J., Pressoir, G., Briggs, W. H., Vroh, B. I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet*. 38, 203–208. doi: 10.1038/ng1702

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y. M., Mao, Y., Xie, C., Smith, H., Luo, L., Xu, S. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics* 169, 2267–2275. doi: 10.1534/genetics.104.033217

Zhong, R., Lee, C., McCarthy, R. L., Reeves, C. K., Jones, E. G., and Ye, Z. H. (2011). Transcriptional activation of secondary wall biosynthesis by rice and maize NAC and MYB transcription factors. *Plant Cell Physiol*. 52, 1856–1871. doi: 10.1093/pcp/pcr123

Zhong, R., and Ye, Z. H. (2014). Complexity of the transcriptional network controlling secondary wall biosynthesis. *Plant Sci*. 229, 193–207. doi: 10.1016/j.plantsci.2014.09.009

Zou, G., Zhai, G., Feng, Q., Yan, S., Wang, A., Zhao, Q., et al. (2012). Identification of QTLs for eight agronomically important traits using an ultra-high-density map based on SNPs generated from high-throughput sequencing in sorghum under contrasting photoperiods. *J. Exp. Bot*. 63, 5451–5462. doi: 10.1093/jxb/ers205

Conflict of Interest Statement: The reviewer ML declared a shared affiliation, though no other collaboration, with several of the authors WT, CW to the handling Editor.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 28 August 2018
doi: 10.3389/fpls.2018.01226

Genome-Wide Association Studies of Photosynthetic Traits Related to Phosphorus Efficiency in Soybean

*Haiyan Lü[1†], Yuming Yang[2†], Haiwang Li[3†], Qijia Liu[1], Jianjun Zhang[1], Junyi Yin[1], Shanshan Chu[1], Xiangqian Zhang[1], Kaiye Yu[1], Lingling Lv[1], Xi Chen[1] and Dan Zhang[1]**

[1] Collaborative Innovation Center of Henan Grain Crops, Henan Agricultural University, Zhengzhou, China, [2] National Center for Soybean Improvement, National Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China, [3] College of Food Science and Technology, Henan University of Technology, Zhengzhou, China

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Jin Zhang,
Nanjing Agricultural University, China
Xingwang Yu,
North Carolina State University,
United States

Correspondence:
Dan Zhang
zhangd@henau.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *30 April 2018*
Accepted: *31 July 2018*
Published: *28 August 2018*

Citation:
Lü H, Yang Y, Li H, Liu Q, Zhang J, Yin J, Chu S, Zhang X, Yu K, Lv L, Chen X and Zhang D (2018) Genome-Wide Association Studies of Photosynthetic Traits Related to Phosphorus Efficiency in Soybean. Front. Plant Sci. 9:1226. doi: 10.3389/fpls.2018.01226

Photosynthesis is the basis of plant growth and development, and is seriously affected by low phosphorus (P) stress. However, few studies have reported for the genetic foundation of photosynthetic response to low P stress in soybean. To address this issue, 219 soybean accessions were genotyped by 292,035 high-quality single nucleotide polymorphisms (SNPs) and phenotyped under normal and low P conditions in 2015 and 2016. These datasets were used to identify quantitative trait nucleotides (QTNs) for photosynthesis-related traits using mrMLM, ISIS EM-BLASSO, pLARmEB, FASTmrMLM, FASTmrEMMA, and pKWmEB methods. As a result, 159 QTNs within 31 genomic regions were found to be associated with four photosynthesis-related traits under different P stress conditions. Among the 31 associated regions, five (*q7-2*, *q8-1*, *q9*, *q13-1,* and *q20-2*) were detected commonly under both normal and low P conditions, indicating the insensitivity of these candidate genes to low P stress; five were detected only under normal P condition, indicating the sensitivity of these candidate genes to low P stress; six were detected only under low P condition, indicating the tolerantness of these candidate genes to low P stress; 20 were reported in previous studies. Around the 159 QTNs, 52 candidate genes were mined. These results provide the important information for marker-assisted breeding in soybean and further reveal the basis for the application of P tolerance to photosynthetic capacity.

Keywords: soybean, photosynthesis-related traits, phosphorus efficiency, multi-locus GWAS, QTNs, candidate gene, mrMLM

INTRODUCTION

Phosphorus (P) is one of the main factors for plant growth because of its influence on cellular phosphorylation events and the synthesis of DNA and RNA (Johnston et al., 2000; Khan et al., 2009; Zhang et al., 2014b; Li et al., 2016). Nevertheless, the availability of P in soil is limited owing to the formation of organic P complexes and the fixation of P by aluminum and ferrum oxides (Vance et al., 2003; Wang et al., 2010). In the past decade, enormous efforts have been made in the dissection of the genetic mechanisms for soybean P efficiency by evaluating factors such as P concentration, root architecture (Ao et al., 2010), biomass (Li et al., 2005), and phosphatase activity (Zhang et al., 2009). Although a series of quantitative trait loci (QTLs) across all 20 chromosomes on the genome have been found to be associated with P efficiency in soybean (SoyBase, https://soybase.org), QTLs underlying photosynthetic response to low P stress have rarely been studied.

Plant productivity relies on photosynthesis, which is sensitive to low P stress (Veneklaas et al., 2012). A number of QTLs associated with photosynthesis-related traits have been detected (Yin et al., 2010a,b; Hao et al., 2012). However, the situation under the low P stress has not been considered. Recently, linkage mapping studies showed a significant genetic relationship between P efficiency and photosynthesis-related traits, such as net photosynthetic rate and transpiration rate (Li et al., 2016). In soybean, however, both the P efficiency and photosynthesis-related traits are complex quantitative traits controlled by polygenes, and most of them are genotype-specific and environment-sensitive. So far, no pleiotropic QTL for the two traits have been reported, mainly because of the relatively low mapping resolution and smaller allele effect sizes.

More recently, genome-wide association study (GWAS) has a great advantage in the dissection of genetic basis of complex traits over linkage analysis: GWAS leverages the greater number of historical recombination events, a greater number of alleles, and broader genetic variation (Yu and Buckler, 2006). Up to now this approach was widely used in multiple crops, for instance, in rice (Huang et al., 2012), soybean (Zhang et al., 2014a,b), maize (Mao et al., 2015; Wang et al., 2016c), and *Arabidopsis thaliana* (van Rooijen et al., 2017).

The most popular method for GWAS is mixed linear model (MLM) method (Zhang et al., 2005; Yu and Buckler, 2006). In the past decade, many MLM-based methods have been proposed to improve computational efficiency, such as CMLM (Zhang et al., 2010) and ECMLM (Li et al., 2014). However, these models are one-dimensional genome scan, which need the correction for multiple tests. The typical Bonferroni correction is often too conservative to identify many important loci with small effects. To address this problem, Wang et al. (2016b) proposed a multi-locus random-SNP-effect mixed linear model (mrMLM) method without Bonferroni correction. And then, a series of multi-locus GWAS methods have been proposed, such as ISIS EM-BLASSO (Tamba et al., 2017), pLARmEB (Zhang et al., 2017), FASTmrEMMA (Wen et al., 2018), FASTmrMLM (Tamba and Zhang, 2018), and pKWmEB (Ren et al., 2018). These methods not only improve the power and accuracy of GWAS but also identify the small-effect quantitative trait nucleotides (QTNs).

To reveal the genetic basis of photosynthetic response to low P stress in soybean, in this study, four photosynthesis-related traits under two P levels were measured for seedling plants in hydroponics across two environments, 219 soybean accessions were genotyped by 292,035 high-quality SNPs from NJAU 355 K Soy SNP array described by Wang et al. (2016a), and the two datasets were used to conduct GWAS for the above four traits. Because of the relatively smaller allelic effects, multi-locus GWAS methods as mentioned above, rather than common GWAS methods based on single marker analysis with a fixed-SNP-effect MLM, were adopted in this study.

Our objectives were: (i) to estimate the genetic variance and heritability of four photosynthesis-related traits under different P conditions; (ii) to investigate the correlations among the four traits under different P levels; (iii) to detect QTNs associated with the above four traits; and (iv) to predict their candidate genes.

MATERIALS AND METHODS

Plant Materials and Hydroponics Experiments

The population for GWAS was comprised of 219 soybean accessions (including 195 landraces and 24 elite varieties) derived from 26 provinces within six ecological regions in China (latitude 53 to 24°N and longitude 134 to 97°E; Wang and Gai, 2002). The 219 soybean accessions were grown hydroponically and measured by two independent experiments in 2015 and 2016 (E1 and E2). Hydroponics experiments and phenotyping were conducted as previously described by Li et al. (2016). The controlled conditions of hydroponics was 28/20°C day/night temperature and 10 h light/14 h dark photoperiod in artificial climate chambers. The surfaces of the seeds were sterilized with chlorine, and then, the seeds were sprouted in sterile vermiculite. Next, regular soybean seedlings, whose cotyledons were expanded completely, were selected. Then, the selected seedlings were moved into modified one-half Hoagland's nutrient solution supplemented with 500 μM P (normal P, KH2PO4) for 3 days. Finally, one half of the seedlings were transferred into modified one-half Hoagland's nutrient solution supplemented with 5 μM P (low P) for 14 days, and the other half remained in the normal P condition as controls.

The photosynthesis-related traits assessed were net photosynthetic rate (Pn, μmol·m^2·s^{-1}), transpiration rate (Tr, g·m^2·h^{-1}), stomatal conductance (Co, mmol·m^{-2}· s^{-1}), and intercellular carbon dioxide concentration (Ci, μL·L^{-1}) under different P conditions (normal P, low P, and the ratio of low/normal P were abbreviated as NP, LP, and L/NP, respectively) in 2015 (E1) and 2016 (E2). A LI-6400 portable photosynthesis system was used to measure the above four traits (Li Cor Inc., Lincoln, NE, USA). The phenotyping used the upper third leaf of three plants, and three replicates were measured per plant. All the traits were measured at the second trifoliolate stage. A total of 12 characteristics were analyzed in this paper: PnNP, PnLP, PnL/NP represent the net photosynthetic rates under normal P, low P, and the ratios of low/normal P, respectively; and TrNP, TrLP, TrL/NP, CoNP, CoLP, CoL/NP, CiNP, CiLP, CiL/NP represent the transpiration rates, stomatal conductance and intercellular carbon dioxide concentrations under normal P, low P, and the ratios of low/normal P, respectively.

Genotyping and Statistical Analysis of the Phenotypes

Two hundred and nineteen soybean accessions were genotyped by 292,035 SNPs derived from NJAU 355 K Soy SNP array described by Wang et al. (2016a). In other words, there was one SNP per 3.3 kb along the 20 soybean chromosomes. In the present study, SNPs with minor allele frequency (MAF) $<$ 0.05 were deleted. Based on this rule, a total of 201,994 SNPs were used for the GWAS.

The ANOVA of the phenotypic data was carried out using the PROC GLM of SAS version 9.2 (SAS Institute, Cary, NC, USA). Genotype and environment were treated as fixed and random, respectively. The broad-sense heritability (h^2) was calculated as: $h^2 = \sigma_g^2/\sigma_p^2$, where σ_g^2 is the genotypic variance, σ_p^2 is the

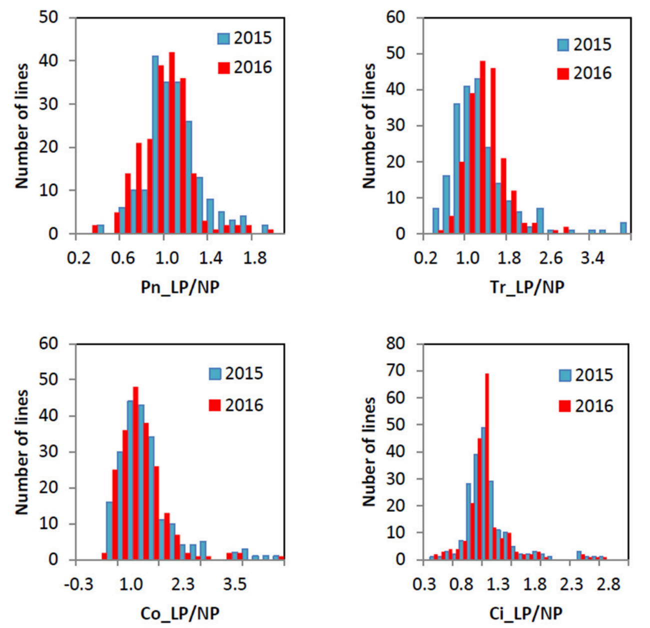


FIGURE 1 | Histogram of the frequency distributions for the four photosynthesis-related traits in soybean under L/NP condition in 2015 and 2016.

phenotypic variance. SPSS Statistics 19.0 (SPSS, Inc., Chicago, IL, USA) was used to analyze the correlation coefficients among the four photosynthesis-related traits under different P conditions in the soybean.

Genome-Wide Association Studies and Prediction of Candidate Genes

Population structure of the 219 soybean accessions each with 201,994 SNPs was calculated using the STRUCTURE package (Pritchard et al., 2009). The relative kinship (K matrix) between a pair of accessions was calculated using the R package mrMLM. GWAS was performed by the R package mrMLM with six multi-locus GWAS methods: mrMLM (Wang et al., 2016b), ISIS EM-BLASSO (Tamba et al., 2017), pLARmEB (Zhang et al., 2017), FASTmrEMMA (Wen et al., 2018), FASTmrMLM (Tamba and Zhang, 2018), and pKWmEB (Ren et al., 2018). In order to get more accurate candidate genes, markers that met the criterion of LOD score ≥ 5 were considered to be significantly associated with the traits.

To mine the candidate genes related to soybean photosynthesis response to low P stress, the predicted genes around significantly associated QTNs were identified based on the annotation in the soybean reference genome (Wm82.a2.v1) in Phytozome v10.3 (http://phytozome.net). Then, the genes with known function annotations underlying soybean photosynthesis-related traits under different P conditions were selected as candidate genes. In addition, we also selected the previously reported QTLs from soybase (https://soybase.org) in the associated genomic regions.

RESULTS

Phenotype for Photosynthesis-Related Traits

All the four traits under different P conditions showed approximately normal distributions (**Figure 1** and **Figure S1**). However, the four traits under the L/NP condition were far away from normal distributions, indicating the existence of major QTNs. The coefficients of variation for the four traits under different P conditions ranged from 13.99~69.22% (**Table 1**). The analysis of variance showed the significant differences for the four traits between genotypes and between environments. The last two results indicated that it is suitable for this population to conduct multi-locus GWAS.

To investigate the correlation among the four photosynthesis-related traits, simple correlations were calculated based on the average values of the two experiments (**Table 2** and **Table S2**). The results showed that Co was very significantly and positively correlated with Tr [$r = 0.886$ (NP) or 0.924 (LP)]; Ci was very significantly and positively correlated with Pn [$r = 0.394$ (NP) or 0.500 (LP)]; Pn was significantly and negatively correlated with Tr and Co ($r = -0.100$ and -0.108), respectively, under normal P condition; Tr was significantly and negatively correlated with Ci ($r = -0.167$) under normal P condition (**Table 2**).

Multi-Locus Genome-Wide Association Studies for Photosynthesis-Related Traits

A total of 201,994 SNPs were selected with MAF ≥ 0.05 from 292,053 high-quality SNPs. The selected SNPs were used to determine the number of sub-populations (*k*) using the software STRUCTURE. As a result, the *k*-value was 3. The above information along with four photosynthesis-related traits under different P conditions (NP, LP, and L/NP) in 2015 (E1) and 2016 (E2) was used to conduct multi-locus GWAS using package mrMLM. For all the traits, QTNs within approximately 5 Mb or less were viewed as caused by one common gene (Visscher et al., 1996; Öckinger et al., 2006; Swanson-Wagner et al., 2009; Wang et al., 2012). As a result, a total of 31 associated regions comprised of 159 QTNs across all the 20 soybean chromosomes, except the 2, 3, 4, 5, and 10 chromosomes, were significantly associated with the related traits at the critical LOD ≥ 5 (**Table 3** and **Figure 2**). All the 31 associated regions were identified by at least three methods. The full list of significant QTNs from the six multi-locus GWAS methods is presented in **Table S2**. Among the 159 QTNs, the numbers of QTNs detected under NP, LP, and L/NP conditions were 59, 64, and 66, respectively; while the numbers of QTNs associated with Co, Tr, Ci, and Pn were 56, 54, 35, and 31, respectively (**Table S2**).

Most of the 31 associated regions were detected under both NP and LP conditions, including the QTNs on chromosomes 1, 6, 7, 8, 9, 13, 14, 15, 18, and 20. Five of them, *q7-2*, *q8-1*, *q9*, *q13-1,* and *q20-2* were more representative than others. *q13-1* was associated with the four photosynthesis-related traits, Pn, Tr, Co, and Ci. *q8-1* was associated with Tr, Co, and Pn. *q9* was associated with Tr, Co, and Ci. Additionally, *q8-1* and *q9* were associated with Tr under the L/NP conditions. *q7-2* was associated with Pn and Ci under LP conditions, while it was also associated with Tr and

TABLE 1 | Descriptive statistical results for photosynthesis-related traits of soybean under different P conditions.

P level[a]	Trait	Year	Mean	Stdev	Skewness	Kurtosis	Minimum	Maximum	CV(%)[b]	Genotype	Environment	h^2(%)[c]
NP	Pn	2015	15.66	6.87	0.83	−0.25	3.06	34.38	43.8	**	**	81.32
		2016	19.00	7.54	0.51	−0.97	4.89	36.17	39.68			
	Tr	2015	3.11	1.61	1.58	3.62	0.85	10.67	51.82	**	**	72.33
		2016	4.79	2.51	0.46	−0.70	1.01	10.73	52.45			
	Co	2015	0.15	0.08	0.78	0.28	0.03	0.39	53.24	**	**	78.66
		2016	0.20	0.12	0.80	0.39	0.04	0.67	58.07			
	Ci	2015	391.97	138.66	1.54	3.11	131.02	972.15	35.37	**	**	74.45
		2016	301.55	43.09	0.29	2.68	131.02	444.70	14.29			
LP	Pn	2015	14.90	6.56	0.72	−0.76	4.10	31.79	44.0	**	**	76.69
		2016	16.99	8.48	0.50	−0.99	2.64	36.50	49.91			
	Tr	2015	3.14	1.84	1.95	5.08	0.80	11.78	58.76	**	**	75.00
		2016	4.53	2.73	0.84	−0.01	0.43	12.56	60.23			
	Co	2015	0.16	0.10	1.44	2.17	0.03	0.53	63.43	**	**	80.83
		2016	0.19	0.13	1.19	1.00	0.02	0.67	69.14			
	Ci	2015	410.08	139.91	1.76	3.71	130.76	954.63	34.12	**	**	69.79
		2016	312.03	43.65	0.68	0.76	213.21	457.05	13.99			
L/NP	Pn	2015	1.02	0.29	2.41	15.05	0.35	3.21	28.7	**	**	47.58
		2016	0.89	0.23	0.59	2.31	0.25	1.88	26.25			
	Tr	2015	1.16	0.64	2.31	7.30	0.26	4.39	55.4	**	**	63.33
		2016	1.03	0.49	2.19	7.55	0.14	3.59	47.26			
	Co	2015	1.29	0.89	2.49	7.47	0.27	5.75	69.2	**	**	85.14
		2016	1.09	0.72	3.09	14.28	0.09	5.75	66.19			
	Ci	2015	1.14	0.44	3.14	13.25	0.37	3.82	38.9	**	**	91.73
		2016	1.12	0.67	3.21	13.97	0.32	3.82	59.62			

***: the 0.01 level of significance.*

[a] P level: NP, LP, and L/NP represent the traits under normal P, low P, and the ratio of low and normal P, respectively.

[b] Coefficient of variation.

[c] broad-sense heritability.

TABLE 2 | Correlations between four photosynthesis-related traits under different P conditions.

LP/NP	Pn	Tr	Co	Ci
Pn	1.000	−0.100*	−0.108*	0.394**
Tr	0.016	1.000	0.886**	−0.167**
Co	0.015	0.924**	1.000	0.023
Ci	0.500**	−0.096	0.073	1.000

Pearson correlation coefficients under NP and LP conditions were listed above and below the diagonal, respectively.

** and **: the 0.05 and 0.01 levels of significance, respectively.*

Co under NP conditions. *q20-2* was associated with Pn and Co under both NP and LP conditions, while it was also associated with Tr and Ci under LP conditions. There were some regions that were uniquely associated with one trait. For example, *q17-2* was associated only with Ci, while *q1-2* and *q8-2* were associated only with Pn, and both loci contain only one QTN. These QTNs probably contribute to the genetic basis of photosynthesis and are probably not significantly influenced by low P stress.

In addition, there were several QTNs identified uniquely under NP or LP conditions and, therefore, they were considered as NP-specific or LP-specific QTNs. For example, *q8-3*, *q11-2*, *q13-2*, *q16-1*, and *q20-3* on chromosomes 8, 11, 13, 16, and 20, respectively, were detected only under NP conditions. In contrast, several LP-specific QTNs, *q1-1*, *q7-1*, *q14-3*, *q15-2*, *q16-2*, and *q19* on chromosomes 1, 7, 14, 15, 16, and 19, respectively, were detected only under LP level, indicating that the genes underlying these QTNs may be more likely to be affected by low P stress. Moreover, most of the 31 loci were detected under L/NP conditions, and the most representative QTNs were *q18-3* and *q20-1*, which were associated with all four photosynthesis-related traits under the L/NP conditions. Further research of these P condition-specific QTNs may supply more understanding to the genetic basis of P tolerance to photosynthetic capacity.

On the other hand, 13 of the 31 associated regions were repeatedly detected more than 6 times across treatments, traits or years, which are major QTNs (**Figure 2**; **Table 3**). These QTNs could be used to assess the effect of low P stress on photosynthesis in further analysis. As shown in **Table 3**, 13 QTNs (*q6*, *q7-2*, *q8-1*, *q8-3*, *q9*, *q13-1*, *q13-2*, *q14-3*, *q18-1*, *q18-2*, *q18-3*, *q20-1*, and *q20-2*) were mapped on chromosomes 6, 7, 8, 9, 13, 14, 18, and 20. In addition, comparative analyses showed that eight major QTNs (*q7-2*, *q8-1*, *q9*, *q13-2*, *q14-3*, *q18-2*, *q18-3*, and *q20-1*) were co-localized with the QTLs identified in previous

TABLE 3 | Details of loci associated with photosynthesis-related traits via multi-locus GWAS in soybean.

Region associated[a]	Chr.[b]	SNP associated[c]	Pos. (bp)[d]	No.[e]	*LOD*	r^2 (%)[f]	Position intervals (bp)	Method[g]	Trait-year-treatment[h]
q1-1	1	AX-93961332	951461	5	6.32	3.35	951461–1344144	1, 3, 5	TrLP_E1, CoLP_E1, TrL/NP_E2, CoL/NP_E1
q1-2	1	AX-93675249	50496989	1	11.37	8.83	50496989	1, 2, 3	PnLP_E2, PnNP_E2
q6	**6**	**AX-93728015**	**12548698**	**7**	**9.18**	**11.51**	**12516126–12966564**	**1, 2, 3, 4, 5**	**TrLP_E1, CoLP_E2, TrNP_E2, CoNP_E2, TrL/NP_E2, CoL/NP_E2**
q7-1	7	AX-93926489	3402895	4	7.68	11.06	1787229– 3402895	1, 3, 5	CiLP_E2, CoL/NP_E1, CiL/NP_E1
q7-2	**7**	**AX-93741303**	**9520676**	**7**	**6.55**	**4.46**	**6578651– 9520676**	**1, 2, 3, 5**	**PnLP_E2, CiLP_E2, TrNP_E2, CoNP_E2, TrL/NP_E1, CoL/NP_E1, CoL/NP_E2**
q8-1	**8**	**AX-93753054**	**11620180**	**7**	**7.82**	**6.41**	**6955757–11620180**	**1, 2, 3, 4, 5**	**PnLP_E2, TrLP_E1, CoLP_E1, PnNP_E1, CoNP_E1, CoNP_E2, TrL/NP_E1,**
q8-2	8	AX-93929582	27620272	1	6.54	3.19	27620272	1, 2, 3, 4	PnLP_E2, PnNP_E2
q8-3	**8**	**AX-93759645**	**43743823**	**6**	**9.87**	**9.64**	**41226360–46081608**	**1, 2, 4, 5, 6**	**PnNP_E1, PnNP_E2, TrNP_E2, CoNP_E2, CiNP_E2, TrL/NP_E1, CoL/NP_E1, CiL/NP_E2**
q9	**9**	**AX-94066868**	**40240035**	**6**	**12.20**	**13.77**	**40188126–42709534**	**1, 2, 3, 4, 5**	**TrLP_E2, CoLP_E2, TrNP_E1, CoNP_E2, CiNP_E1, TrL/NP_E1, TrL/NP_E2, CiNP_E2**
q11-1	11	AX-94084631	6262749	4	8.13	7.09	6262749–9894114	2, 4, 5	PnNP_E1, PnNP_E2, CoL/NP_E2, CiL/NP_E1
q11-2	11	AX-94091690	32599188	4	6.92	9.87	32540598–34020885	2, 3, 4, 5	TrNP_E1, CoNP_E1, CiNP_E2, CoL/NP_E2, CiL/NP_E2
q12	12	AX-93796430	1697221	2	8.18	7.24	613090–1697221	1, 3, 4, 5	CiNP_E2, CiL/NP_E2
q13-1	**13**	**AX-94104819**	**18590366**	**4**	**28.95**	**6.31**	**15071765–18590366**	**1, 2, 3,**	**TrLP_E1, CoLP_E1, CiLP_E1, PnNP_E1, CiNP_E1, CoNP_E1, CoL/NP_E2**
q13-2	**13**	**AX-94287210**	**31003637**	**5**	**10.38**	**1.28**	**29481274–31003637**	**1, 3, 4, 5**	**PnNP_E2, TrNP_E2, PnL/NP_E1, TrL/NP_E1, CiL/NP_E1, CiL/NP_E2**
q14-1	14	AX-93820315	1059966	5	9.24	7.03	1059966–2165525	1, 2, 4, 5, 6	PnLP_E2, PnNP_E2, TrL/NP_E2, CoL/NP_E2, CiL/NP_E2
q14-2	14	AX-94288085	4897088	4	21.44	7.25	4897088–7755174	1, 3, 4, 5	CoLP_E2, CiLP_E2, TrNP_E2, CoNP_E2, TrL/NP_E2
q14-3	**14**	**AX-94129538**	**47514182**	**7**	**7.27**	**5.19**	**46008634–47723841**	**2, 3, 4, 5**	**PnLP_E1, CoLP_E1, PnL/NP_E1, PnL/NP_E2, TrL/NP_E1, CiL/NP_E2**
q15-1	15	AX-94134672	12611721	3	6.26	2.32	12227172–12611721	1, 3, 4	CoLP_E2, CoNP_E1, CoNP_E2, PnL/NP_E1
q15-2	15	AX-93841986	35259050	2	7.39	5.01	33719705–35259050	1, 2, 4, 5	PnLP_E2, TrLP_E2, CoLP_E2
q16-1	16	AX-93946322	179538	5	6.73	5.10	147918–2435036	1, 2, 4, 5	TrNP_E2, CoNP_E1, CoNP_E2, CiNP_E2
q16-2	16	AX-93947209	35210924	7	10.99	7.30	32784699–37003959	1, 2, 4, 5	TrLP_E1, TrLP_E2, CoLP_E1, CoLP_E2, CoL/NP_E2
q17-1	17	AX-94155900	5718961	4	11.38	9.51	5718961–5761052	2, 3, 5	CoLP_E2, PnNP_E1, TrNP_E2, CiL/NP_E1
q17-2	17	AX-94159333	15188572	2	7.19	8.91	15188572–15492731	2, 3, 4, 5	CiLP_E1, CiNP_E1
q17-3	17	AX-93866265	37701731	5	8.32	9.24	37615577–39224452	1, 2, 3, 4	TrLP_E1, PnNP_E1, TrNP_E2, TrL/NP_E2, CoL/NP_E2
q18-1	**18**	**AX-94166205**	**2982489**	**7**	**15.10**	**12.61**	**674420–3925002**	**2, 3, 4, 5**	**PnLP_E1, TrLP_E1, CiLP_E1, CiLP_E2, CiNP_E1, TrL/NP_E2**
q18-2	**18**	**AX-93871255**	**9956907**	**6**	**9.86**	**15.88**	**5443584-9956907**	**1, 2, 3, 4, 5**	**PnLP_E1, TrLP_E1, TrLP_E2, CoLP_E2, PnNP_E2, CoL/NP_E1**
q18-3	**18**	**AX-93883305**	**53037663**	**15**	**13.24**	**10.91**	**50663235- 55727445**	**1, 2, 3, 5, 6**	**PnLP_E2, TrLP_E2, CoLP_E1, PnNP_E2, CoNP_E1, CoNP_E2, PnL/NP_E1, TrL/NP_E1, CoL/NP_E1, CiL/NP_E1, CiL/NP_E2**
q19	19	AX-93886366	2991135	4	8.17	9.02	2991135- 3374702	1, 4, 5, 6	TrLP_E1, TrLP_E2, CoLP_E1, CoLP_E2
q20-1	**20**	**AX-94197533**	**1364621**	**6**	**8.20**	**6.22**	**455608-3145850**	**2, 3, 4, 5**	**CoLP_E1, PnL/NP_E1, TrL/NP_E1, TrL/NP_E2, CoL/NP_E1, CiL/NP_E2**
q20-2	**20**	**AX-93956837**	**35089898**	**10**	**11.48**	**16.82**	**35089898-39987379**	**1, 2, 3, 5, 6**	**PnLP_E1, TrLP_E1, CoLP_E1, CiLP_E1, PnNP_E1, CoNP_E1, PnL/NP_E2, TrL/NP_E1, TrL/NP_E2, CoL/NP_E2**
q20-3	20	AX-93910666	45122261	4	8.91	7.38	44049520- 46908248	1, 2, 3, 4	TrNP_E1, CoNP_E1, CiNP_E2, CoL/NP_E2

[a] *QTN named by chromosome.*

[b] *Chromosome.*

[c] *QTNs that were significantly associated with the trait.*

[d] *QTN position (bp) on soybean genome assembly Glycine max Wm82.a1.v1.1 (www.phytozome.net).*

[e] *The number of significant QTNs detected in the region.*

[f] *The proportion of phenotypic variance explained by each QTN.*

[g] *The mrMLM, pKWmEB, pLARmEB, FASTmrMLM, ISIS EM-BLASSO, and FASTmrEMMA were marked from 1 to 6, respectively.*

[h] *The trait-year-treatment combination of QTN, for example, Pn, net photosynthetic rate; Tr, transpiration rate; Ci, intercellular carbon dioxide concentration; and Co, stomatal conductance; followed by the treatments and environments. NP, LP and L/NP denote normal-P, low-P and the ratio of low/normal P condition, respectively. E1 and E2 denote 2015 and 2016, respectively.*

The bold values denote major QTL.

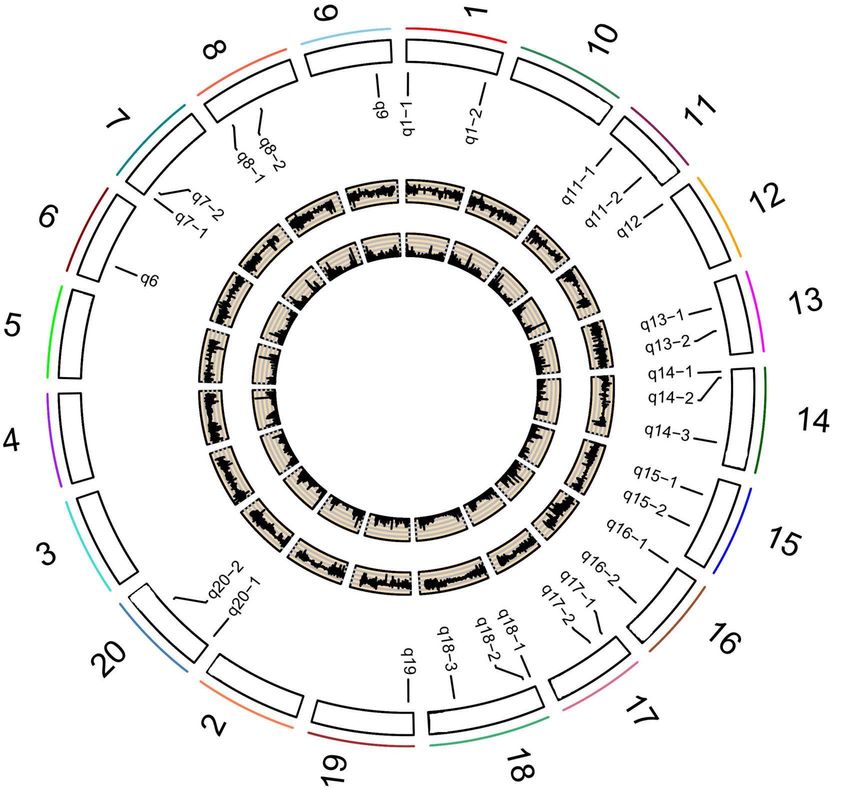


FIGURE 2 | Soybean chromosomes and QTLs for the studied traits under different P conditions. The outside/inside wheat-colored circle indicates the LOD/ r^2 value curve for the studied traits across environments. The outermost circle indicates the 20 soybean chromosomes; QTLs for the studied traits under different P conditions.

TABLE 4 | Summary of six multi-locus GWAS analysis for the four traits.

Case	mrMLM	pKWmEB	pLARmEB	FASTmrMLM	ISIS EM-BLASSO	FASTmrEMMA
No. QTN[a]	46	55	52	44	43	9
No. region[b]	24	24	24	23	25	5
LOD	5.02~11.48	5.11~13.24	5.12~28.95	5.03~10.99	5.08~11.38	5.02~8.26
r^2 (%)[c]	3.35~16.82	3.49~12.27	0.01~13.77	0.62~9.82	2.11~16.04	5.06~9.14

[a] *The number of detected QTNs.*

[b] *The number of associated regions.*

[c] *The proportion of phenotypic variance explained by each QTN (%).*

reports (Zhang et al., 2009, 2016; Li et al., 2016), including the QTL harboring the P efficiency-related gene, *GmACP1* (Zhang et al., 2014b). These eight QTNs most likely play important roles for P efficiency in soybean. For instance, the major QTN *q8-1*, where the acid phosphatase encoding gene *GmACP1* is located and underlying variation in Pn, Tr, and Co, was stably detected across traits and environments. The co-localization of *GmACP1* with *q8-1* demonstrates the high accuracy of the GWAS results in this study.

Prediction and Preliminary Validation of Candidate Genes

Although it is not easy to compare the results in different studies with different genetic maps, we determined whether the 31 associated regions in the present study were situated at or near the same position as previously identified QTLs by comparing the chromosomal locations of these QTLs (https://soybase.org). Twenty of the 31 regions were reported in previous studies (**Table S3**), and some of them were associated with leaflet-related traits (Yamanaka et al., 2001; Jun et al., 2014; Shim et al., 2015), such as *q7-2*, *q14-2*, *q18-2*, and *q20-2*, which could be closely related with photosynthesis.

To identify candidate genes affecting each trait, we re-investigated the 159 QTNs detected in our study based on the annotation of the soybean reference genome W82.a2.v1. As a result, 52 annotated genes were found and listed in **Table S4**. Most of them were previously associated with P efficiency. For example, the gene cluster within the *q8-1* region on chromosome

8 (*Glyma.08G114800*, *Glyma.08G115400*, *Glyma.08G123200*, *Glyma.08G129200*, *Glyma.08G150800*) was near the key P efficiency-related gene, *GmACP1* (Zhang et al., 2014b). Another gene cluster within the *q13-2* region on chromosome 13 (*Glyma.13G181600*, *Glyma.13G192100*, *Glyma.13G194500*, and *Glyma.13G196600*) was near the protein kinase gene, *Glyma.13G161900* (Zhang et al., 2016). In particular, the gene *Glyma.13G196600*, encoding NADPH: quinine oxidoreductase, might participate in the metabolic processes involving phosphate and photosynthesis.

The major gene cluster of *q18-1* on chromosome 18 has three annotated genes in the region encoding DNA polymerase alpha 2 (*Glyma.18G009300*), anaphase-promoting complex/cyclosome 2 (*Glyma.18G036900*), and ALWAYS EARLY4 (*Glyma.18G040400*), which is near a rubisco activase gene *Glyma.18G036400* (Li et al., 2016); these could have significant effects on the regulation of photosynthetic capacity in the soybean. In addition, there was also a single annotated gene that had been reported previously. *Glyma.12G023100*, within the *q12* region on chromosome 12, encodes a Transmembrane amino acid transporter family protein, which is physically close to ribulose-bisphosphate carboxylases gene *Glyma.12G061600* (Li et al., 2016).

DISCUSSION

Comparison of Six Multi-Locus GWAS Methods

With the development of advanced genomic sequencing technologies, GWAS has become a widely used method and is popular for the genetic dissection of variation in complex traits. While most complex traits are dominated by major genes plus polygenes, the common GWAS using a one-dimensional scanning model might not be able to detect associations with the variation of polygenes because of the limitation of the model. A better alternative is the multi-locus model GWAS (Wang et al., 2016b). In the present study, six multi-locus GWAS methods were used, and a total of 159 QTNs were found to be associated with the four photosynthesis-related traits under different P conditions (**Table 4** and **Table S2**). Furthermore, 41 of the 159 QTNs were detected by at least two methods and all the 31 associated regions were detected by at least three methods. In comparing the six multi-locus GWAS methods, we found that only nine QTNs had been detected by FASTmrEMMA, while more than 40 QTNs were detected by each of the other five methods.

The maximum LOD scores were more than 10 except for those from FASTmrEMMA, which was 8.26, smaller than the other five methods. The maximum LOD score of pLARmEB (28.95) was significantly larger than the LOD scores from the other methods. Moreover, the minimum r^2 (%) was 0.01 from pLARmEB, which may be meaningless. Meanwhile, the minimum r^2 (%) from FASTmrEMMA was 5.06, which was significantly higher than those from the other methods, meaning that FASTmrEMMA might detect major QTNs with the larger effects. This outcome explains why there were fewer associated QTNs from FASTmrEMMA than from the other five methods.

Novel QTNs and Potential Candidate Genes of Interest

Among the 13 major QTNs, five (*q6*, *q8-3*, *q13-1*, *q18-1*, and *q20-1*), which have not been reported in previous studies, were considered as novel QTNs for photosynthesis response to low P stress. It is worth noting that *q20-1* was associated with all four photosynthesis-related traits under the L/NP conditions. Thus *q20-1* might represent another important novel QTN related to Photosynthesis. In addition, two annotated genes within the *q20-1* region encoding a Mitochondrial substrate carrier family protein (*Glyma.20G004600*) and a Cyclophilin-like peptidyl-prolyl cis-trans isomerase family protein (*Glyma.20G005600*) were found in our study. If possible, more research on these genes might reveal their genetic mechanisms in future.

Another major QTN, *q13-1*, was associated with the four photosynthesis-related traits under both NP and LP conditions. This QTN was also reported previously for seed methionine content and seed cysteine content (Panthee et al., 2006a,b). Furthermore, one annotated gene *Glyma.13G053400*, within the *q13-1* region on chromosome 13, which encodes a Mitochondrial substrate carrier family protein, was listed in **Table S4**. Thus, this QTN could be a promising candidate locus for further study of low P stress on photosynthetic efficiency.

Some annotated genes weren't reported previously to be associated with phosphate and photosynthetic metabolic processes. For instance, one gene *Glyma.14G029100*, within the *q14-1* region on chromosome 14, encodes sucrose phosphate synthase 3F. Two annotated genes encoding a phosphate transporter (*Glyma.11G087800*) and a phospholipase (*Glyma.11G230100*) might be involved in the metabolic process of phosphate and photosynthesis.

Based on 292,035 high-quality SNPs in 219 soybean accessions, 159 QTNs within 31 regions were identified to be associated with four photosynthesis-related traits under different P conditions. Importantly, genetic improvement simultaneously for phosphorus efficiency and photosynthesis in soybean might be carried out by selecting for a single large-effect QTN. The associated regions and candidate genes detected in the present study could be further tested for marker-assisted breeding of soybean varieties for the application of P tolerance to photosynthetic capacity.

AUTHOR CONTRIBUTIONS

DZ and HLü conceived and designed the experiments. YY, HLi, SC, XZ, KY, LL, XC, and WW performed the experiments. HLü, JZ, JY, and QL performed data analyses. HLü and DZ wrote the manuscript. All authors have read and approved the final version of the manuscript.

ACKNOWLEDGMENTS

This research was supported by the China Postdoctoral Science Foundation (2017T100532, 2015M580630), the National Natural Science Foundation of China (31701309), the Science and technology Innovation Talents Projects of the

Education Department of Henan Province (15HASTIT034), and Key scientific research Program of the Higher Education Institutions of Henan Province (17A110024, 18A110020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01226/full#supplementary-material

Figure S1 | Histogram of the frequency distributions for the four photosynthesis-related traits of soybean under NP and LP conditions in 2015 and 2016.

Table S1 | Correlation coefficients among the four photosynthesis-related traits under different P conditions.

Table S2 | The detailed list of QTNs for photosynthesis-related traits under different P conditions across two environments using six multi-locus GWAS methods.

Table S3 | The overlapped QTLs between present and previous studies.

Table S4 | Predicted candidate genes associated with photosynthesis-related traits under different *P* conditions.

REFERENCES

Ao, J., Fu, J., Tian, J., Yan, X., Liao, H., Ma, J., et al. (2010). Genetic variability for root morph-architecture traits and root growth dynamics as related to phosphorus efficiency in soybean. *Funct. Plant Biol.* 37, 304–312. doi: 10.1071/FP09215

Hao, D., Chao, M., Yin, Z., and Yu, D. (2012). Genome-wide association analysis detecting significant single nucleotide polymorphisms for chlorophyll and chlorophyll fluorescence parameters in soybean (*Glycine max*) landraces. *Euphytica* 186, 919–931. doi: 10.1007/s10681-012-0697-x

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., et al. (2012). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* 44, 32–39. doi: 10.1038/ng.1018

Johnston, A., Steen, I., and Association, E. F. M. (2000). *Understanding phosphorus and its use in agriculture*. Brussels: European Fertilizer Manufacturers Association.

Jun, T., Freewalt, K., Michel, A. P., and Mian, R. (2014). Identification of novel QTL for leaf traits in soybean. *Plant Breeding* 133, 61–66. doi: 10.1111/pbr.12107

Khan, A. A., Jilani, G., Akhtar, M. S., Naqvi, S. M. S., and Rasheed, M. (2009). Phosphorus solubilizing bacteria: occurrence, mechanisms and their role in crop production. *J. Agric. Biol. Sci.* 1, 48–58.

Li, H., Yang, Y., Zhang, H., Chu, S., Zhang, X., Yin, D., et al. (2016). A genetic relationship between phosphorus efficiency and photosynthetic traits in soybean as revealed by QTL analysis using a high-density genetic map. *Front. Plant Sci.* 7:924. doi: 10.3389/fpls.2016.00924

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y. M., Todhunter, R. J., et al. (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biol.* 12:73. doi: 10.1186/s12915-014-0073-5

Li, Y., Wang, Y., Tong, Y., Gao, J., Zhang, J., and Chen, S. (2005). QTL mapping of phosphorus deficiency tolerance in soybean (*Glycine max* L. Merr.). *Euphytica* 142, 137–142. doi: 10.1007/s10681-005-1192-4

Mao, H., Wang, H., Liu, S., Li, Z., Yang, X., Yan, J., et al. (2015). A transposable element in a *nac* gene is associated with drought tolerance in maize seedlings. *Nat. Commun.* 6:8326. doi: 10.1038/ncomms9326

Öckinger, J., Serrano-Fernández, P., Möller, S., Ibrahim, S. M., Olsson, T., and Jagodic, M. (2006). Definition of a 1.06-Mb region linked to neuroinflammation in humans, rats and mice. *Genetics* 173, 1539–1545. doi: 10.1534/genetics.106.057406

Panthee, D. R., Pantalone, V. R., Sams, C. E., Saxton, A. M., West, D. R., Orf, J. H., et al. (2006a). Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. *Theor. Appl. Genet.* 112, 546–553. doi: 10.1007/s00122-005-0161-6

Panthee, D. R., Pantalone, V. R., Saxton, A. M., West, D. R., and Sams, C. (2006b). Genomic regions associated with amino acid composition in soybean. *Mol. Breeding* 17, 79–89. doi: 10.1007/s11032-005-2519-5

Pritchard, J. K., Wen, X., and Falush, D. (2009). *Documentation for STRUCTURE Software: Version 2.3*. Chicago, IL: The University of Chicago Press.

Ren, W., Wen, Y., Dunwell, J. M., and Zhang, Y. M. (2018). pKWmEB: integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Shim, H. C., Ha, B. K., Yoo, M., and Kang, S. T. (2015). Detection of quantitative trait loci controlling UV-B resistance in soybean. *Euphytica* 202, 109–118. doi: 10.1007/s10681-014-1233-y

Swanson-Wagner, R., DeCook, R., Jia, Y., Bancroft, T., Ji, T., Zhao, X., et al. (2009). Paternal dominance of trans-eQTL influences gene expression patterns in maize hybrids. *Science* 326, 1118–1120. doi: 10.1126/science.1178294

Tamba, C., Ni, Y., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tamba, C., and Zhang, Y. M. (2018). A fast mrMLM algorithm for multi-locus genome-wide association studies. *bioRxiv* 7. doi: 10.1101/341784

van Rooijen, R., Kruijer, W., Boesten, R., van Eeuwijk, F. A., Harbinson, J., and Aarts, M. G. M. (2017). Natural variation of yellow seedling1affects photosynthetic acclimation of *Arabidopsis thaliana*. *Nat. Commun.* 8:1421. doi: 10.1038/s41467-017-01576-3

Vance, C. P., Uhdestone, C., and Allan, D. L. (2003). Phosphorus acquisition and use: critical adaptations by plants for securing a nonrenewable resource. *New Phytol.* 157, 423–447. doi: 10.1046/j.1469-8137.2003.00695.x

Veneklaas, E. J., Lambers, H., Bragg, J., Finnegan, P. M., Lovelock, C. E., Plaxton, W. C., et al. (2012). Opportunities for improving phosphorus-use efficiency in crop plants. *New Phytol.* 195, 306–320. doi: 10.1111/j.1469-8137.2012.04190.x

Visscher, P., Thompson, R., and Haley, C. (1996). Confidence intervals in QTL mapping by bootstrapping. *Genetics* 143, 1013–1020.

Wang, J., Chu, S., Zhang, H., Zhu, Y., Cheng, H., and Yu, D. (2016a). Development and application of a novel genome-wide SNP array reveals domestication history in soybean. *Sci. Rep.* 6:20728. doi: 10.1038/srep20728

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016b). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wang, X., Wang, H., Liu, S., Ferjani, A., Li, J., Yan, J., et al. (2016c). Genetic variation in zmvpp1 contributes to drought tolerance in maize seedlings. *Nat. Genet.* 48, 1233–1241. doi: 10.1038/ng.3636

Wang, X., Wurmser, C., Pausch, H., Jung, S., Reinhardt, F., Tetens, J., et al. (2012). Identification and dissection of four major QTL affecting milk fat content in the german holstein-friesian population. *PLoS ONE* 7:e40711. doi: 10.1371/journal.pone.0040711

Wang, X. R., Shen, J. B., and Liao, H. (2010). Acquisition or utilization, which is more critical for enhancing phosphorus efficiency in modern crops?. *Plant Sci.* 179, 302–306. doi: 10.1016/j.plantsci.2010.06.007

Wang, Y., and Gai, J. (2002). Study on the ecological regions of soybean in China.?. Ecological environment and representative varieties. *J. appl. Ecol.* 13, 71–75.

Wen, Y. J., Zhang, H., Ni, Y. N., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Yamanaka, N., Ninomiya, S., Hoshi, M., Tsubokura, Y., Yano, M., Nagamura, Y., et al. (2001). An informative linkage map of soybean reveals QTLs for flowering time, leaflet morphology and regions of segregation distortion. *DNA Res.* 8, 61–72. doi: 10.1093/dnares/8. 2.61

Yin, Z., Meng, F., Song, H., He, X., Xu, X., and Yu, D. (2010a). Mapping quantitative trait loci associated with chlorophyll a fluorescence parameters in soybean (*Glycinemax* (L) Merr). *Planta* 231, 875–885. doi: 10.1007/s00425-009-1094-0

Yin, Z., Meng, F., Song, H., Wang, X., Xu, X., and Yu, D. (2010b). Expression quantitative trait loci analysis of two genes encoding rubisco activase in soybean. *Plant Physiol.* 152, 1625–1637. doi: 10.1104/pp.109.148312

Yu, J., and Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. *Curr. Opin. Biotechnol.* 17, 155–160. doi: 10.1016/j.copbio.2006.02.003

Zhang, D., Cheng, H., Geng, L., Kan, G., Cui, S., Meng, Q., et al. (2009). Detection of quantitative trait loci for phosphorus deficiency tolerance at soybean seedling stage. *Euphytica* 167, 313–322. doi: 10.1007/s10681-009-9880-0

Zhang, D., Kan, G., Hu, Z., Cheng, H., Zhang, Y., Wang, Q., et al. (2014a). Use of single nucleotide polymorphisms and haplotypes to identify genomic regions associated with protein content and water-soluble protein content in soybean. *Theor. Appl. Genet.* 127, 1905–1915. doi: 10.1007/s00122-014-2348-1

Zhang, D., Li, H., Wang, J., Zhang, H., Hu, Z., Chu, S., et al. (2016). High-density genetic mapping identifies new major loci for tolerance to low-phosphorus stress in soybean. *Front. Plant Sci.*.7:372. doi: 10.3389/fpls.2016.00372

Zhang, D., Song, H., Cheng, H., Hao, D., Wang, H., Kan, G., et al. (2014b). The acid phosphatase-encoding gene *GmACP1* contributes to soybean tolerance to low-phosphorus stress. *PLoS Genet.* 10:e1004061. doi: 10.1371/journal.pgen.1004061

Zhang, J., Feng, J. Y., Ni, Y. N., Wen, Y. J., Niu, Y., Tamba, C., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multi-locus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y. M., Mao, Y. C., Xie, C., Smith, H., Luo, L., and Xu, S. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics* 169, 2267–2275. doi: 10.1534/genetics.104.033217

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer JZ declared a shared affiliation, though no other collaboration, with one of the authors YY to the handling Editor.

ORIGINAL RESEARCH
published: 03 September 2018
doi: 10.3389/fpls.2018.01276

Genome-Wide Association Studies Reveal Genetic Variation and Candidate Genes of Drought Stress Related Traits in Cotton (*Gossypium hirsutum* L.)

Sen Hou, Guozhong Zhu, Yuan Li, Weixi Li, Jie Fu, Erli Niu, Lechen Li, Dayong Zhang and Wangzhen Guo*

State Key Laboratory of Crop Genetics & Germplasm Enhancement, Hybrid Cotton R & D Engineering Research Center, Ministry of Education, Nanjing Agricultural University, Nanjing, China

Cotton is an important industrial crop worldwide and upland cotton (*Gossypium hirsutum* L.) is most widely cultivated in the world. Due to ever-increasing water deficit, drought stress brings a major threat to cotton production. Thus, it is important to reveal the genetic basis under drought stress and develop drought tolerant cotton cultivars. To address this issue, in present study, 319 upland cotton accessions were genotyped by 55,060 single nucleotide polymorphisms (SNPs) from high-density CottonSNP80K array and phenotyped nine drought tolerance related traits. The two datasets were used to identify quantitative trait nucleotides (QTNs) for the above nine traits using multi-locus random-SNP-effect mixed linear model method. As a result, a total of 20 QTNs distributed on 16 chromosomes were found to be significantly associated with six drought tolerance related traits. Of the 1,326 genes around the 20 QTNs, 205 were induced after drought stress treatment, and 46 were further mapped to Gene ontology (GO) term "response to stress." Taken genome-wide association study (GWAS) analysis, RNA-seq data and qRT-PCR verification, four genes, *RD2* encoding a response to desiccation 2 protein, *HAT22* encoding a homeobox-leucine zipper protein, *PIP2* encoding a plasma membrane intrinsic protein 2, and *PP2C* encoding a protein phosphatase 2C, were proposed to be potentially important for drought tolerance in cotton. These results will deepen our understanding of the genetic basis of drought stress tolerance in cotton and provide candidate markers to accelerate the development of drought-tolerant cotton cultivars.

Keywords: upland cotton, drought stress, genome-wide association study, single nucleotide polymorphism (SNP), RNA-sequencing

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Zhongxu Lin,
Huazhong Agricultural University,
China
Baohong Zhang,
East Carolina University, United States

***Correspondence:**
Wangzhen Guo
moelab@njau.edu.cn

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *10 May 2018*
Accepted: *14 August 2018*
Published: *03 September 2018*

Citation:
Hou S, Zhu G, Li Y, Li W, Fu J, Niu E, Li L, Zhang D and Guo W (2018) Genome-Wide Association Studies Reveal Genetic Variation and Candidate Genes of Drought Stress Related Traits in Cotton (Gossypium hirsutum L.). Front. Plant Sci. 9:1276. doi: 10.3389/fpls.2018.01276

INTRODUCTION

Cotton (*Gossypium* spp.) is the most important natural fiber crop and is also a significant oilseed crop. The upland cotton (*Gossypium hirsutum* L.), which accounts for 95% of the annual cotton production worldwide, is the most cultivated species. However, cotton production is limited by various abiotic and biotic stresses. Of them, drought stress becomes a major threat to substantial loss of cotton yield due to the ever-increasing scarcity of water around the world (Pettigrew, 2004). There is a urgent need to ascertain the molecular and genetic basis underlying the cotton response to drought stress and to develop cotton cultivars with improved drought tolerance.

Quantitative trait loci (QTLs) mapping was an effective tool which was generally used to reveal the genetic basis of complex quantitative traits in cotton (Li C. et al., 2013; Fang et al., 2014; Wang et al., 2014). However, using traditional molecular markers, such as restriction fragment length polymorphism (RFLP) and simple sequence repeat (SSR), only a few QTLs related to drought stress were discovered in a wide region (Levi et al., 2009; Saleem et al., 2015; Zheng et al., 2016) because of the narrow genetic diversity and low map density in modern upland cotton accessions (Fang et al., 2017; Huang et al., 2017; Sun et al., 2017). In addition, these QTLs have not been applied to cotton breeding either. Recently, with the development of functional genomics and transcriptomics, a plenty of genes were found to be involved in drought resistance, including protein kinases, transcription factors and some structural genes (Ashraf et al., 2018). And these factors are related to different signal transduction pathway in response to drought stress. For example, *GbMYB5* and *GhWRKY17* are positively involved in response to drought stress (Yan et al., 2014; Chen et al., 2015), while a complete MAP kinase cascade that phosphorylates and activates *GhWRKY59* is involved in abscisic acid (ABA)-independent signaling pathway to regulate cotton drought responses (Li et al., 2017). And overexpression of *GhNAC2* can enhance root growth and improve tolerance to drought in transgenic cotton and *Arabidopsis* (Gunapati et al., 2016). Nevertheless, the functional verification for only few genes related to drought tolerance was reported. How to excavate more drought stress related genes accurately and effectively and to utilize them for breeding drought-tolerant cotton cultivars remain a big challenge.

Compared with traditional molecular markers, single nucleotide polymorphisms (SNPs) are the most abundant DNA variation distributed along the genome, with high density, bi-allelic and co-dominant characteristics. Recent years, the application of SNP arrays (Hulse-Kemp et al., 2015; Cai et al., 2017), sequencing and re-sequencing (Li et al., 2015; Zhang T. et al., 2015) for upland cotton accessions made it possible to improve the resolution of genetic maps and the accuracy of QTL mapping. Genome-wide association study (GWAS) is an effective method, which can associate phenotypes with genotypes in natural populations and reveal vast natural allelic variations and candidate genes based on linkage disequilibrium (LD), and have been widely used in crop plants, such as rice (Huang et al., 2010; Yano et al., 2016), maize (Li H. et al., 2013; Yang et al., 2014), and soybean (Zhang J. et al., 2015). Based on the rapid-developed studies for genome-wide SNPs in cotton, GWAS for several important agronomic traits has been performed. Using CottonSNP63K SNP array and 719 diverse *G. hirsutum* accessions, GWAS was conducted by integrating different environment tests of fiber quality traits with the SNP genotyping data, and forty-six SNPs were found to be significantly associated with five fiber quality traits (Sun et al., 2017). Using the same CottonSNP63K SNP array, 503 *G. hirsutum* accessions were genotyped for a GWAS with sixteen agronomic traits, and a total of 324 SNPs and 160 candidate quantitative trait nucleotide (QTN) regions were found to be significantly associated with these agronomic traits (Huang et al., 2017). With genome-wide resequencing for 318 cotton landraces and modern improved accessions or lines, 119 associated loci, including 71 for yield-related traits, 45 for fiber qualities and three for resistance to *Verticillium* wilt were identified by GWAS (Fang et al., 2017). Similarly, with resequencing 267 cotton accessions, 19 candidate loci for fiber-quality-related traits were reported (Wang et al., 2017). Recently, through integrating genotyping variation and phenotyping data of 13 fiber-related traits across 12 environments for 419 diverse *G. hirsutum* accessions, 7,383 unique SNPs were found to be significantly associated with these traits. The results showed that more associated loci were identified for fiber quality than fiber yield, and fiber genes in the D subgenome were more than those in the A subgenome (Ma et al., 2018). These studies indicated that GWAS was suitable for detecting QTNs of complex traits in plants. Nevertheless, GWAS based on SNP markers and using large natural populations for drought tolerance related traits has not been reported in cotton.

To determine the key QTN regions and candidate genes significantly associated with cotton drought response, we deployed GWAS of drought stress through integrating the genotypic data of upland cotton accessions by the high-density CottonSNP80K array analysis with their various phenotypic data response to drought stress. Further, the candidate genes were screened by integrating GWAS and gene expression data with qRT-PCR confirmation. This study will provide not only elite genetic resources with candidate SNPs but also key genes to accelerate the drought stress improvement of upland cotton.

MATERIALS AND METHODS

Plant Materials

A total of 319 upland cotton accessions, with 306 cultivars/lines collected from China and 13 landraces introduced from the United States, were used in this study. The accessions in China were mainly collected from four different ecological growing regions: the Yellow River region (YRR, 183), the Yangtze River region (YtRR, 82), the Northwestern inland region (NIR, 22), the Northern China region, the specifically early maturation region (NSEMR, 16), and three unknown origin cotton accessions (**Supplementary Table S1**).

Phenotypic Analysis

In 2015 and 2017, the 319 upland cotton accessions were planted in the green house with hydroponics and in the phytotron with soil culture in Nanjing Agricultural University, Nanjing, Jiangsu Province, China, respectively. Pilot experiments were performed to screen the suitable concentrations of Polyethylene glycol 6000 (PEG 6000) for investigating the drought tolerance capacities of different upland cotton accessions. As a result, 15% PEG 6000 was used as drought stress treatment for water culture and 10% PEG 6000 for soil culture analysis.

All the accessions were grown in 1/2 diluted Hoagland solution (Hoagland and Arnon, 1950). At the stage of seedlings with five leaves, the experimental plants were treated with 15% PEG 6000 as drought stress, and plants only grown in 1/2 diluted Hoagland solution as control. After 48 h, the indicators related to drought tolerance were measured, including plant height (PH), shoot dry matter (SDM), and root dry matter (RDM). At the same time, top second leaf of the plants were sampled to measure proline content (PC), superoxide dismutase activities (SOD), malonaldehyde content (MDA) and soluble sugar content (SS). In addition, seeds from 319 upland cotton accessions were planted in nursery soil with 10% PEG 6000 solution as drought stress treatment and only watering as control. After 7 days, hypocotyl length (HL) and germination percentage (GP) were measured. For water culture experiments, we selected six seedlings with the relatively uniform growth for each treatment. Each two as a biological replicate, and together three biological replicates were set to investigate the traits, including PH, SDM, RDM, PC, SOD, MDA, and SS. For soil culture experiments, we selected 12 well-developed seeds as a biological replicate for each treatment, with three biological replicates to measure HL and GP.

PH was measured by the length from cotyledonary node to top of the plant. SDM and RDM were measured by weight of aboveground and underground part of the plant after drying at 65°C, respectively. PC was determined by acidic ninhydrin reaction as previously described by Bates et al. (1973). SOD activities were determined by measure inhibition of photochemical reduction of nitro blue tetrazolium (NBT) according to the method of Beyer and Fridovich (1987). MDA content was determined by a modified thiobarbituric acid (TBA) reaction (Hodges et al., 1999). SS content was determined by the colorimetric method using anthrone reagent according to Odjegba and Fasidi (2006). HL was measured by the length of hypocotyl. GP was calculated by ratio of germinated seeds number and planted seeds number.

Factor Analysis

For the drought tolerance evaluation, factor analysis was performed using SPSS software[1]. Kaiser–Meyer–Olkin (KMO) measurement and Bartlett's statistic were calculated to determine the selected variables. Factors were extracted by the cumulative-contribution-rate-more-than-85% rule and comprehensive evaluation of the drought tolerance based on factor scores (Chen, 2013).

SNP Genotyping

Genomic DNA of the 319 cotton accessions was extracted according to the method described by Paterson et al. (1993). A CottonSNP80K array containing 77,774 SNPs (Cai et al., 2017), was used to genotype the 319 accessions. Qualified DNA was hybridized to the array following the Illumina protocols. The Illumina iScan array scanner was used to scan arrays, and GenomeStudio Genotyping software (V2011.1, Illumina, Inc.) was employed to cluster SNP alleles and genotyping. All 77,774 SNPs were tested and manually adjusted as described by Cai et al. (2017). The SNP data set with a calling rate < 0.9 and MAF < 0.05 was further filtered, and the high quality data was used for subsequent analysis.

Population Characteristics and Linkage Disequilibrium Analysis

PLINK V1.90 software[2] was used to conduct the similarity analysis and clustering of 319 cotton accessions. Based on the distance matrix data (1-IBS, identity-by-state), phylogenetic trees were constructed using TASSEL 5.0 software[3], and visually edited by Figtree software[4]. The IBS matrix data was used to conduct principal component analysis (PCA). The correlation coefficient (r^2) of alleles was calculated to measure linkage disequilibrium (LD) in each group level using PLINK V1.90, and LD blocks containing SNP loci associated with target traits were generated using the R software package "LD heatmap."

GWAS and Identification of Candidate Genes

A total of 55,060 SNPs (calling rate ≥ 0.9 and MAF ≥ 0.05) were used for GWAS. To explore the SNP-trait association, multi-locus random-SNP-effect mixed linear model (mrMLM) (Wang S.B. et al., 2016) was employed using the R package "mrMLM" with the following parameters: Critical *P*-value in rMLM: 0.001; Search radius of candidate gene (Kb): 100; Critical LOD score in mrMLM: 3. And the Q+K model was used. Population structure (Q) matrix was calculated using admixture 1.3 with $k = 3$, and kinship (K) matrix was calculated by the R package "mrMLM". Putative candidate genes were identified flanking 500 Kb of peak SNPs (the most significant SNPs). Gene ontology (GO) analysis was implemented using AgriGO[5], and candidate genes in "response to stress" terms were selected for further analysis.

Transcriptome Sequencing and Quantitative Real-Time PCR Analysis

The seedlings of upland cotton genetic standard line, *G. hirsutum* acc. TM-1, with two simple leaves and one heart-shaped leaf, was treated with 15% PEG. The cotton leaf samples were collected in different time-points after treating 0, 6, 12, 24, 48, and 72 h, respectively. Total RNA was extracted from these samples using the Biospin plant total RNA extraction kit (Bioer Technology Co., Ltd.). After pre-processing the RNA-seq data with an NGS QC toolkit (Patel and Jain, 2012), the reads were mapped to the *G. hirsutum* TM-1 genome using a Tophat spliced aligner with default parameters (Trapnell et al., 2009). The genome-matched reads from each library were assembled with Cufflinks (Trapnell et al., 2012). Cuffmerge was then used to merge the individual transcript assemblies into a single transcript set. Lastly, Cuffdiff was used to detect differentially expressed genes (DEGs) with a cutoff of 0.05 *q*-value. Three biological replicates from each sample were used for all RNA-seq experiments.

[1] http://www.spss.com.cn/

[2] http://www.cog-genomics.org/plink2/

[3] https://tassel.bitbucket.io/

[4] http://tree.bio.ed.ac.uk/

[5] http://bioinfo.cau.edu.cn/agriGO/analysis.php

For quantitative real-time PCR (qRT-PCR) analysis, first-strand cDNA was synthesized using the reverse transcription polymerase reaction system (Promega, United States) and adjusted to ∼100 ng/μL with a One Drop Spectrophotometer OD-1000+ (OneDrop, Nanjing, China). qRT-PCR was deployed on an ABI 7500 real-time PCR system[6]. The qRT-PCR amplification program previously described by Provenzano and Mocellin (2007) was used. The relative expression level was calculated using the $2^{-\Delta CT}$ method (Livak and Schmittgen, 2001) with three biological and technical replicates, respectively. The expression level of *GhHis3* (Accession No. AF024716) was used as an internal control (Gutierrez et al., 2008). All the primers were summarized in **Supplementary Table S2**.

Statistical Analysis

Correlation analysis among drought tolerance traits was performed using SPSS software, * and ** present the significant differences at the 5% and 1% levels, respectively. qRT-PCR data was analyzed using Excel software and shown as the mean ± SD. Multiple comparison in one-way ANOVA was conducted by LSD method at the 0.05 and 0.01 levels, which were marked by * and **, respectively.

RESULTS

Phenotypic Variation in Drought Tolerance Related Traits

To evaluate the phenotypic variation under drought stress in the natural population, the seeds or seedlings of 319 upland cotton accessions were treated in PEG stress and in well-watered controls. Nine traits related to drought stress tolerance were measured, including HL and GP at germinating stage; PH, SDM, RDM, PC, SOD activities, MDA content and SS content at seedling stage, respectively. The mean and extreme values of five drought-tolerance traits were lower under drought stress than that in control plants and four traits showed higher values under drought stress condition than that in control plants. The coefficients of variation (CV, %) of nine drought tolerance related traits ranged from 11.77 (HL) to 95.89 (PC) under drought stress, and 11.79 (HL) to 92.70 (PC) under well-watered condition, respectively. Furthermore, it showed higher CV value under drought stress treatment than that in control for most drought-tolerance related traits, indicating the wide variation under drought stress among cotton accessions used in this study (**Table 1**).

The relative values of the nine drought-tolerance traits were further calculated using the ratio of the phenotypic effects value under drought stress and that under well-watered control condition. Relative values of each trait were conformed to Gaussian distribution (**Supplementary Figure S1**). To explore the relationships among nine drought-tolerance traits, correlation analysis was conducted. As a result, relative PH (RPH), relative SDM (RSDM) and relative RDM (RRDM) showed significant and positive correlation each other. We also detected a significant and positive correlation among different biochemical index, involved in relative PC (RPC), relative MDA (RMDA), relative SS (RSS), and relative SOD (RSOD). In addition, relative HL (RHL) showed significant and positive correlation with relative GP (RGP) and RSDM. However, both RHL and RSDM showed a significant and negative correlation with RPC, RMDA, and RSS, respectively (**Table 2**).

Comprehensive Evaluation of Drought Tolerance

In order to identify the drought tolerance of 319 cotton accessions, factor and cluster analyses were performed with relative values of nine drought-tolerance traits. The KMO value was 0.634 (>0.5), and the Bartlett's statistic value $p < 0.05$, indicating that the raw data was suitable for factor analysis. A six-factor solution that accounted for 89.45% of the total variance was obtained (**Supplementary Table S3**) (Chen, 2013). Factor 1 represented the biochemical index factor at seedling stage, including RPC, RMDA, and RSS. Factor 2 was regarded as the physiological index factor, including RPH and RSDM. Factors 3–6 represented RSOD, RHL, RGP, and RRDM, respectively (**Supplementary Table S4**). In addition, the F factor composite score for the drought-tolerance of each cotton accession was calculated by six F factors. Based on drought tolerance capacity with different F factors, cluster analysis showed that the 319 upland cotton accessions were divided into four groups. Totally, 16, 75, 207, and 21 accessions were clustered into advanced, medium, sensitive and extremely sensitive types to drought stress tolerance with F factor ranging from 0.818 to 1.938, 0.239 to 0.742, −0.637 to 0.209, −1.385 to −0.655, respectively (**Supplementary Table S1**).

Genetic Variation Based on SNPs

We genotyped 319 upland cotton accessions using the CottonSNP80K array. GenomeStudio software (V2011.1, Illumina, Inc.) was used to genotype with a manual corrected clustering file (Cai et al., 2017). The genotypic data revealed that these cotton accessions possessed a high average call rate of 99.39%. With low-quality (call rate < 90% and minor allele frequency < 0.05) loci filtered, a final set of 55,060 SNPs was obtained, with 30,075 and 24,985 SNPs in the At and Dt subgenomes, respectively. These SNP markers were distributed over the entire genome, expect chromosomes A02, A03, and A04 with less SNP density. In addition, the polymorphism information content (PIC) values ranged from 0.263 to 0.389 among chromosomes, and the mean PIC value of the At and Dt subgenomes was 0.338 and 0.334, respectively (**Table 3**).

Population Structure and Linkage Disequilibrium

Principal component analysis and neighbor-joining tree were conducted to infer population stratification. A pairwise distance matrix derived from a modified Euclidean distance for all polymorphic SNPs was calculated to construct neighbor-joining trees using TASSEL 5.0 software. As a result, the 319 accessions could be clustered into four groups, which contained 61, 33,

[6] http://www.lifetechnologies.com

TABLE 1 | Statistics of various traits related to drought tolerance.

Traits	Control					PEG treatment				
	Minimum	Maximum	Mean	*SD*	CV(%)	Minimum	Maximum	Mean	*SD*	CV (%)
PH (cm)	19.46	38.57	28.11	3.96	14.11	14.47	30.41	23.00	3.32	14.43
SDM (g)	0.20	0.74	0.42	0.10	24.57	0.04	0.49	0.28	0.08	30.06
RDM (g)	0.03	0.19	0.09	0.03	36.52	0.01	0.11	0.05	0.02	40.35
PC (μg/g.FW)	2.96	253.23	42.12	39.04	92.70	14.42	4451.77	833.79	799.55	95.89
SOD (U/g.FW)	14.81	163.95	75.23	27.20	36.15	19.16	635.79	156.96	76.48	48.73
MDA (nmoL/g.FW)	8.68	34.79	15.37	3.82	24.87	10.73	86.80	26.55	11.12	41.87
SS (mg/g.FW)	3.16	22.84	6.30	1.86	29.55	4.89	72.32	20.57	11.78	57.27
HL (cm)	3.07	7.33	5.22	0.62	11.79	2.50	6.08	4.25	0.50	11.77
GP (%)	19.44	100.00	79.07	13.98	17.68	13.89	97.22	73.24	13.99	19.10

PH, plant height; SDM, shoot dry matter; RDM, root dry matter; PC, proline content; SOD, superoxide dismutase activity; MDA, malonaldehyde content; SS, soluble sugar content; HL, hypocotyl length; GP, germination percentage; SD, standard deviations; CV, coefficient of variation; FW, fresh weight.

TABLE 2 | Correlation analysis of drought tolerance traits.

	RPH	RSDM	RRDM	RPC	RSOD	RMDA	RSS	RHL	RGP
RPH	–								
RSDM	0.733**	–							
RRDM	0.262**	0.203**	–						
RPC	−0.063	−0.106*	0.051	–					
RSOD	−0.154**	−0.073	0.011	0.242**	–				
RMDA	−0.091	−0.111*	0.112*	0.621**	0.555**	–			
RSS	−0.074	−0.121*	0.002	0.554**	0.333**	0.596**	–		
RHL	0.041	0.142**	−0.015	−0.145**	−0.030	−0.181**	−0.128*	–	
RGP	−0.005	−0.069	0.000	0.024	0.011	−0.007	−0.042	0.127*	–

RPH, relative plant height; RSDM, relative shoot dry matter; RRDM, relative root dry matter; RPC, relative proline content; RSOD, relative superoxide dismutase activity; RMDA, relative malonaldehyde content; RSS, relative soluble sugar content; RHL, relative hypocotyl length; RGP, relative germination percentage.
** and **: the 5% and 1% levels of significance, respectively.*

87, and 138 accessions, respectively. We found that the four clustered groups had no relationships with their geographic origin (**Figure 1A** and **Supplementary Table S1**). Further, clustering data in the phylogenetic tree matched to PCA results well (**Figure 1B**). We also performed linkage disequilibrium (LD) analysis using PLINK software and evaluated Pairwise LD using squared allele frequency correlations (r^2). The LD rate declining to half its maximum value was 980 Kb, with >1000 Kb in At and 760 Kb in Dt subgenome, respectively (**Figure 2**).

Genome-Wide Association Studies

The GWAS was conducted for the nine traits related to drought tolerance using mrMLM method with Q+K model. Twenty SNPs were found to be significantly associated with six drought tolerance related traits, with both 10 SNPs located on At and Dt subgenome (**Table 4**). In detail, there were three SNPs located on chromosomes A10, D01, and D13 for RRDM, explaining 22.66% of the total phenotypic variation. For RHL, six SNPs were identified on chromosomes A11, A04, A09, D04, D06, and D09, explaining 50.94% of the total phenotypic variation. Two SNPs for RPC were detected on chromosomes A11 and D07, explaining 11.51% of the total phenotypic variation. Two SNPs for RSS were identified on chromosomes D11 and D12, explaining 40.61% of the total phenotypic variation. Four SNPs for RSDM were identified on chromosomes A03, A05, A06, and D06, explaining 29.85% of the total phenotypic variation. For RPH, three SNPs were identified on chromosomes A05, A08, and D12, explaining 26.08% of the total phenotypic variation. The widespread associated loci in different chromosomes indicated that the genetic basis of drought tolerance is complex.

Candidate Genes Associated With Significant SNPs Region

Candidate genes involved in the 20 significant SNP loci were further mined by referring the LD value in the study. With flanking 500 kb of the significantly associated SNPs and *G. hirsutum* TM-1 genome (Zhang T. et al., 2015) as reference, 1,326 candidate genes with 630 in At and 696 in Dt subgenome were identified. The number of candidate genes associated with six traits was predicted. We found 398 candidate genes from significant SNP regions associated with RHL, while only 70 candidate genes with RPC (**Table 4**). These results implied that HL was involved in a complex process regulated by more regulators while the regulation of PC was relatively specific. GO analysis indicated that 1,226 candidate genes could be mapped to GO background

TABLE 3 | Summary of high quality SNPs by genotyping analysis using CottonSNP80K array.

Chr.	Total SNPs	Filtered SNPs	Chr. size (Mb)	Density of SNP (Kb/SNP)	PIC
A01	3500	2365	99.88	42.23	0.372
A02	1996	1410	83.45	59.18	0.349
A03	2466	1792	100.26	55.95	0.348
A04	1434	1055	62.91	59.63	0.362
A05	3384	2485	92.05	37.04	0.352
A06	4698	2525	103.17	40.86	0.287
A07	3070	2132	78.25	36.7	0.340
A08	7773	4967	103.63	20.86	0.286
A09	3621	2440	75	30.74	0.332
A10	2964	2035	100.87	49.57	0.330
A11	2897	1890	93.32	49.38	0.325
A12	3040	1994	87.48	43.87	0.355
A13	4340	2985	79.96	26.79	0.356
D01	2339	1852	61.46	33.19	0.359
D02	2985	2350	67.28	28.63	0.378
D03	1889	1321	46.69	35.34	0.263
D04	1272	1019	51.45	50.49	0.355
D05	2041	1596	61.93	38.8	0.339
D06	4037	3191	64.29	20.15	0.316
D07	3472	2617	55.31	21.13	0.326
D08	2898	2256	65.89	29.21	0.389
D09	2938	2147	51	23.75	0.296
D10	2130	1701	63.37	37.25	0.338
D11	1866	1409	66.09	46.91	0.335
D12	2562	1911	59.11	30.93	0.324
D13	2162	1615	60.53	37.48	0.322

Chr, chromosome; PIC, polymorphism information content.

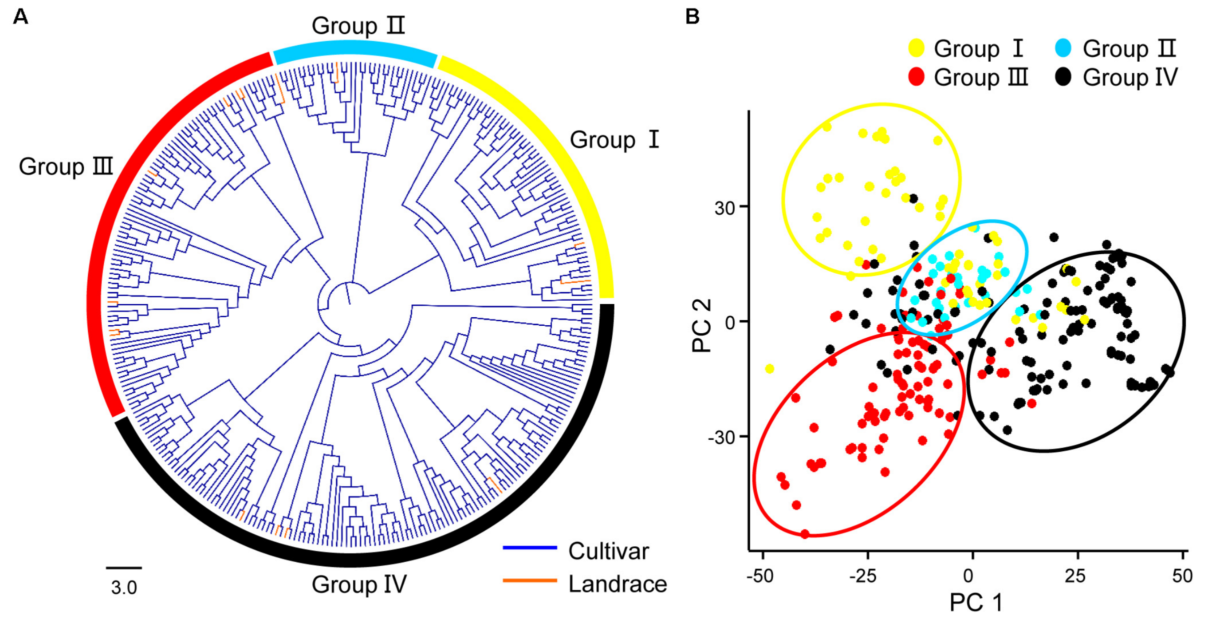


FIGURE 1 | Population structure of 319 upland cotton accessions. **(A)** Neighbor-joining tree of 319 cotton accessions in the panel. Cultivars and Landraces are shown by blue and orange line, respectively. **(B)** PCA plots of the accessions.

TABLE 4 | Summary of SNPs associated with drought tolerance traits.

Traits	SNP IDs	Chr.	Pos. (Mb)	QTN effect	LOD score	r^2 (%)	Number of candidate genes
RRDM	TM36896	A10	100.64	−0.6333	3.41	14.57	59
	TM82142	D13	58.86	−0.5046	3.71	4.82	103
	TM50155	D01	59.85	−0.3164	4.47	3.27	91
RHL	TM37191	A11	6.02	0.0631	4.86	14.14	84
	TM9833	A04	60.55	0.0246	6.21	8.60	83
	TM55926	D04	10.80	0.0192	4.23	4.30	44
	TM30039	A09	2.63	0.0244	3.40	2.76	51
	TM72632	D09	44.33	−0.0398	4.21	16.88	81
	TM62940	D06	60.15	−0.0253	3.71	4.26	55
RPC	TM66782	D07	54.23	−19.9393	4.34	6.35	67
	TM38632	A11	59.82	−17.9695	3.47	5.16	3
RSS	TM77502	D12	3.61	−1.1742	3.89	19.60	69
	TM75380	D11	4.17	1.2065	3.67	21.01	93
RSDM	TM7846	A03	90.78	1.0222	3.56	5.03	33
	TM10434	A05	8.42	1.217	4.92	6.28	130
	TM13658	A06	1.84	0.597	4.73	2.59	76
	TM59389	D06	8.06	−1.1709	3.26	15.95	49
RPH	TM11090	A05	23.64	−0.0826	4.53	5.54	48
	TM29675	A08	96.79	−0.0545	3.26	2.11	63
	TM77685	D12	6.01	0.0905	3.27	18.43	44

Chr, chromosome; Pos, position; QTN, quantitative trait nucleotide. RRDM, relative root dry matter; RHL, relative hypocotyl length; RPC, relative proline content; RSS, relative soluble sugar content; RSDM, relative shoot dry matter; RPH, relative plant height.

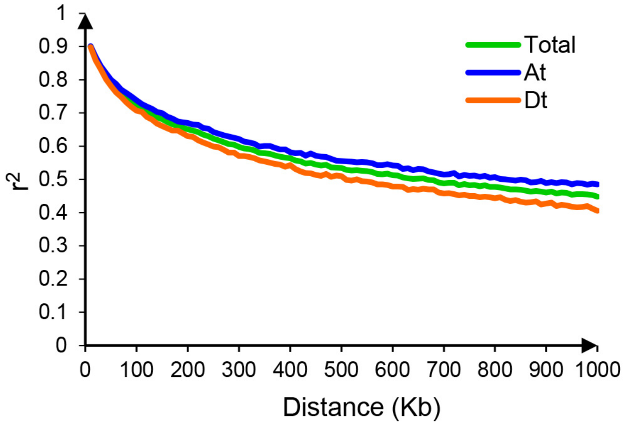


FIGURE 2 | Genome-wide linkage disequilibrium (LD) decay in all cotton accessions. Different colors show the LD decay estimated in different subgenomes.

in *G. hirsutum*, which were involved in several biological processes significantly associated with drought stress, such as root system development, regulation of transport, water transport and response to stress. Further, we focused on the GO term "response to stress" (SR), which contained 189 candidate genes (**Figure 3A**). Of them, many genes had been reported to play important roles in drought tolerance, such as *RD2*, *PIP2*, *PP2C,* and *LEA* (Aroca et al., 2006; Samota et al., 2017; Baek et al., 2018; Magwanga et al., 2018), and some key transcription factors, *WRKY*, *NAC,* and *MYB* (Wei et al., 2017; Kiranmai et al., 2018; Negi et al., 2018).

Transcriptome Analysis and Identification of Elite Alleles

RNA-seq analysis was performed to further explore elite alleles which contributed to drought tolerance. RNA samples were collected from leaves of *G. hirsutum* acc. TM-1 at 0, 6, 12, 24, 48, and 72 h post treatment of 15% PEG. In total, 18 separate libraries were generated with three biological replicates of each sample. The reads generated by the Illumina Hiseq2000 were initially processed to remove adapter sequences and low-quality bases. Approximately 0.82 billion valid reads, each 150 nucleotides long, and roughly 40.8 Gb of nucleotides were obtained. We investigated the expression level of all 1,326 candidate genes [$\log_2(\text{RPKM}+1) > 1$] from GWAS analysis, and found that 205 were differential expression genes (DEGs) with significant induced expression under drought stress condition compared with untreated control. Among these DEGs, 46 were annotated to "response to stress" in GO dataset and other 159 were novel candidate genes (**Figure 3A** and **Supplementary Table S5**). For these 46 DEGs, up-regulated genes are more than down-regulated genes after stress-treated time points (**Figure 3B**). Some up-regulated genes were positively related to stress tolerance such as *RD2* (Samota et al., 2017), and a few down-regulated gene were reported to be negatively related to stress tolerance such as *PIP2* (Macho et al., 2018). In addition, other 143 candidate genes involved in GO term "response to stress" were not induced by drought stress (**Figure 3A** and **Supplementary Figure S2**).

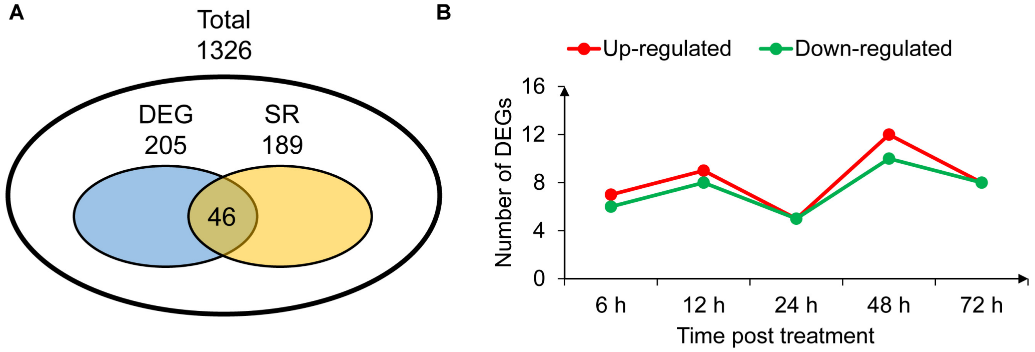


FIGURE 3 | Transcriptome analysis and identification of elite alleles. **(A)** Distribution of the differentially expressed genes (DEGs) and response to stress genes (SR). **(B)** Statistics of the up-regulated and down-regulated DEGs annotated as response to stress.

We combined the GWAS and drought-induced RNA-seq data to explore elite alleles involved in drought tolerance. As a result, four genes were further identified for potential drought tolerance. Within the association signal at D12: 3609663 which explaining approximately 19.60% of the phenotypic variation of RSS, we identified 69 candidate genes. The RNA-seq data showed that one of these genes, which encoded a response to desiccation 2 protein and named as *RD2* (Gh_D12G0260), was continuously up-regulated in all five time points, especially in 72 h post PEG treatment (**Figure 4**). Another gene within association signal at D11: 4173831, which explaining approximately 21.01% of the phenotypic variation of RSS, was also continuously up-regulated after PEG treatment. This gene encodes a homeobox-leucine zipper protein, named *HAT22* (Gh_D11G0526), which has been reported to be related to plant stress tolerance (Liu et al., 2016) (**Figure 5**). Within the association signal at

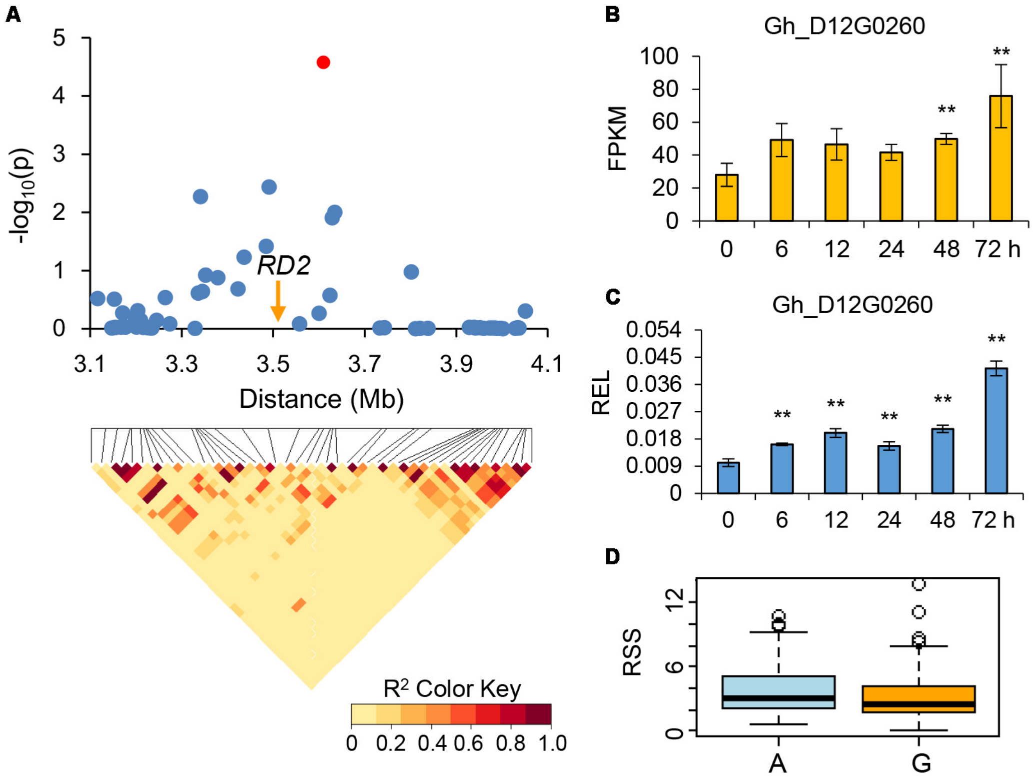


FIGURE 4 | Genome-wide association study (GWAS) for drought-tolerance and identification of the candidate gene *RD2* on chromosome D12. **(A)** Local Manhattan plot (top) and LD heat map (bottom). The red dot indicates the SNP related to the drought-tolerance trait. The arrow indicates the location of *RD2*. **(B)** The expression level of the candidate gene *RD2* calculated via RNA-Seq. **(C)** The relative expression level (REL) of the candidate gene *RD2* calculated via qRT-PCR. **(D)** Differences of relative soluble sugar content (RSS) among two haplotypes. ** means the 1% level of significance.

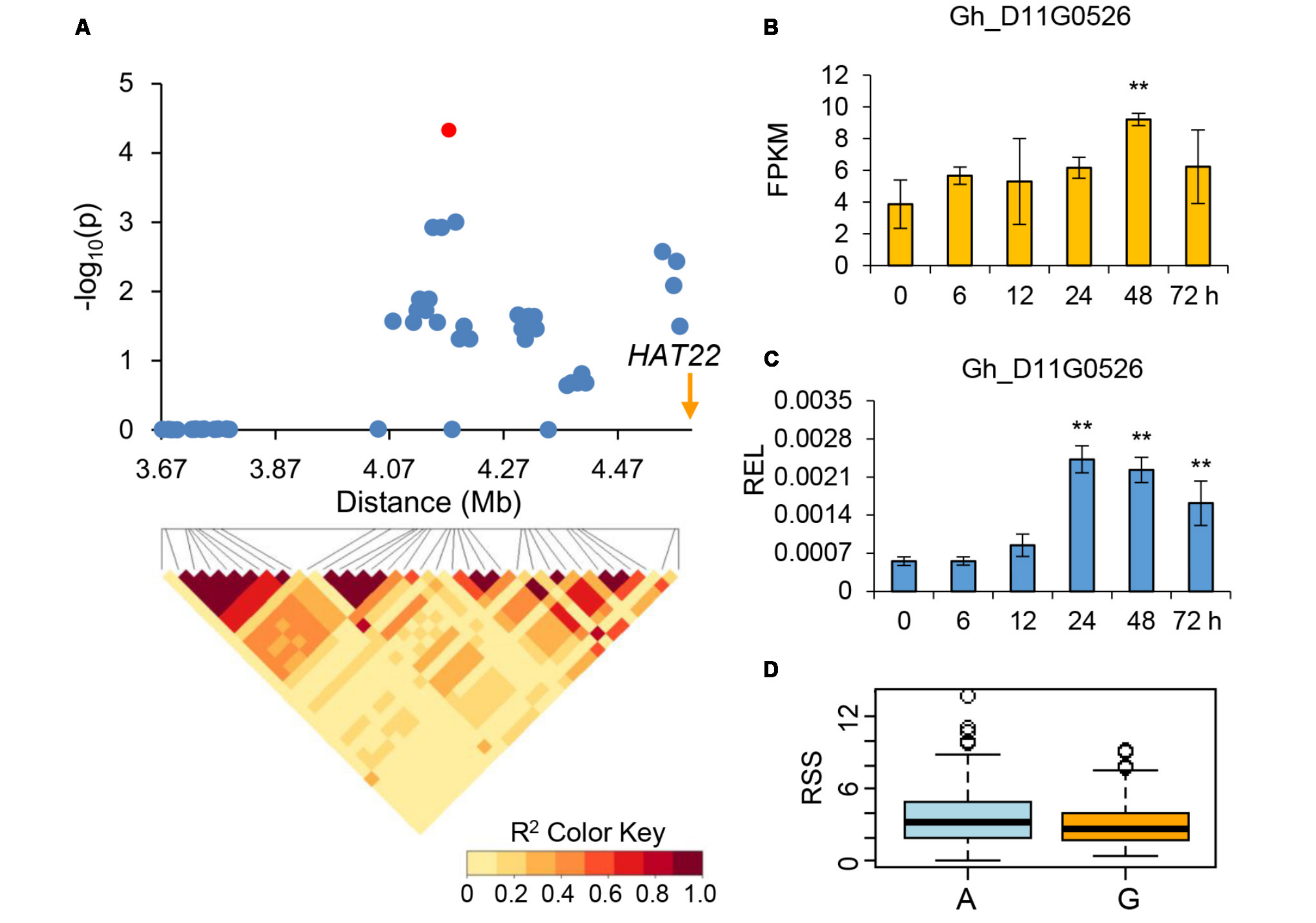


FIGURE 5 | Genome-wide association study for drought-tolerance and identification of the candidate gene *HAT22* on chromosome D11. **(A)** Local Manhattan plot (top) and LD heat map (bottom). The red dot indicates the SNP related to the drought-tolerance trait. The arrow indicates the location of *HAT22*. **(B)** The expression level of the candidate gene *HAT22* calculated via RNA-Seq. **(C)** The relative expression level (REL) of the candidate gene *HAT22* calculated via qRT-PCR. **(D)** Differences of relative soluble sugar content (RSS) among two haplotypes. ** means the 1% level of significance.

D01: 59846909, which explaining approximately 3.27% of the phenotypic variation of RRDM, we identified a gene down-regulated after drought stress, encoding a plasma membrane intrinsic protein 2 (*PIP2*, Gh_D01G2086), which was involved in root water uptake and tissue hydraulic conductance (Macho et al., 2018) (**Figure 6**). Another gene within association signal at D04: 10799426, explaining approximately 4.30% of the phenotypic variation of RHL, was also down-regulated after PEG treatment. This gene encodes a protein phosphatase 2C, named *PP2C* (Gh_D04G0612), which negatively regulate ABA signaling and stress responses (Baek et al., 2018) (**Figure 7**). In order to validate the reliability of the RNA-seq data, we conducted qPCR assay and confirmed their expression patterns, which were kept consistent in all four candidate genes (**Figures 4–7**).

DISCUSSION

Drought is a serious global problem restricting agricultural development. Previous studies on cotton drought-tolerance mainly pay attention to a single period of cotton development or a small number of indexes or accessions (Chen et al., 2013; Zheng et al., 2014; Ranjan and Sawant, 2015). In this study, 319 upland cotton accessions were collected with a high geographical diversity for genome-wide association studies. A total of nine traits related to drought-tolerance were measured containing two traits at the germinating stage (HL and GP), and seven traits at the seedling stages (PH, SDM, RDM, PC, SOD, MDA, and SS). Proline content has the largest CV, which means that proline content is sensitive to drought stress, and can be regarded as one of the most important trait related to drought tolerance. When suffered drought stress, plants can adapt in three ways: physiological responses such as reducing growth rates, molecular responses such as the increased expression in ABA biosynthetic genes, and biochemical responses such as accumulation of stress metabolites like proline, glutathione, glycinebetaine, polyamines and so on (Fang and Xiong, 2015). Accumulation of proline content, which can increase plant cell's osmotic pressure and retard plant losing water, is an efficient mechanism to improve drought tolerance. Hypocotyl length

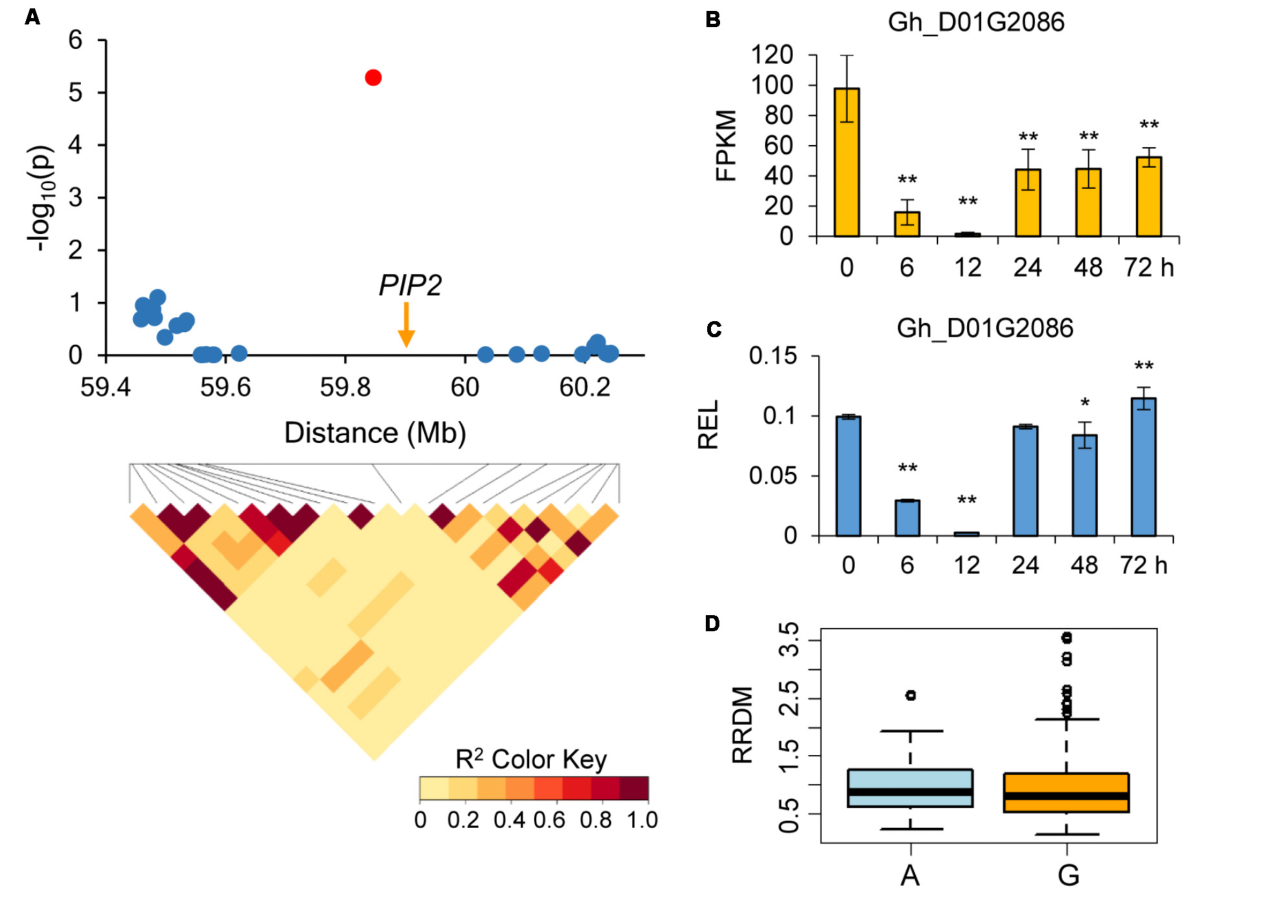


FIGURE 6 | Genome-wide association study for drought-tolerance and identification of the candidate gene *PIP2* on chromosome D01. **(A)** Local Manhattan plot (top) and LD heat map (bottom). The red dot indicates the SNP related to the drought-tolerance trait. The arrow indicates the location of *PIP2*. **(B)** The expression level of the candidate gene *PIP2* calculated via RNA-Seq. **(C)** The relative expression level (REL) of the candidate gene *PIP2* calculated via qRT-PCR. **(D)** Differences of relative root dry matter (RRDM) among two haplotypes. * and ** mean the 5% and 1% levels of significance, respectively.

associated to the most of SNPs and genes (**Table 4**), indicates that germinating stage is one of the most sensitive stages of cotton, just like sesame (Li et al., 2018). Based on drought tolerance capacity with different F factors, cluster analysis grouped the 319 upland cotton accessions as four types: 16 advanced drought-tolerant accessions, 75 medium drought-tolerant accessions, 207 drought-sensitive accessions and 21 extremly drought-sensitive accessions. In our previous report for salt-tolerance of 304 upland cotton accessions, we detected that 43 accessions were advanced salt-tolerance, 114 medium salt-tolerance, and 147 salt-sensitive (Du et al., 2016). Compared to the salt tolerance, there were relatively few drought-tolerant accessions and need to be further improved for drought tolerance in cotton breeding.

On the basis of the phylogenetic and PCA analysis, we classified the 319 accessions into four groups. However, it showed no obvious relationship with their geographic origin and this result was consistent with most previous study in Upland cotton (Cai et al., 2017; Sun et al., 2017). LD analysis of upland cotton in this study showed that the LD rate declining to half its maximum value was 980 Kb, it is longer than most other crops, such as rice (167 Kb) (Huang et al., 2012) and soybean (420 Kb) (Zhou et al., 2015). The speed of LD decay determines the capacity and resolution of marker-trait association mapping, and the causes of LD mainly including mutation, population bottlenecks, founder effects, drift, selection, migration and population admixture (Morrell et al., 2005; Mackay and Powell, 2007). We speculated that the short history and relatively high rate of self-fertilization of upland cotton breeding in China led to the slower LD decay, implying the more narrow genetic diversity of upland cotton accessions.

Drought tolerance is a complex trait and is regulated by polygenes with small effect. Common GWAS methods are all based on a fixed-SNP-effect mixed linear model (MLM) and single marker analysis, which require Bonferroni correction for multiple tests. When the number of markers is extremely large, the test is too strict. In cotton, combined the resequenced SNP data and phenotyping variation data, GWAS was performed using EMMAX method and the significance threshold was estimated as approximately $P = 10^{-6}$. As a result, three loci associated with resistance to *Verticillium wilt* were identified (Fang et al., 2017), implying that mixed linear model with single marker analysis is too strict in GWAS for complex trait such as

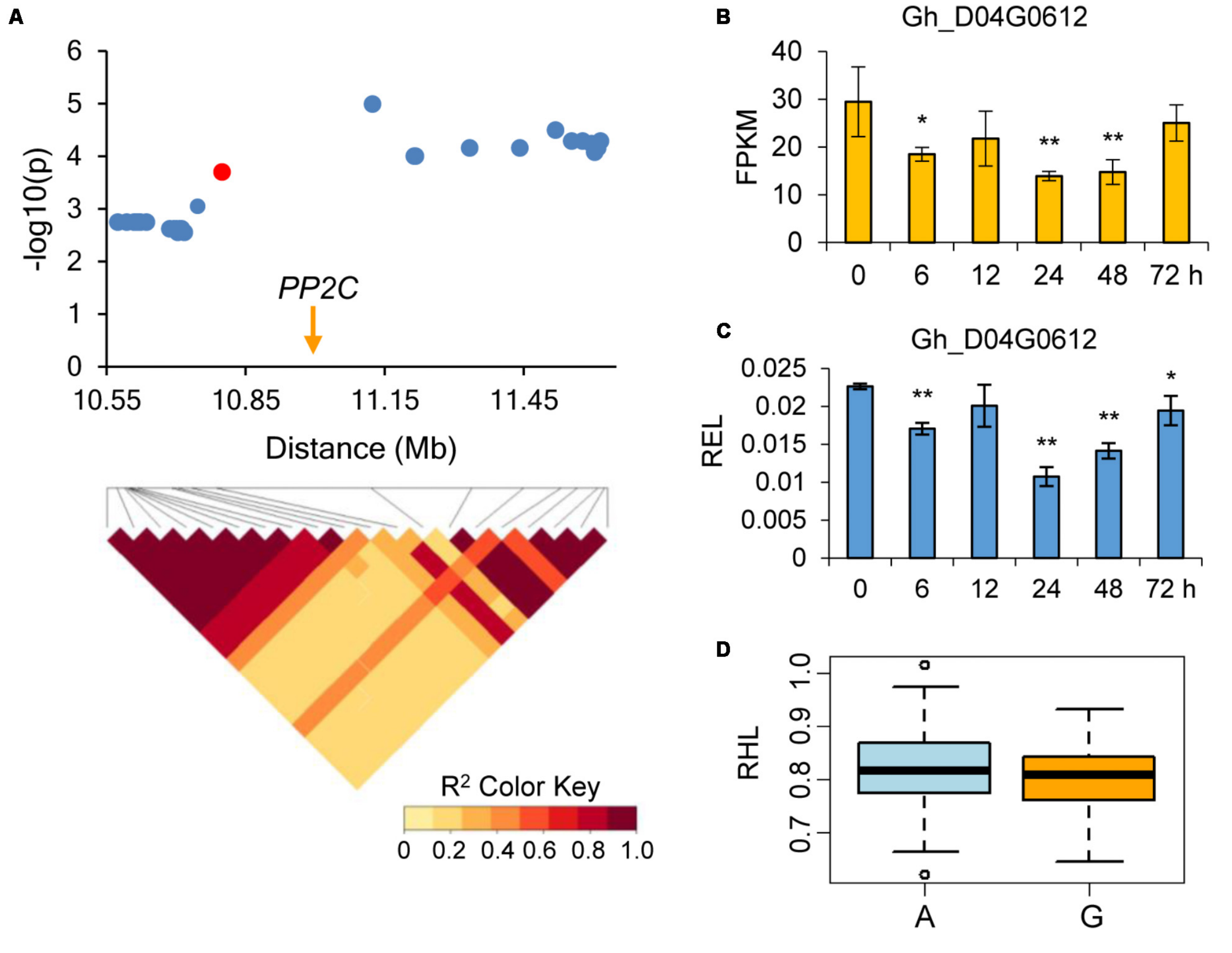


FIGURE 7 | Genome-wide association study for drought-tolerance and identification of the candidate gene *PP2C* on chromosome D04. **(A)** Local Manhattan plot (top) and LD heat map (bottom). The red dot indicates the SNP related to the drought-tolerance trait. The arrow indicates the location of *PP2C*. **(B)** The expression level of the candidate gene *PP2C* calculated via RNA-Seq. **(C)** The relative expression level (REL) of the candidate gene *PP2C* calculated via qRT-PCR. **(D)** Differences of relative hypocotyl length (RHL) among two haplotypes. * and ** mean the 5% and 1% levels of significance, respectively.

disease resistance. Multi-locus mixed linear model compresses the markers by the rMLM method and used the selected SNPs to further associate with traits by mrMLM method. Also due to the multi-locus nature, no multiple test correction is needed. So it shows the good effect for complex traits. Here, we used mrMLM of Wang S.B. et al. (2016) and the CottonSNP80K array, genome-wide association studies of drought-tolerance traits with natural population of upland cotton accessions were conducted, and 20 QTNs for drought-tolerance traits were identified. These associated loci were widely distributed across the entire genome and the candidate genes around the loci involved in many biological process, such as root system development, regulation of transport, water transport and response to stress. It demonstrates that drought stress response is controlled by multiple loci and numerous genes.

Compared to yield and fiber quality traits of cotton (Fang et al., 2017; Sun et al., 2017; Ma et al., 2018), the number of reported loci associated with drought tolerance is much fewer. Previous studies have identified several drought-tolerance QTLs in cotton. Based on the progenies from the cross of *G. hirsutum* cv. Siv'on and *G. barbadense* cv. F-177, a total of 33 QTLs were identified under water-limited environments, including five QTLs for different physiological traits, 11 for plant productivity and 17 for fiber quality, respectively (Saranga et al., 2001). In another study, a vast number of QTLs from 42 different studies were surveyed by comprehensive meta QTL analysis, including 132 QTLs for fiber strength, 26 for boll weight, six for gossypol, four for fruiting banch number, five for osmotic potential, three for chlorophyll and so on (Said et al., 2013). However, compared to the GWAS analysis based on high density SNPs, the low density markers and the wide QTL regions showed limitation in identification and utilization of elite genes, especially for marker-assisted selection (MAS) breeding (Zhou et al., 2015). The present study makes progresses in revealing loci related to drought-tolerance traits and identifying SNP loci and candidate genes for drought tolerance. In other studies, using 240 maize accessions and high-density markers, 61 significant SNPs related to drought tolerance were detected by GWAS analysis (Thirunavukkarasu et al., 2014). In rice, through integrating 175 upland rice accessions with 150,325 SNPs, 13 SNP markers and 50 genes associated with yield under drought conditions were identified, further, 10 genes related to drought and abiotic stress tolerance were verified (Pantalião et al., 2016). These studies could be exploited to discover drought tolerance mechanism and contribute to breeding the drought tolerance varieties in crops.

Genetic basis of drought tolerance is complex. Previous studies have reported many genes that are responsive to drought stress in many plants (Song X. et al., 2016; Wang N. et al., 2016; Ma et al., 2017). However, it is difficult to identify candidate genes from the enormous gene pool. In present study, we performed GWAS to identify elite QTNs in natural population and further screen candidate genes by combining with RNA-seq data. As a result, 46 candidate genes with both annotated as response to stress and differential expression under drought stress were selected. Of these genes, *RD2*, encoding a response to desiccation protein, was a key gene for cotton drought tolerance. In rice, expression analyses showed that both *RD1* and *RD2* genes up-regulated under drought stress due to seed-priming, and *RD2* was increased more significantly than *RD1* in tolerance to drought stress, especially on priming with paclobutrazol in drought-tolerant plant and with salicylic acid in drought-sensitive plant (Samota et al., 2017). Another candidate gene *HAT22*, which also named *ABIG1* and encoded a homeobox protein, is a member of the HD-Zip II family. Expression of *HAT22* mRNA increased under drought and ABA treatment. There was less leaf yellowing in *HAT22* mutants than wild type plants with drought conditions. Moreover, some stress related genes such as ABA and ethylene response loci were regulated by *HAT22* (Liu et al., 2016). In the plasma membrane, PIPs are the most plentiful aquaporins with two types, PIP1 and PIP2. There are five and eight isoforms of PIP1 and PIP2 in *Arabidopsis thaliana*, respectively. Generally, PIP1 proteins behave a low efficiency for water transport while PIP2 proteins in plant have a high capacity for water transport (Maurel et al., 2008). Previous studies showed that PP2Cs played a crucial role in regulation of signal transduction pathways. PP2Cs are negative regulators of stress-induced MAPK pathways, ABA signaling and receptor kinase signaling. In turn, expression of *PP2C* was transcriptionally controlled by developmental signals, ABA and stress response (Schweighofer et al., 2004). Abscisic acid (ABA) is an important plant hormone, and regulates plant development and resistance to biotic and abiotic stresses. It controls transpirational water loss by regulating the stomatal opening and closure to resist drought stress. Moreover, ABA can increase plant cell's osmotic pressure, and expedite chlorophyll breakdown and leaf senescence (Wilkinson and Davies, 2002). By detecting the 46 candidate genes, we found most of them involved in ABA signal pathway and were reported to be related to drought response, such as *PIP2*, *HK1*, *GOLS1*, and *ADC2* (Zhai et al., 2010; Héricourt et al., 2016; Song C. et al., 2016; Macho et al., 2018), indicating ABA signal pathway play crucial roles in response to drought tolerance in cotton. In summary, identification of more drought-tolerance related genes/QTLs enlarges new insight into mechanisms of drought response, and high-throughput genotyping platforms are powerful tools for complex traits dissection and development of drought tolerance varieties in future cotton breeding.

CONCLUSION

We genotyped 319 upland cotton accessions using the CottonSNP80K array and phenotyped nine drought-tolerance related traits. By SNP-trait GWAS, we identified 20 SNPs significantly associated with drought-tolerance traits. Integrating the GWAS and RNA-seq data with qRT-PCR verification, we identified four candidate genes *RD2*, *HAT22*, *PIP2*, and *PP2C* for improving drought stress. Our study provides valuable information to explore molecular mechanisms underlying cotton drought tolerance, and supplies new resource for the improvement of drought-tolerance in future cotton breeding efforts.

AUTHOR CONTRIBUTIONS

WG conceived the study. SH and YL contributed to phenotyping investigation. GZ and WL performed the GWAS and RNA-Seq analysis. SH, JF, EN, LL, and DZ contributed to the experimental data analysis and discussion. SH, GZ, DZ, and WG wrote the manuscript. All authors read and approved the manuscript.

FUNDING

This program was financially supported in part by the Fundamental Research Funds for the Central Universities (KYYJ201601 and KYYJ201701), six talent peaks project in Jiangsu province (2015-NY-002), Qing Lan Project for Science and Technology Innovation Team in Jiangsu Province (No. 6), and Jiangsu Collaborative Innovation Center for Modern Crop Production project (No. 10).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01276/full#supplementary-material

FIGURE S1 | Phenotypic distributions for drought-tolerance traits. RPH, relative plant height; RSDM, relative shoot dry matter; RRDM, relative root dry matter; RPC, relative proline content; RSOD, relative superoxide dismutase activity; RMDA, relative malonaldehyde content; RSS, relative soluble sugar content; RHL, relative hypocotyl length; RGP, relative germination percentage. Gaussian curve is shown in orange line for each trait.

FIGURE S2 | Expression pattern of the candidate genes involved in drought response. (A) Genes were annotated as stress response by Go analysis and also differential expression under drought stress. (B) Genes were not annotated as stress response by Go analysis but differential expression under drought stress.

(C) Genes were annotated as stress response but not induced expression under drought stress. The scales represent the fold change value of gene expression level.

TABLE S1 | Information on 319 cotton accessions used in this study.

TABLE S2 | Primers used in this study.

TABLE S3 | Eigenvalues of six principal components and their contribution and accumulative contribution.

TABLE S4 | Loading matrix of each in six principal components.

TABLE S5 | Information on 46 candidate genes related to stress response with significantly differential expression under drought stress.

REFERENCES

Aroca, R., Ferrante, A., Vernieri, P., and Chrispeels, M. J. (2006). Drought, abscisic acid and transpiration rate effects on the regulation of PIP aquaporin gene expression and abundance in *Phaseolus vulgaris* plants. *Ann. Bot.* 98, 1301–1310. doi: 10.1093/aob/mcl219

Ashraf, J., Zuo, D., Wang, Q., Malik, W., Zhang, Y., Abid, M. A., et al. (2018). Recent insights into cotton functional genomics: progress and future perspectives. *Plant Biotechnol. J.* 16, 699–713. doi: 10.1111/pbi.12856

Baek, W., Lim, C. W., and Lee, S. C. (2018). A DEAD-box RNA helicase, RH8, is critical for regulation of ABA signaling and the drought stress response via inhibition of PP2CA activity. *Plant Cell Environ.* 41, 1593-1604. doi: 10.1111/pce.13200

Bates, L., Waldren, R., and Teare, I. (1973). Rapid determination of free proline for water-stress studies. *Plant Soil* 39, 205–207. doi: 10.1007/BF00018060

Beyer, W., and Fridovich, I. (1987). Assaying for superoxide dismutase activity: some large consequences of minor changes in conditions. *Anal. Biochem.* 161, 559–566. doi: 10.1016/0003-2697(87)90489-1

Cai, C., Zhu, G., Zhang, T., and Guo, W. (2017). High-density 80 K SNP array is a powerful tool for genotyping *G. hirsutum* accessions and genome analysis. *BMC Genomics* 18:654. doi: 10.1186/s12864-017-4062-2

Chen, T., Li, W., Hu, X., Guo, J., Liu, A., and Zhang, B. (2015). A cotton MYB transcription factor, GbMYB5, is positively involved in plant adaptive response to drought stress. *Plant Cell Physiol.* 56, 917–929. doi: 10.1093/pcp/pcv019

Chen, Y. (2013). Analyzing blends of herbivore-induced volatile organic compounds with factor analysis: revisiting "cotton plant, *Gossypium hirsutum* L., defense in response to nitrogen fertilization*"*. *J. Econ. Entomol.* 106, 1053–1057. doi: 10.1603/EC12178

Chen, Y., Liu, Z., Feng, L., Zheng, Y., Li, D., and Li, X. (2013). Genome-wide functional analysis of cotton (*Gossypium hirsutum*) in response to drought. *PLoS One* 8:e80879. doi: 10.1371/journal.pone.0080879

Du, L., Cai, C., Wu, S., Zhang, F., Hou, S., and Guo, W. (2016). Evaluation and exploration of favorable qtl alleles for salt stress related traits in cotton. *PLoS One* 11:e0151076. doi: 10.1371/journal.pone.0151076

Fang, D., Jenkins, J., Deng, D., McCarty, J., Li, P., and Wu, J. (2014). Quantitative trait loci analysis of fiber quality traits using a random-mated recombinant inbred population in Upland cotton (*Gossypium hirsutum* L.). *BMC Genomics* 15:397. doi: 10.1186/1471-2164-15-397

Fang, L., Wang, Q., Hu, Y., Jia, Y., Chen, J., Liu, B., et al. (2017). Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* 49, 1089–1098. doi: 10.1038/ng.3887

Fang, Y., and Xiong, L. (2015). General mechanisms of drought response and their application in drought resistance improvement in plants. *Cell Mol. Life Sci.* 72, 673–689. doi: 10.1007/s00018-014-1767-0

Gunapati, S., Naresh, R., Ranjan, S., Nigam, D., Hans, A., Verma, P. C., et al. (2016). Expression of *GhNAC2* from *G. herbaceum*, improves root growth and imparts tolerance to drought in transgenic cotton and Arabidopsis. *Sci. Rep.* 6:24978. doi: 10.1038/srep24978

Gutierrez, L., Mauriat, M., Guénin, S., Pelloux, J., Lefebvre, J. F., Louvet, R., et al. (2008). The lack of a systematic validation of reference genes: a serious pitfall undervalued in reverse transcription-polymerase chain reaction (RT-PCR) analysis in plants. *Plant Biotechnol. J.* 6, 609–618. doi: 10.1111/j.1467-7652.2008.00346.x

Héricourt, F., Chefdor, F., Djeghdir, I., Larcher, M., Lafontaine, F., Courdavault, V., et al. (2016). Functional divergence of poplar histidine-aspartate kinase HK1 paralogs in response to osmotic stress. *Int. J. Mol. Sci.* 17:2061. doi: 10.3390/ijms17122061

Hoagland, D., and Arnon, D. (1950). The water culture method for growing plants without soil. *Calif. Agric. Exp. Stn. Circ.* 347, 357–359.

Hodges, D., Forney, C., Prange, R., and DeLong, J. (1999). Improving the thiobarbituric acid-reactive-substances assay for estimating lipid peroxidation in plant tissues containing anthocyanin and other interfering compounds. *Planta* 207, 604–611. doi: 10.1017/S0031182002002585

Huang, C., Nie, X., Shen, C., You, C., Li, W., Zhao, W., et al. (2017). Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genomewide association study using high-density SNPs. *Plant Biotechnol. J.* 15, 1374–1386. doi: 10.1111/pbi.12722

Huang, X., Kurata, N., Wei, X., Wang, Z. X., Wang, A., Zhao, Q., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490, 497–501. doi: 10.1038/nature11532

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y. et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967. doi: 10.1038/ng.695

Hulse-Kemp, A. M., Lemm, J., Plieske, J., Ashrafi, H., Buyyarapu, R., Fang, D. D., et al. (2015). Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium* spp. *G3* 5, 1187–1209. doi: 10.1534/g3.115.018416

Kiranmai, K., Lokanadha Rao, G., Pandurangaiah, M., Nareshkumar, A., Amaranatha, Reddy, V., Lokesh, U., et al. (2018). A novel WRKY transcription factor, MuWRKY3 (*Macrotyloma uniflorum* Lam. Verdc.) enhances drought stress tolerance in transgenic Groundnut (*Arachis hypogaea* L.) Plants. *Front. Plant Sci.* 9:346. doi: 10.3389/fpls.2018.00346

Levi, A., Paterson, A., Barak, V., Yakir, D., Wang, B., Chee, P., et al. (2009). Field evaluation of cotton near-isogenic lines introgressed with QTLs for productivity and drought related traits. *Mol. Breed.* 23, 179–195. doi: 10.1007/s11032-008-9224-0

Li, C., Wang, X., Dong, N., Zhao, H., Xia, Z., Wang, R., et al. (2013). QTL analysis for early-maturing traits in cotton using two upland cotton (*Gossypium hirsutum* L.) crosses. *Breed. Sci.* 63, 154–163. doi: 10.1270/jsbbs.63.154

Li, H., Peng, Z. Y., Yang, X. H., Wang, W. D., Fu, J. J., Wang, J. H., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45, 43–50. doi: 10.1038/ng.2484

Li, D., Dossa, K., Zhang, Y., Wei, X., Wang, L., Zhang, Y., et al. (2018). GWAS uncovers differential genetic bases for drought and salt tolerances in sesame at the germination stage. *Genes* 9:87 doi: 10.3390/genes9020087

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., et al. (2015). Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* 33, 524–530. doi: 10.1038/nbt.3208

Li, F., Li, M., Wang, P., Cox, K. L. Jr., Duan, L., Dever, J. K., et al. (2017). Regulation of cotton (*Gossypium hirsutum*) drought responses by mitogen-activated protein (MAP) kinase cascade-mediated phosphorylation of GhWRKY59. *New Phytol.* 215, 1462–1475. doi: 10.1111/nph.14680

Liu, T., Longhurst, A. D., Talavera-Rauh, F., Hokin, S. A., and Barton, M. K. (2016). The Arabidopsis transcription factor ABIG1 relays ABA signaled growth inhibition and drought induced senescence. *eLife* 5:e13768. doi: 10.7554/eLife.13768.

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Ma, H., Chen, J., Zhang, Z., Ma, L., Yang, Z., Zhang, Q., et al. (2017). MAPK kinase 10.2 promotes disease resistance and drought tolerance by activating different MAPKs in rice. *Plant J.* 92, 557–570. doi: 10.1111/tpj.13674

Ma, Z. Y., He, S. P., Wang, X. F., Sun, J. L., Zhang, Y., Zhang, G. Y., et al. (2018). Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* 50, 803–813. doi: 10.1038/s41588-018-0119-7

Macho, R. M., Herrera, R. M., Brejcha, R., Schäffner, A. R., Tanaka, N., Fujiwara, T., et al. (2018). Boron toxicity reduces water transport from root to shoot in Arabidopsis plants. Evidence for a reduced transpiration rate and expression of major pip aquaporin genes. *Plant Cell Physiol.* 59, 836–844. doi: 10.1093/pcp/pcy026

Mackay, I., and Powell, W. (2007). Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci.* 12, 57–63. doi: 10.1016/j.tplants.2006.12.001

Magwanga, R. O., Lu, P., Kirungu, J. N., Lu, H., Wang, X., Cai, X., et al. (2018). Characterization of the late embryogenesis abundant (LEA) proteins family and their role in drought stress tolerance in upland cotton. *BMC Genet.* 19:6. doi: 10.1186/s12863-017-0596-1

Maurel, C., Verdoucq, L., Luu, D., and Santoni, V. (2008). Plant aquaporins: membrane channels with multiple integrated functions. *Annu. Rev. Plant Biol.* 59, 595–624. doi: 10.1146/annurev.arplant.59.032607.092734

Morrell, P. L., Toleno, D. M., Lundy, K. E., and Clegg, M. T. (2005). Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2442–2447. doi: 10.1073/pnas.0409804102

Negi, S., Tak, H., and Ganapathi, T. R. (2018). A banana NAC transcription factor (MusaSNAC1) impart drought tolerance by modulating stomatal closure and H2O2 content. *Plant Mol. Biol.* 96, 457–471. doi: 10.1007/s11103-018-0710-4

Odjegba, V., and Fasidi, I. (2006). Effects of heavy metals on some proximate composition of *Eichhornia crassipes*. *J. Appl. Sci. Environ. Mgt.* 10, 83–87. doi: 10.4314/jasem.v10i1.17309

Pantalião, G., Narciso, M., Guimarães, C., Castro, A., Colombari, J., Breseghello, F., et al. (2016). Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. *Genetica* 144, 651–664. doi: 10.1007/s10709-016-9932-z

Patel, R. K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 7:e30619. doi: 10.1371/journal.pone.0030619

Paterson, A. H., Brubaker, C. L., and Wendel, J. F. (1993). A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. *Plant Mol. Biol. Rep.* 11, 122–127. doi: 10.1007/BF02670470

Pettigrew, W. T. (2004). Physiological consequences of moisture deficit stress in cotton. *Crop Sci.* 44, 1265–1272. doi: 10.2135/cropsci2004.1265

Provenzano, M., and Mocellin, S. (2007). Complementary techniques: validation of gene expression data by quantitative real time PCR. *Adv. Exp. Med. Biol.* 593, 66–73. doi: 10.1007/978-0-387-39978-2_7

Ranjan, A., and Sawant, S. (2015). Genome-wide transcriptomic comparison of cotton (*Gossypium herbaceum*) leaf and root under drought stress. *3Biotech* 5, 585–596. doi: 10.1007/s13205-014-0257-2

Said, J. I., Lin, Z., Zhang, X., Song, M., and Zhang, J. (2013). A comprehensive meta QTL analysis for fiber quality, yield, yield related and morphological traits, drought tolerance, and disease resistance in tetraploid cotton. *BMC Genomics* 14:776. doi: 10.1186/1471-2164-14-776

Saleem, M., Malik, T., Shakeel, A., and Ashraf, M. (2015). QTL mapping for some important drought tolerant traits in upland cotton. *J. Anim. Plant Sci.* 25, 502–509.

Samota, M. K., Sasi, M., Awana, M., Yadav, O. P., Amitha Mithra, S. V., Tyagi, A., et al. (2017). Elicitor-induced biochemical and molecular manifestations to improve drought tolerance in rice (*Oryza Sativa* L.) through seed-priming. *Front. Plant Sci.* 8:934. doi: 10.3389/fpls.2017.00934

Saranga, Y., Menz, M., Jiang, C., Wright, R., Yakir, D., and Paterson, A. (2001). Genomic dissection of genotype x environment interactions conferring adaptation of cotton to arid conditions. *Genom. Res.* 11, 1988–1995. doi: 10.1101/gr.157201

Schweighofer, A., Hirt, H., and Meskiene, I. (2004). Plant PP2C phosphatases: emerging functions in stress signaling. *Trends Plant Sci.* 9, 236–243. doi: 10.1016/j.tplants.2004.03.007

Song, C., Chung, W., and Lim, C. (2016). Overexpression of heat shock factor gene HsfA3 increases galactinol levels and oxidative stress tolerance in Arabidopsis. *Mol. Cells* 39, 477–483. doi: 10.14348/molcells.2016.0027

Song, X., Zhang, Y., Wu, F., and Zhang, L. (2016). Genome-wide analysis of the maize (*Zea may* L.) CPP-like gene family and expression profiling under abiotic stress. *Genet. Mol. Res.* 15:gmr.15038023. doi: 10.4238/gmr.15038023

Sun, Z., Wang, X., Liu, Z., Gu, Q., Zhang, Y., Li, Z., et al. (2017). Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in *Gossypium hirsutum* L. *Plant Biotechnol. J.* 15, 982–996. doi: 10.1111/pbi.12693

Thirunavukkarasu, N., Hossain, F., Arora, K., Sharma, R., Shiriga, K., Mittal, S., et al. (2014). Functional mechanisms of drought tolerance in subtropical maize (*Zea mays* L.) identified using genome-wide association mapping. *BMC Genomics* 15:1182. doi: 10.1186/1471-2164-15-1182

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Wang, M., Li, C., and Wang, Q. (2014). Quantitative trait loci mapping and genetic dissection for lint percentage in upland cotton (*Gossypium hirsutum*). *J. Genet.* 93, 371–378. doi: 10.1007/s12041-014-0385-9

Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q., et al. (2017). Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.* 49, 579–587. doi: 10.1038/ng.3807

Wang, N., Liu, Y., Cong, Y., Wang, T., Zhong, X., Yang, S., et al. (2016). Genome-wide identification of soybean U-box E3 ubiquitin ligases and roles of GmPUB8 in negative regulation of drought stress response in Arabidopsis. *Plant Cell Physiol.* 28, 72–88. doi: 10.1093/pcp/pcw068

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wei, Q., Zhang, F., Sun, F., Luo, Q., Wang, R., Hu, R., et al. (2017). A wheat MYB transcriptional repressor TaMyb1D regulates phenylpropanoid metabolism and enhances tolerance to drought and oxidative stresses in transgenic tobacco plants. *Plant Sci.* 265, 112–123. doi: 10.1016/j.plantsci.2017.09.020

Wilkinson, S., and Davies, W. (2002). ABA-based chemical signaling: the co-ordination of responses to stress in plants. *Plant Cell Environ.* 25, 195-210. doi: 10.1046/j.0016-8025.2001.00824.x

Yan, H., Jia, H., Chen, X., Hao, L., An, H., and Guo, X. (2014). The cotton WRKY transcription factor GhWRKY17 functions in drought and salt stress in transgenic *Nicotiana benthamiana* through ABA signaling and the modulation of reactive oxygen species production. *Plant Cell Physiol.* 55, 2060–2076. doi: 10.1093/pcp/pcu133

Yang, N., Lu, Y. L., Yang, X. H., Huang, J., Zhou, Y., Ali, F., et al. (2014). Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. *PLoS Genet.* 10:e1004573. doi: 10.1371/journal.pgen.1004573

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., et al. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet.* 48, 927–934. doi: 10.1038/ng.3596

Zhai, H., Bai, X., Zhu, Y., Li, Y., Cai, H., Ji, W., et al. (2010). A single-repeat R3-MYB transcription factor MYBC1 negatively regulates freezing tolerance in Arabidopsis. *Biochem. Biophys. Res. Commun.* 394, 1018–1023. doi: 10.1016/j.bbrc.2010.03.114

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537. doi: 10.1038/nbt.3207

Zhang, J., Singh, A., Mueller, D. S., and Singh, A. K. (2015). Genome-wide association and epistasis studies unravel the genetic architecture of sudden death syndrome resistance in soybean. *Plant J.* 84, 1124–1136. doi: 10.1111/tpj.13069

Zheng, J., Oluoch, G., Riaz Khan, M. K., Wang, X., Cai, X., Zhou, Z., et al. (2016). Mapping QTLs for drought tolerance in an F2:3 population from an inter-specific cross between *Gossypium tomentosum* and *Gossypium hirsutum*. *Genet. Mol. Res.* 15 gmr.15038477. doi: 10.4238/gmr.15038477.

Zheng, M., Meng, Y., Yang, C., Zhou, Z., Wang, Y., and Chen, B. (2014). Protein expression changes during cotton fiber elongation in response to drought stress and recovery. *Proteomics* 14, 1776–1795. doi: 10.1002/pmic.201300123

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33, 408–414. doi: 10.1038/nbt.3096

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer ZL and handling Editor declared their shared affiliation at time of review.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 05 September 2018
doi: 10.3389/fpls.2018.01311

Genome-Wide Association Mapping of Starch Pasting Properties in Maize Using Single-Locus and Multi-Locus Models

Yang Xu†, *Tiantian Yang*†, *Yao Zhou, Shuangyi Yin, Pengcheng Li, Jun Liu, Shuhui Xu, Zefeng Yang* and Chenwu Xu**

Key Laboratory of Crop Genetics and Physiology of Jiangsu Province, Key Laboratory of Plant Functional Genomics of Ministry of Education, Co-Innovation Center for Modern Production Technology of Grain Crops, Yangzhou University, Yangzhou, China

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Julong Wei,
Nanjing Agricultural University, China
Huihui Li,
Institute of Crop Sciences (CAAS),
China
Yanru Cui,
University of California, Riverside,
United States

***Correspondence:**
Zefeng Yang
zfyang@yzu.edu.cn
Chenwu Xu
cwxu@yzu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: 13 June 2018
Accepted: 20 August 2018
Published: 05 September 2018

Citation:
Xu Y, Yang T, Zhou Y, Yin S, Li P, Liu J, Xu S, Yang Z and Xu C (2018) Genome-Wide Association Mapping of Starch Pasting Properties in Maize Using Single-Locus and Multi-Locus Models. Front. Plant Sci. 9:1311. doi: 10.3389/fpls.2018.01311

Maize starch plays a critical role in food processing and industrial application. The pasting properties, the most important starch characteristics, have enormous influence on fabrication property, flavor characteristics, storage, cooking, and baking. Understanding the genetic basis of starch pasting properties will be beneficial for manipulation of starch properties for a given purpose. Genome-wide association studies (GWAS) are becoming a powerful tool for dissecting the complex traits. Here, we carried out GWAS for seven pasting properties of maize starch with a panel of 230 inbred lines and 145,232 SNPs using one single-locus method, genome-wide efficient mixed model association (GEMMA), and three multi-locus methods, FASTmrEMMA, FarmCPU, and LASSO. We totally identified 60 quantitative trait nucleotides (QTNs) for starch pasting properties with these four GWAS methods. FASTmrEMMA detected the most QTNs (29), followed by FarmCPU (19) and LASSO (12), GEMMA detected the least QTNs (7). Of these QTNs, seven QTNs were identified by more than one method simultaneously. We further investigated locations of these significantly associated QTNs for possible candidate genes. These candidate genes and significant QTNs provide the guidance for further understanding of molecular mechanisms of starch pasting properties. We also compared the statistical powers and Type I errors of the four GWAS methods using Monte Carlo simulations. The results suggest that the multi-locus method is more powerful than the single-locus method and a combination of these multi-locus methods could help improve the detection power of GWAS.

Keywords: maize, starch, pasting properties, GWAS, multi-locus

INTRODUCTION

Maize (*Zea mays* L.) is the world's most important crop for food, feed and industrial materials. Starch is the principal constituent of maize kernels, which accounts for approximately 70% of the kernel weight (Liu N. et al., 2016). Benefitting from its characteristics such as slow tendency of retrogradation and low pasting temperature (PTP), maize starch serves as an essential ingredient for industrial production of food, and has been widely used to thicken sauces or soups and make corn syrup and other sugars (Yang Z. et al., 2014). Recently, great progress has been made in dissection of starch content in maize kernels (Wang et al., 2015; Li et al., 2018). However, further improvements in starch quality are needed to meet demands of food processing and industrial application. The pasting properties are important characteristics of starch, determining the starch quality and functionality. Dissection the genetic basis of pasting properties will facilitate the improvement of starch quality in maize.

Genome-wide association studies (GWAS) provide the opportunity to decipher genetic architectures of complex traits in crops (Zhu et al., 2008). Owing to the rapid linkage disequilibrium (LD) decay and abundant diversity, maize is an ideal species to perform GWAS. GWAS have successfully analyzed many important traits, such as kernel oil biosynthesis, plant height and disease resistance in maize (Kump et al., 2011; Li et al., 2013). Some statistical models have been developed to conduct GWAS. Mixed linear model (MLM) has become the most popular approach with the ability to consider population structure and family relatedness (Zhang et al., 2005; Yu et al., 2006). Based on the MLM framework, some single-locus approaches have been proposed to alleviate the heavy computational burden, such as EMMAX (Kang et al., 2008), P3D (Zhang et al., 2010), FaST-LMM (Lippert et al., 2011), and genome-wide efficient mixed model association (GEMMA) (Zhou and Stephens, 2012). However, the single-locus model testing one locus at a time fails to match the true genetic model of complex traits that are controlled by numerous loci simultaneously. Additionally, multiple test corrections for critical values are usually required to control false positive rates for single-locus GWAS. The commonly used Bonferroni correction is so conservative that lots of true loci may be neglected. To overcome these problems, multi-locus GWAS methods have been recommended because these methods consider the information of all loci simultaneously and multiple test corrections are not required because of the multi-locus nature (Wang et al., 2016). Several multi-locus methods, such as FASTmrEMMA, ISIS EM-BLASSO, FASTmrMLM, pLARmEB, pKWmEB, LASSO, and FarmCPU, have been proved to be more powerful than single-locus methods (Liu X.L. et al., 2016; Tamba et al., 2017; Xu et al., 2017; Zhang et al., 2017; Ren et al., 2018; Wen et al., 2018).

There have been a few studies focusing on the genetic basis of pasting properties in maize starch. Zhang et al. (2004) suggested that *SSIIa* of maize affected the starch structure and physiochemical properties. Wilson et al. (2004) used association mapping to evaluate six candidate genes involved in starch synthesis and found that *ae1* and *sh*2 were associated with starch pasting properties. Xu et al. (2014a) detected the associations of sequence variants of the *ZmBT1* gene with seven pasting properties. Yang Z. et al. (2014) identified seven quantitative trait nucleotides (QTNs) in coding regions of *Zmisa2* underlying pasting properties of maize starch and proposed that these markers may be potentially utilized for marker-assisted selection. However, all of the above studies were based on specific candidate genes involved in kernel starch biosynthesis. Therefore, more comprehensive studies are required to further understand the molecular mechanisms of starch pasting properties. To our knowledge, GWAS for pasting properties of maize starch have not been reported up to now.

In this study, a worldwide collection of 230 inbred lines were genotyped with 145,232 SNPs using genotyping-by-sequencing (GBS) technology. Starch pasting properties including peak viscosity (PV), trough viscosity (TV), final viscosity (FV), breakdown viscosity (BD), setback viscosity (SB), pasting time (PT), and PTP were measured for the 230 lines using the Rapid Visco Analyser (RVA). The main objectives of this study were to (i) identify loci that are significantly associated with pasting properties of maize starch using single-locus and multi-locus GWAS methods, and (ii) compare three multi-locus methods (FASTmrEMMA, LASSO, and FarmCPU) with one single-locus method (GEMMA) in terms of their detection powers and Type I errors.

MATERIALS AND METHODS

Plant Materials

In this study, an association panel of 230 maize lines collected from the tropical, subtropical or temperate zone, representing a wide range of diversity, was used for GWAS. All the materials were planted with a randomized block design of three repetitions in the field of Sanya, Hainan province. At the four-leaf stage, young leave tissues were collected from each line and preserved at −80°C. DNA was extracted from the freeze-dried leave tissues with the modified CTAB method (Fulton et al., 1995). After harvest, mature kernels of five randomly selected plants in each line were collected and used for evaluation of starch pasting properties.

Genotyping

The panel of 230 maize inbred lines was genotyped using a GBS strategy. The ApeK1 restriction enzyme was used for library preparation, and GBS was performed on an Illumina platform by Novogene Bioinformatics Institute, Beijing, China. After quality control, a total of 145,232 high-quality SNPs with minor allele frequency (MAF) above 2% and missing rate below 20% remained to perform GWAS.

Measurement of Starch Pasting Properties

The pasting properties of maize starch were evaluated using RVA (Model 3D, Newport Scientific, Sydney, NSW, Australia). Three grams of starch obtained from each line was mixed with 25 ml of distilled water in the RVA canister. The RVA profile took a heat–hold–cool temperature cycle as follows: (1) set at 50°C as the starting temperature and maintained for 1 min; (2) heated to 95°C and held at 95°C for 2.5 min; and (3) cooled to 50°C and kept at 50°C for 1.4 min. The total processing time was about 12 min. The pasting properties were determined using a fixed paddle rotation at the speed of 160 r/m. The RVA parameters were recorded in centipose (cP). The pasting parameters obtained from the pasting curve including PV, TV, FV, PTP, PT and their derived parameters, BD and SB were recorded for all the inbred lines. The average value of three biological replicates from each line was obtained for data analysis.

Genome-Wide Association Analysis

In this study, GWAS were performed in the association panel composed of 230 diverse maize inbred lines with 145,232 high-quality SNPs. The decay distance of LD across the whole genome was determined by software PopLDdecay[1]. Principle component analysis (PCA) was used to control for population structure. Both single-locus and multi-locus methods were used to identify significant QTNs for seven starch properties. GEMMA was used for single-locus GWAS, and FASTmrEMMA (Wen et al., 2018), LASSO (Xu et al., 2017), and FarmCPU (Liu X.L. et al., 2016) were used for multi-locus GWAS. GEMMA was developed based on the framework of MLM, which takes advantage of eigen decomposition to substantially increase the computational speed. GEMMA was implemented in the software GEMMA. FASTmrEMMA is a multi-locus two-stage GWAS method, combining the MLM and the expectation and maximization empirical Bayes (EMEB) method. In the first stage, the marker effects were treated as random and a small number of markers were selected, and then in the second stage, the selected markers were fitted into a multi-locus model and estimated using the EMEB method. FASTmrEMMA was implemented in the R package mrMLM. LASSO is a powerful multi-locus approach, but it lacks a default method to perform a significance test. Here, we used our previously proposed Bayesian algorithm to approximately estimate the variance of each marker effect and then used a Wald test to obtain the significant test for each marker. Details about this algorithm were given in Xu et al. (2017). The LASSO method was implemented in the R package glmnet and our own R program. The FarmCPU method is a commonly used GWAS method at present, which effectively eliminates confounding and improves statistical power for MLM methods by using the fixed effect model and random effect model iteratively. FarmCPU was implemented in the R package FarmCPU. All parameters were set at default values. The significantly associated QTNs were determined by the LOD value exceeding three for FASTmrEMMA and LASSO, and the *P*-value less than $1/m$ (m is the number of markers) for GEMMA and FarmCPU. To mine candidate genes based on the detected QTNs for the pasting properties, we used gene annotation and ontology information available in maizeGDB[2] and Phytozome database[3].

Simulation Experiments

To investigate the powers and Type I errors of the single-locus and multi-locus GWAS methods, we carried out a Monte Carlo simulation experiment using the genotypic data of 230 maize inbred lines. We assigned eight QTL located on the first eight chromosomes. The assigned QTL totally explained 56% of the phenotypic variation. The detailed description of the eight QTL is presented in **Table 1**. Both the polygenic variance and residual error variance were set at one. The population structure effect was added according to the first five principal components determined from the genotypic data. These principal components contributed to 10% of the phenotypic variance. The phenotype was simulated with the contribution of the genetic effect of simulated QTL, polygenic effect, residual effect, and population structure effect. The simulations were replicated 200 times and the four GWAS methods, FASTmrEMMA, FarmCPU, LASSO, and GEMMA, had been used to analyze the simulated data. The statistical power for a simulated QTL was defined as the fraction of the 200 replicates where the LOD score of the QTL was larger than three for the FASTmrEMMA and LASSO methods and the *P*-value of the QTL was less than $1/m$ for GEMMA and FarmCPU. Type I error was defined as the ratio of false positives out of all markers not assigned a QTL effect. Each QTL within 1 kb of the assigned QTL was counted as a real QTL.

RESULTS

Phenotypic Variations and Heritability

The descriptive statistics of the seven pasting properties for the 230 maize inbred lines are listed in **Table 2**. The average values for PV, TV, BD, FV, SB, PT, and PTP are 1,200.22, 1,004.04, 196.18, 1,980.36, 976.33, 5.46, and 81.24 with the standard deviations 334.56, 225.29, 141.77, 427.17, 314.40, 0.42, and 2.14, respectively. Substantial variations among genotypes are observed for the seven pasting properties, and pasting properties vary significantly among different lines. Also, variance components were estimated using the restricted maximum likelihood (REML) analysis (Xu et al., 2014b). The narrow sense heritability, defined as the ratio of additive genetic variance to total phenotypic variance, ranges from 0.46 for PT to 0.77 for TV (**Table 2**). These results indicate that the phenotypic variations of starch pasting properties are mainly affected by genetic factors, and therefore this panel can be used for further genetic analyses. To determine the correlation among the seven pasting properties, the Pearson's correlation coefficients were calculated. The results of the correlation analysis are illustrated in **Figure 1**. All the pairwise correlations between any two pasting properties exhibit significantly positive or negative correlations except three correlations between PT and TV, between PT and SB, and between SB and PTP.

TABLE 1 | Information for the eight simulated QTL.

QTL	Chromosome	Position (bp)	MAF	Effect	$R^{2,a}$ (%)
QTL1	1	14898058	0.335	0.569	4
QTL2	2	19326559	0.16	0.842	4
QTL3	3	20532172	0.307	1.041	6
QTL4	4	13181343	0.452	0.667	6
QTL5	5	15819352	0.204	0.935	8
QTL6	6	27154881	0.378	0.766	8
QTL7	7	16672999	0.085	1.564	10
QTL8	8	22685122	0.217	1.028	10

[a] *Proportion of the total phenotypic variation explained by the QTL.*

[1] https://github.com/BGI-shenzhen/PopLDdecay

[2] http://maizegdb.org

[3] https://phytozome.jgi.doe.gov

TABLE 2 | Phenotypic performance, variance component, and heritability of seven pasting properties of maize starch.

	Mean ± SD	Range	Genetic variance	Residual variance	Heritability	*F* value
PV (cP)	1,200.22 ± 334.56	494.5–2,272	115,459.13	50,143.19	0.70	2.11**
TV (cP)	1,004.04 ± 225.29	463.5–1,737	63,071.08	18,425.03	0.77	2.28**
BD (cP)	196.18 ± 141.77	2.5–783.5	15,269.62	11,891.91	0.56	1.96**
FV (cP)	1980.36 ± 427.17	920–3411	155519.10	99828.04	0.61	2.12**
SB (cP)	976.33 ± 314.40	319–1930	108176.00	41637.66	0.72	2.40**
PT (min)	5.46 ± 0.42	4.6–7	0.11	0.13	0.46	3.26**
PTP (°C)	81.24 ± 2.14	75.55–87.28	3.40	2.86	0.54	2.71**

**Indicates significance level at *P* < 0.01. *PV, peak viscosity; TV, trough viscosity; BD, breakdown viscosity; FV, final viscosity; SB, setback viscosity; PT, pasting time; PTP, pasting temperature.*

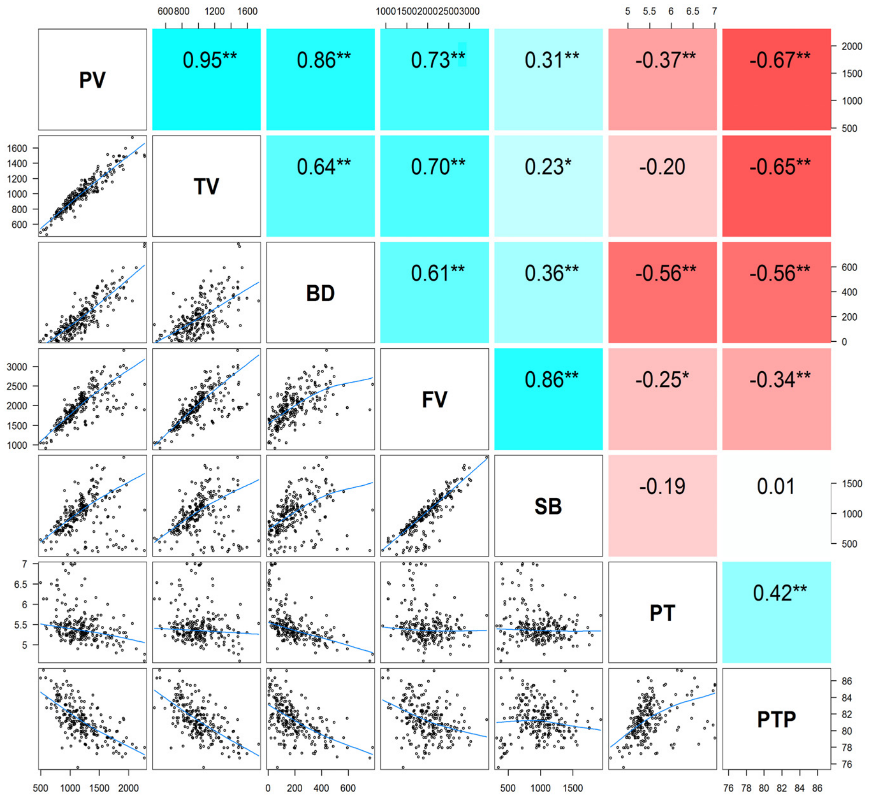


FIGURE 1 | The pairwise correlation analysis among seven pasting properties of maize starch. *Upper diagonal*: Pearson correlation coefficients between every two traits; *Lower diagonal*: Scatter plots of correlations between every two traits. *Asterisk* (*) indicates significance level at *P* < 0.05; *Double asterisks* (**) indicates significance level at *P* < 0.01. PV, peak viscosity; TV, trough viscosity; BD, breakdown viscosity; FV, final viscosity; SB, setback viscosity; PT, pasting time; PTP, pasting temperature.

Population Structure and Linkage Disequilibrium

In this study, PCA was used to correct for population structure. PCA plots of this association population are illustrated in **Figure 2**. According to the scree plot, the variance of principle component score decreases quickly until the fifth principle component (**Figure 2B**). Therefore, we selected the first five principal components to control the population structure. All filtered SNPs were used to determine LD decay. A monotonic decrease in LD is found with increasing distance (**Figure 3**). At r^2 = 0.2, the overall LD decay decreases dramatically to 10 kb. The genome-wide LD decay distance is about 250 kb at the cut-off of r^2 = 0.1.

GWAS for Starch Pasting Properties

In this study, GWAS were conducted for 230 maize inbred lines with 145,232 SNPs using four methods and the results are listed in **Table 3**. A total of 60 significant QTNs are identified for seven starch properties from the four GWAS methods. FASTmrEMMA detects the most QTNs (29), followed by FarmCPU (19) and LASSO (12), GEMMA detected the least QTNs (7). The numbers of significant QTNs detected for starch

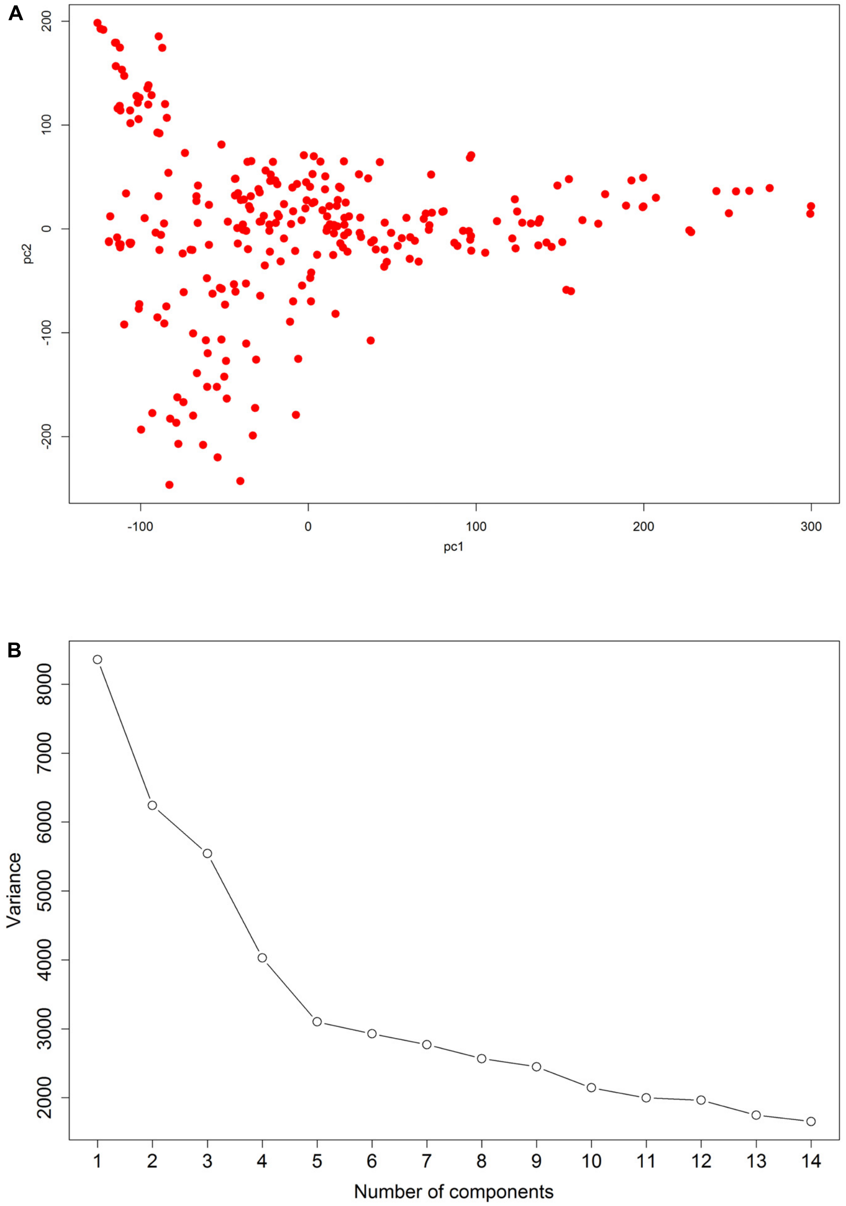


FIGURE 2 | Genetic structure of maize inbred lines. **(A)** Plot of the first two principal components of 230 inbred lines. **(B)** Scree plot showing the selection of principal components for GWAS.

properties PV, TV, BD, FV, SB, PT, and PTP are 14, 10, 8, 12, 12, 6, and 5, respectively, from all the four methods. The corresponding numbers of the significant QTNs are 8, 6, 5, 3, 6, 1, and 4 from FASTmrEMMA; 2, 4, 2, 7, 4, 2, and 1 from FarmCPU; 2, 2, 0, 1, 3, 3, and 1 from LASSO; and 2, 0, 1, 1, 1, 2, and 0 from GEMMA. The largest QTN detected by FASTmrEMMA, FarmCPU, LASSO, and GEMMA explains 9.35, 14.96, 1.03, and 12.03(%) of the phenotypic variation, respectively. Among these significant QTNs, seven QTNs appear to control more than one trait (pleiotropic effect). For example, three QTNs (SNP_2_190495578, SNP_9_138239683, and SNP_4_89429269) have significant effects on PV and TV. Both SNP_6_109456130 and SNP_7_160060597 are associated with PV and BD. The correlations between PV and TV as well as between PV and BD are significant.

When comparing the results across different methods, only seven common QTNs are identified by more than one method simultaneously. Among these QTNs, SNP_2_9506602 is detected across three GWAS methods (FarmCPU, LASSO, and GEMMA); SNP_9_103241537 and one pleiotropic QTN (SNP_4_89429269) are identified by FASTmrEMMA and FarmCPU simultaneously; SNP_4_144401228 is detected by FASTmrEMMA and LASSO; SNP_9_109684667 is detected by LASSO and GEMMA; SNP_3_12888452 is detected by FarmCPU

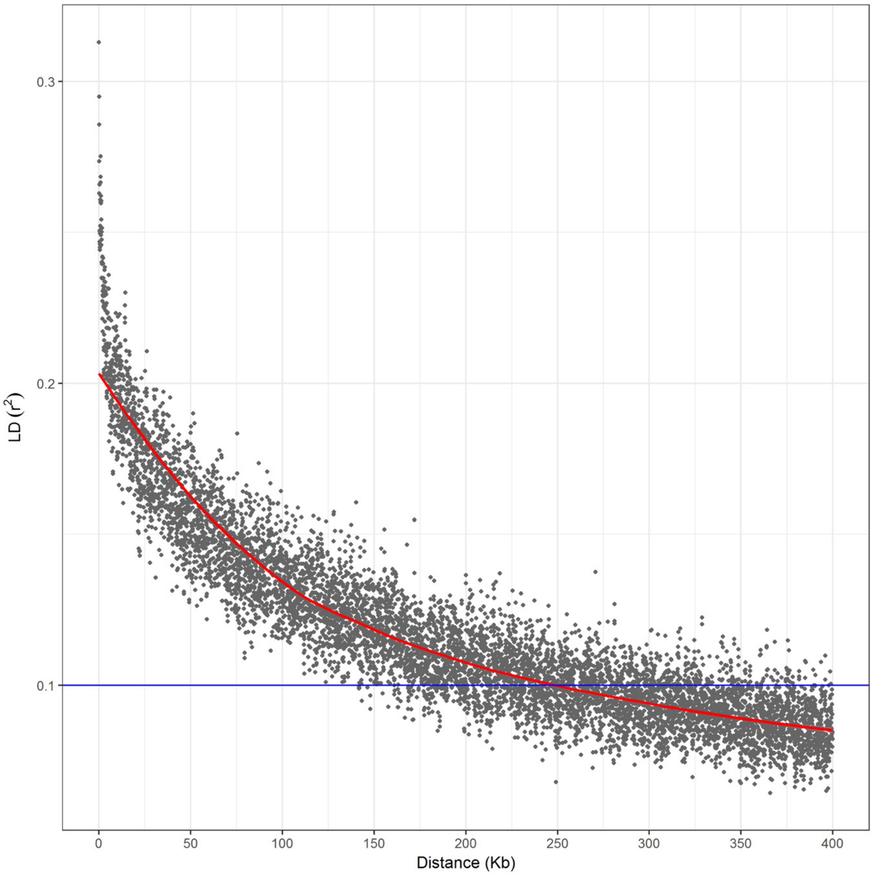


FIGURE 3 | Linkage disequilibrium decay across the whole genome of the association panel. The *blue horizontal line* shows the LD threshold for the association panel (r^2 = 0.1).

and LASSO. SNP_7_173235732 associated with SB and FV are detected by FASTmrEMMA and FarmCPU, respectively. Note that the estimated effects and R^2 values (proportion of phenotypic variance explained by the QTL) of the co-identified QTNs detected by different methods are completely different, whereas the signs of effects for these co-identified QTNs for the same trait are consistent. For example, the estimated effects of SNP_2_9506602 are −0.177, −0.057, and −0.244, and the corresponding R^2 values are 5.14, 0.53, and 9.74(%) for trait PT when using FarmCPU, LASSO, and GEMMA, respectively. All the three methods demonstrate that this QTN has the negative effect on PT.

Simulation Studies for GWAS

Simulation experiments were performed to compare the statistical powers and Type I errors of the four GWAS methods. The statistical powers of detecting the simulated QTL calculated based on 200 simulations are given in **Table 4**. The average powers for FASTmrEMMA, FarmCPU, LASSO, and GEMMA were 55.19, 43.31, 53.69, and 40.44(%), respectively, indicating the highest average power of FASTmrEMMA. However, different methods may be suitable for detection of different QTL. For example, FASTmrEMMA has the highest power for detecting QTL1, QTL4, QTL5, QTL6, and QTL8 but the lowest power for detecting QTL7. LASSO is the best method for detecting QTL2 and QTL3, whereas it is the worst method for detecting QTL1. GEMMA has the lowest power of detecting all the simulated QTL, but it is the most efficient method for detecting QTL7. Type I errors for all the four methods are also listed in **Table 4**. LASSO has the lowest Type I error, followed by GEMMA and FASTmrEMMA, and FarmCPU has the highest Type I error. The Type I errors of the four methods are under 0.0001 with the same order of magnitude. Overall, the Type I errors are well controlled for all the four approaches, and the three multi-locus approaches are more powerful than the single-locus approach.

DISCUSSION

In this study, we compared statistical powers of FASTmrEMMA, FarmCPU, LASSO, and GEMMA using real and simulation data. Simulation experiments based on the genotypic data of 230 maize inbred lines illustrate that the multi-locus approach is more powerful than single-locus approach in most cases, especially for loci with small effect that explain less than six percent of phenotypic variance. Although single-locus methods have been

TABLE 3 | Significantly associated QTNs identified by four GWAS methods for seven pasting properties of maize starch.

Trait	Marker	Alleles	Chr	Pos	FASTmrEMMA		FarmCPU		LASSO		GEMMA	
					Effect	R^2 (%)	Effect	R^2 (%)	Effect	R^2 (%)	Effect	R^2 (%)
PV	SNP_2_190495578#	C/A	2	190495578	256.70	7.05						
	SNP_2_42359599	A/G	2	42359599	−131.53	2.87						
	SNP_2_51001688	A/C	2	51001688	−191.54	5.03						
	SNP_3_171824570#	G/T	3	171824570	126.17	2.46						
	SNP_4_64845133	A/G	4	64845133	193.04	2.81						
	SNP_4_89429269#	G/A	4	89429269	134.34	3.09						
	SNP_5_26160368	T/A	5	26160368							−234.93	11.13
	SNP_5_26160478	C/A	5	26160478							−234.29	11.64
	SNP_6_109456130#	A/C	6	109456130			−116.84	7.12				
	SNP_6_164038368	C/A	6	164038368					−51.80	0.46		
	SNP_7_160060597#	A/G	7	160060597	174.52	2.79						
	SNP_8_147208913	C/T	8	147208913			−133.38	5.05				
	SNP_9_138239683#	C/A	9	138239683					−68.33	1.03		
	SNP_9_58569771	C/T	9	58569771	210.21	3.01						
TV	SNP_2_190495578#	C/A	2	190495578	148.11	5.18						
	SNP_2_75175274	A/G	2	75175274	−163.45	3.19						
	SNP_4_144401228*	A/G	4	144401228	−141.31	4.68			−30.63	0.88		
	SNP_4_89429269#*	G/A	4	89429269	103.79	4.07	54.80	4.56				
	SNP_5_168661067	G/C	5	168661067			−76.75	4.19				
	SNP_8_104430223	T/C	8	104430223	163.65	7.19						
	SNP_9_138239602	G/C	9	138239602			−102.16	4.94				
	SNP_9_138239683#	C/A	9	138239683					−44.03	0.94		
	SNP_10_12091187	A/G	10	12091187	−103.99	3.72						
	SNP_10_142948941	A/G	10	142948941			48.76	3.85				
BD	SNP_1_241610826	C/T	1	241610826			−33.08	4.33				
	SNP_1_825561	C/T	1	825561	93.47	3.56						
	SNP_4_146006182	G/A	4	146006182	73.78	3.74						
	SNP_6_109456130#	A/C	6	109456130			−58.28	9.86				
	SNP_7_160060597#	A/G	7	160060597	76.06	2.95						
	SNP_9_142242612	C/T	9	142242612	−49.54	2.53						
	SNP_10_138051694	G/C	10	138051694							82.38	10.28
	SNP_10_9143566	G/T	10	9143566	−83.04	5.87						
FV	SNP_1_283390691	T/C	1	283390691			106.83	5.53				
	SNP_2_51001706	C/T	2	51001706	−273.07	6.02						
	SNP_5_160490300	A/G	5	160490300			−142.99	5.16				
	SNP_5_160866262	T/C	5	160866262			167.30	6.97				
	SNP_5_213796937	A/G	5	213796937			−96.09	3.71				
	SNP_6_107223456	G/C	6	107223456							−206.80	12.03
	SNP_6_115373488	G/A	6	115373488	175.34	3.14						
	SNP_7_173235732#	T/G	7	173235732			130.23	5.04				
	SNP_8_124259102	A/C	8	124259102					−64.27	0.51		
	SNP_8_154309867	G/T	8	154309867			150.77	6.61				
	SNP_9_113510544	G/A	9	113510544	264.83	3.48						
	SNP_9_83760699	A/T	9	83760699			143.42	3.60				
SB	SNP_1_168229057	C/T	1	168229057					−40.57	0.44		
	SNP_2_27401698	G/T	2	27401698					−37.77	0.31		
	SNP_2_46177221	G/A	2	46177221	147.63	2.52						
	SNP_6_104663091	A/C	6	104663091			115.82	9.30				
	SNP_6_124651063	G/A	6	124651063			−63.76	3.81				
	SNP_6_158401136	G/C	6	158401136			−72.23	3.41				
	SNP_7_173235732#	T/G	7	173235732	186.23	4.73						
	SNP_7_48994000	A/G	7	48994000	−166.31	5.17						

(Continued)

TABLE 3 | Continued

Trait	Marker	Alleles	Chr	Pos	FASTmrEMMA		FarmCPU		LASSO		GEMMA	
					Effect	R^2 (%)	Effect	R^2 (%)	Effect	R^2 (%)	Effect	R^2 (%)
	SNP_8_38060255	T/C	8	38060255	−148.06	2.91						
	SNP_9_103241537*	G/C	9	103241537	213.43	9.35	81.24	5.44				
	SNP_9_109684667*	C/A	9	109684667					41.66	0.54	181.03	10.19
	SNP_10_143879663	G/A	10	143879663	163.20	4.53						
PT	SNP_2_79885513	T/C	2	79885513							0.31	10.34
	SNP_2_9506602*	T/C	2	9506602			−0.18	5.14	−0.06	0.53	−0.24	9.74
	SNP_3_219463585	T/A	3	219463585			−0.10	4.67				
	SNP_4_211011498	T/C	4	211011498					−0.07	0.54		
	SNP_5_59630329	G/C	5	59630329					−0.07	0.58		
	SNP_8_22655499	T/A	8	22655499	0.34	5.00						
PTP	SNP_2_80464203	C/T	2	80464203	1.36	2.24						
	SNP_3_12888452*	G/C	3	12888452			1.58	14.96	0.27	0.42		
	SNP_3_171824570#	G/T	3	171824570	−1.26	6.02						
	SNP_4_193530385	T/G	4	193530385	−1.36	3.71						
	SNP_4_71048778	A/G	4	71048778	−1.27	3.32						

*#Indicates the QTN identified across different traits. *Indicates the QTN identified across different methods.*

TABLE 4 | Statistical powers (%) of eight simulated QTL and Type I error rates for four GWAS methods drawn from 200 replicated simulation experiments.

QTL	FASTmrEMMA	FarmCPU	LASSO	GEMMA
QTL1	52.5	22.5	5.5	6
QTL2	20.5	15.5	39	0
QTL3	55	46	62	41
QTL4	47	11.5	9	2.5
QTL5	92	61	83.5	60
QTL6	58	57	51	50
QTL7	20.5	65	92.5	94
QTL8	96	68	87	70
Type I error	6.99E-05	7.17E-05	4.70E-05	6.58E-05

widely used to identify genetic variants in many crop species, they neglect the overall effects of multiple loci and suffer from the problem of multiple test corrections for critical values. Several investigators have compared statistic powers of multi-locus and single-locus methods and demonstrated that multi-locus methods perform better than single-locus methods. Wen et al. (2018) compared FASTmrEMMA with single-locus approaches including EMMA, SUPER, CMLM, and ECMLM using a series of simulation studies and found that FASTmrEMMA has the highest power and accuracy. Xu et al. (2017) showed that the multi-locus LASSO method has higher statistical power and lower Type I error than GEMMA. Liu X.L. et al. (2016) demonstrated that FarmCPU improves statistical power compared to GLM, MLM, CMLM, FaST-LMM-Select across multiple species, such as *Arabidopsis thaliana*, human and maize. In previous simulation studies, Bonferroni multiple test correction was used for single-locus method. However, it may be too strict to use Bonferroni correction ($0.05/m$) as the cut-off as not all loci are independent (Yang N. et al., 2014). To avoid missing the relevant loci, we replaced Bonferroni correction by a less stringent criterion ($1/m$) for GEMMA. The results of simulation showed that Type I error of GEMMA with $1/m$ as the cut-off was well controlled and similar to that of three multi-locus methods. Additionally, the permutation method is commonly used to adjust for multiple tests, which yields reliable outcome but requires a lot of time for huge samples (Churchill and Doerge, 1994). Fortunately, no multiple test correction is required for FASTmrEMMA and LASSO because all markers are fitted to a single model and all effects are estimated and tested simultaneously.

In the real data analysis, a total of 29, 19, 12, and 7 significant QTNs were identified for seven pasting properties of maize starch using FASTmrEMMA, FarmCPU, LASSO, and GEMMA, respectively. FASTmrEMMA detected the most QTNs, while GEMMA detected the least, which was consistent with the results of the simulation that FASTmrEMMA performed the best for detection of most QTL and GEMMA performed the worst. Unexpectedly, there was no significant QTN detected by these four methods simultaneously, and only seven QTNs were detected by more than one method. This situation could be explained by the simulation studies. From the simulation results, none of these methods were found to achieve very high power for detecting all the simulated QTL and different methods may be suitable for identification of different QTL. For example, FASTmrEMMA possessed good performance for most QTL, whereas it was not efficient for simulated QTL7 with the largest effect and lowest MAF. LASSO performed well for detecting large QTL but poorly for small QTL. Each method has its own advantages and limitations. LASSO is computationally efficient, but fails to handle a large number of markers. FASTmrEMMA is powerful in detection of QTL and accurate in effect estimation of QTL. However, FASTmrEMMA is a two-step combined method. The first step is to select a small fraction of makers and then apply these markers to perform multi-locus analyses in the second step. This method has an issue in how to determine the suitable thresholds in the first step. To improve the power of GWAS, it is better to use a combination of these methods, and the QTL detected by multiple methods may be more reliable. Recently, Zhang et al. (2018) and Ma et al. (2018) also proposed that using a combination of multiple multi-locus methods could improve the efficiency for detecting the QTL underlying lodging resistance-related and regeneration-related traits of maize. Genome-wide association studies have been applied to dissect genetic architectures of several complex traits in maize (Xiao et al., 2017). However, no previous studies have focused on GWAS for starch pasting properties in maize. Here, we performed GWAS for seven pasting properties in a panel of 230 maize inbred lines genotyped with 145,232 SNPs and identified 60 significant QTNs using single-locus and multi-locus GWAS methods. Notably, the detected loci may not be the real causative loci due to false positives caused by LD or population structure. To understand the molecular basis of pasting properties, we further investigated locations of associated QTNs for possible candidate genes. The candidate genes within 250 kb downstream and upstream of the identified QTNs and their orthologs in Arabidopsis and rice are presented in **Supplementary Table S1**. According to functional annotations, these candidate genes were primarily categorized as protein kinases, glycosyltransferases, glycosidases, hydrolases, and transcription factors. The transcription factors included E2F, BHLH, TFIIH, MYB, bZIP, and HSF superfamily. Some of the candidate genes or their homologous genes are known genes linked to starch biosynthesis. For example, GRMZM2G032628 (*ae1*) encodes starch branching enzyme, which is a downstream gene involved in the final product of starch biosynthesis (Dolezal et al., 2014). It was reported that *ae1* was significantly associated with pasting properties of maize starch (Wilson et al., 2004). The homologous gene *SUS3* of GRMZM2G392988 in Arabidopsis has been reported to be involved in starch biosynthesis within seed coat and embryo (Angeles-Nunez and Tiessen, 2010). Several candidate genes are annotated as glycosyltransferases, which formed the important catalytic mechanism to synthesize and break the glycosidic bonds in oligosaccharides, disaccharides, and polysaccharides (Li et al., 2018). To better understand the potential biological functions of these candidate genes, we performed the gene ontology (GO) analysis for these genes using clusterProfiler (Yu et al., 2012). The GO analysis revealed that these genes were significant enriched in 16 GO terms (*P*-value <0.01), which were classified into three main types containing biological process, molecular function, and cellular component (**Supplementary Figure S1**). Under the first type, the most significant GO terms are gluconeogenesis process and hexose biosynthetic process, which play important roles in starch biosynthesis. Under the second type, these genes were significant related to chorismate synthase activity and glucose-6-phosphate isomerase activity. Under the third type, several genes were involved in photosystem. We also found that some candidate genes were involved in multiple functions. For example, GRMZM2G065083 are involved in gluconeogenesis process, hexose biosynthetic and metabolic process and glucose-6-phosphate isomerase activity. However, these genes were not found to be known genes involved in starch biosynthesis pathway, indicating that our study of the molecular mechanisms underlying pasting properties of maize starch is incomplete. These identified QTNs and candidate genes provide foundation for further functional studies to dissect the genetic mechanism manipulating maize pasting properties.

CONCLUSION

In this study, single-locus and multi-locus GWAS methods were used to identify loci associated with starch pasting properties in maize. A total of 60 significant QTNs were detected for seven pasting properties, of which 29, 19, 12, and 7 QTNs were detected using FASTmrEMMA, FarmCPU, LASSO, and GEMMA, respectively. These QTNs could be utilized for further genetic and breeding studies to regulate starch pasting properties. Additionally, we compared four GWAS methods for their detection powers and Type I errors based on simulation studies and found that the multi-locus method is more powerful than the single-locus method and the combination of these multi-locus methods could help improve the statistical power of current GWAS.

AUTHOR CONTRIBUTIONS

CX and ZY designed the research plan. YX, TY, YZ, SY, PL, JL, and SX performed the experiments. YX and PL analyzed the data. YX wrote the paper. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Key Technology Research and Development Program of MOST (2016YFD0100303), the National Natural Science Foundation of China (31801028, 31601810, and 81700822), the China Postdoctoral Science Foundation (137070445), the Yangzhou University Innovation Cultivation Foundation (135030066), the Priority Academic Program Development of Jiangsu Higher Education Institutions, the Natural Science Foundations of Jiangsu Province (BK20150010), the Innovative Research Team of Ministry of Agriculture, and the Qing-Lan Project of Jiangsu Province.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01311/full#supplementary-material

FIGURE S1 | Distribution of significant GO terms (*P*-value < 0.01).

TABLE S1 | Candidate genes for QTNs significantly associated with starch pasting properties in maize and their homologous genes in Arabidopsis and rice.

REFERENCES

Angeles-Nunez, J. G., and Tiessen, A. (2010). Arabidopsis sucrose synthase 2 and 3 modulate metabolic homeostasis and direct carbon towards starch synthesis in developing seeds. *Planta* 232, 701–718. doi: 10.1007/s00425-010-1207-9

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963–971.

Dolezal, A. L., Shu, X., Obrian, G. R., Nielsen, D. M., Woloshuk, C. P., Boston, R. S., et al. (2014). *Aspergillus flavus* infection induces transcriptional and physical changes in developing maize kernels. *Front. Microbiol.* 5:384. doi: 10.3389/fmicb.2014.00384

Fulton, T. M., Chunwongse, J., and Tanksley, S. D. (1995). Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Mol. Biol. Report.* 13, 207–209.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Kump, K. L., Bradbury, P. J., Wisser, R. J., Buckler, E. S., Belcher, A. R., Oropeza-Rosas, M. A., et al. (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* 43, 163–168. doi: 10.1038/ng.747

Li, C., Huang, Y., Huang, R., Wu, Y., and Wang, W. (2018). The genetic architecture of amylose biosynthesis in maize kernel. *Plant Biotechnol. J.* 16, 688–695. doi: 10.1111/pbi.12821

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45, 43–50. doi: 10.1038/ng.2484

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* 8, 833–U894. doi: 10.1038/Nmeth.1681

Liu, N., Xue, Y., Guo, Z., Li, W., and Tang, J. (2016). Genome-wide association study identifies candidate genes for starch content regulation in maize kernels. *Front. Plant Sci.* 7:1046. doi: 10.3389/fpls.2016.01046

Liu, X. L., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. W. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12:e1005767. doi: 10.1371/journal.pgen.1005767

Ma, L., Liu, M., Yan, Y., Qing, C., Zhang, X., Zhang, Y., et al. (2018). Genetic dissection of maize embryonic callus regenerative capacity using multi-locus genome-wide association studies. *Front. Plant Sci.* 9:561. doi: 10.3389/fpls.2018.00561

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2018). pKWmEB: integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wang, T. T., Wang, M., Hu, S. T., Xiao, Y. N., Tong, H., Pan, Q. C., et al. (2015). Genetic basis of maize kernel starch content revealed by high-density single nucleotide polymorphism markers in a recombinant inbred line population. *BMC Plant Biol.* 15:288. doi: 10.1186/S12870-015-0675-2

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Wilson, L. M., Whitt, S. R., Ibanez, A. M., Rocheford, T. R., Goodman, M. M., and Buckler, E. S. (2004). Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell* 16, 2719–2733. doi: 10.1105/tpc.104.025700

Xiao, Y. J., Liu, H. J., Wu, L. J., Warburton, M., and Yan, J. B. (2017). Genome-wide association studies in maize: praise and stargaze. *Mol. Plant* 10, 359–374. doi: 10.1016/j.molp.2016.12.008

Xu, S., Yang, Z., Zhang, E., Jiang, Y., Pan, L. A., Chen, Q., et al. (2014a). Nucleotide diversity of maize zmbt1 gene and association with starch physicochemical properties. *PLoS One* 9:e103627. doi: 10.1371/journal.pone.0103627

Xu, S., Zhu, D., and Zhang, Q. (2014b). Predicting hybrid performance in rice using genomic best linear unbiased prediction. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12456–12461. doi: 10.1073/pnas.1413750111

Xu, Y., Xu, C., and Xu, S. (2017). Prediction and association mapping of agronomic traits in maize using multiple omic data. *Heredity* 119, 174–184. doi: 10.1038/hdy.2017.27

Yang, N., Lu, Y., Yang, X., Huang, J., Zhou, Y., Ali, F., et al. (2014). Genome wide association studies using a new nonparametric model reveal the genetic architecture of 17 agronomic traits in an enlarged maize association panel. *PLoS Genet.* 10:e1004573. doi: 10.1371/journal.pgen.1004573

Yang, Z., Zhang, E., Jiang, Y., Xu, S., Pan, L., Chen, Q., et al. (2014). Sequence polymorphisms in Zmisa2 gene are significantly associated with starch pasting and gelatinization properties in maize (*Zea mays* L.). *Mol. Breed.* 34, 1833–1842. doi: 10.1007/s11032-014-0142-z

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, X., Colleoni, C., Ratushna, V., Sirghie-Colleoni, M., James, M. G., and Myers, A. M. (2004). Molecular characterization demonstrates that the *Zea mays* gene sugary2 codes for the starch synthase isoform SSIIa. *Plant Mol. Biol.* 54, 865–879. doi: 10.1007/s11103-004-0312-1

Zhang, Y., Liu, P., Zhang, X., Zheng, Q., Chen, M., Ge, F., et al. (2018). Multi-locus genome-wide association study reveals the genetic architecture of stalk lodging resistance-related traits in maize. *Front. Plant Sci.* 9:611. doi: 10.3389/fpls.2018.00611

Zhang, Y. M., Mao, Y., Xie, C., Smith, H., Luo, L., and Xu, S. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics* 169, 2267–2275. doi: 10.1534/genetics.104.033217

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Zhu, C., Gore, M., Buckler, E. S., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome* 1, 5–20. doi: 10.3835/plantgenome2008.02.0089

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Plant Science

ORIGINAL RESEARCH
published: 13 September 2018
doi: 10.3389/fpls.2018.01067

GWAS Analysis and QTL Identification of Fiber Quality Traits and Yield Components in Upland Cotton Using Enriched High-Density SNP Markers

Ruixian Liu[1,2†], Juwu Gong[1,2†], Xianghui Xiao[1,2†], Zhen Zhang[2], Junwen Li[2], Aiying Liu[2], Quanwei Lu[2,3], Haihong Shang[2], Yuzhen Shi[2], Qun Ge[2], Muhammad S. Iqbal[2], Xiaoying Deng[2], Shaoqi Li[2], Jingtao Pan[2], Li Duan[2], Qi Zhang[2], Xiao Jiang[2], Xianyan Zou[2], Abdul Hafeez[2], Quanjia Chen[1], Hongwei Geng[1], Wankui Gong[2]* and Youlu Yuan[1,2]**

[1] Xinjiang Research Base, State Key Laboratory of Cotton Biology, Xinjiang Agricultural University, Urumqi, China, [2] State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang, China, [3] School of Biotechnology and Food Engineering, Anyang Institute of Technology, Anyang, China

It is of great importance to identify quantitative trait loci (QTL) controlling fiber quality traits and yield components for future marker-assisted selection (MAS) and candidate gene function identifications. In this study, two kinds of traits in 231 $F_{6:8}$ recombinant inbred lines (RILs), derived from an intraspecific cross between Xinluzao24, a cultivar with elite fiber quality, and Lumianyan28, a cultivar with wide adaptability and high yield potential, were measured in nine environments. This RIL population was genotyped by 122 SSR and 4729 SNP markers, which were also used to construct the genetic map. The map covered 2477.99 cM of *hirsutum* genome, with an average marker interval of 0.51 cM between adjacent markers. As a result, a total of 134 QTLs for fiber quality traits and 122 QTLs for yield components were detected, with 2.18–24.45 and 1.68–28.27% proportions of the phenotypic variance explained by each QTL, respectively. Among these QTLs, 57 were detected in at least two environments, named stable QTLs. A total of 209 and 139 quantitative trait nucleotides (QTNs) were associated with fiber quality traits and yield components by four multilocus genome-wide association studies methods, respectively. Among these QTNs, 74 were detected by at least two algorithms or in two environments. The candidate genes harbored by 57 stable QTLs were compared with the ones associated with QTN, and 35 common candidate genes were found. Among these common candidate genes, four were possibly "pleiotropic." This study provided important information for MAS and candidate gene functional studies.

Keywords: upland cotton, QTL, multilocus GWAS, QTN, candidate gene, fiber quality traits, yield components

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Hongde Qin,
Hubei Academy of Agricultural Sciences, China
Xinlian Shen,
Jiangsu Academy of Agricultural Sciences (JAAS), China

Correspondence:
Hongwei Geng
hw_geng@163.com
Wankui Gong
gongwankui@caas.cn
Youlu Yuan
yuanyoulu@caas.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *03 May 2018*
Accepted: *02 July 2018*
Published: *13 September 2018*

Citation:
Liu R, Gong J, Xiao X, Zhang Z, Li J, Liu A, Lu Q, Shang H, Shi Y, Ge Q, Iqbal MS, Deng X, Li S, Pan J, Duan L, Zhang Q, Jiang X, Zou X, Hafeez A, Chen Q, Geng H, Gong W and Yuan Y (2018) GWAS Analysis and QTL Identification of Fiber Quality Traits and Yield Components in Upland Cotton Using Enriched High-Density SNP Markers. Front. Plant Sci. 9:1067. doi: 10.3389/fpls.2018.01067

INTRODUCTION

Cotton is an important cash crop that provides major natural fiber supply for textile industry and human daily life. Four species in *Gossypium*, namely *G. herbaceum* (A1), *G. arboreum* (A2), *G. hirsutum* (AD1), and *G. barbadense* (AD2), are cultivated ones. *G. hirsutum* ($2n = 4x = 52$, genome size: 2.5 Gb) (Li et al., 2014, 2015; Wendel and Grover, 2015; Zhang et al., 2015), also called upland cotton, has a high yield potential, whereas fair fiber quality attributes (Cai et al., 2014), thus making it most widely cultivated and utilized worldwide, approximately accounting for 95% of global cotton fiber production (Chen et al., 2007). Along with the progress of technologies in textile industry and improvement of human living standard, the demand for cotton fiber supply not only increases in quantity but also is required in a diverse combination of various qualities such as high strength, natural color, various lengths, and fineness. Fiber quality traits and yield components are quantitative and controlled by multiple genes (Said et al., 2013), yet most of which were negatively correlated with each other (Shen et al., 2007; Wang H. et al., 2015). Therefore, it is difficult to improve all these traits simultaneously by traditional breeding programs, even after time-consuming and laborious efforts were put (Shen et al., 2005; Lacape et al., 2009; Jamshed et al., 2016; Zhang et al., 2016). The rapid development of applied genome research provides an effective tool for improving plant breeding efficiency, a typical example of which is the marker-assisted selection (MAS) and genome selection through the molecular markers closely linked to target genes or quantitative trait loci (QTLs).

Currently, plenty of intraspecific segregating populations of *G. hirsutum* are constructed targeting various traits in upland cotton, and many QTLs are identified, including those for fiber quality traits (Shen et al., 2005; Sun F.D. et al., 2012; Fang et al., 2014; Xu et al., 2014; Tan et al., 2015; Wang H. et al., 2015; Jamshed et al., 2016; Li et al., 2016; Yang et al., 2016; Liu et al., 2017; Zhang Z. et al., 2017), yield components (Xia et al., 2014; Wang H. et al., 2015; Zhang et al., 2016; Liu et al., 2017), drought tolerances (Levi et al., 2011), disease resistances (Jiang et al., 2009; Ulloa et al., 2013; Zhao et al., 2014; Palanga et al., 2017), early maturity (Stiller et al., 2004; Li et al., 2012, 2013), and plant morphological traits (Tang and Xiao, 2014; Qi et al., 2017).

A genome-wide association studies (GWAS) is also an effective approach for connecting phenotypes and genotypes in plants, and helps us to avoid the difficulty of screening large biparental mapping populations, so it is widely applied to various studies (Thornsberry et al., 2001; Flint-Garcia et al., 2005; Maccaferri et al., 2005; Eizenga et al., 2006; Zhu et al., 2008; Jia et al., 2014; Nie et al., 2016) to identify quantitative trait nucleotides (QTNs) for complex traits (Zhao et al., 2011; Fernandes et al., 2012; Segura et al., 2012; Spindel et al., 2015). It has been successfully applied to *Arabidopsis thaliana* (Atwell et al., 2010; Horton et al., 2012), rice (Huang et al., 2010; Zhao et al., 2011), corn (Kump et al., 2011; Samayoa et al., 2015), and soybean (Dhanapal et al., 2015; Zeng et al., 2017), and many QTNs and their candidate genes have been identified for various ecological and agricultural traits. More recently, it has also been used in cotton (Abdurakhmonov et al., 2008; Kantartzi and Stewart, 2008; Zeng et al., 2009; Cai et al., 2014; Mei et al., 2013; Zhang et al., 2013; Su et al., 2016; Huang et al., 2017; Sun et al., 2017). To better understand the genetic architecture of fiber quality traits and yield components in upland cotton, we genotyped an intraspecific recombinant inbred lines (RILs) using enriched high-density markers of both single-nucleotide polymorphisms (SNPs) based on the CottonSNP80K arrays (Cai et al., 2017) and simple sequence repeats (SSRs). To obtain reliable QTLs and their candidate genes, we tried to use two strategies. One was linkage-map-based QTL mapping, in which a high coverage genetic linkage map was constructed with HighMap software and QTLs were mapped using composite interval mapping (CIM); the other was GWAS along with four multilocus GWAS methods (Wang et al., 2016; Tamba et al., 2017; Wen et al., 2017; Zhang J. et al., 2017). The results in the study could be worthy for further studies not only in molecular-assisted breeding through MAS but also in functional gene validations, which is of great significance to the improvement of cotton fiber quality and yield.

MATERIALS AND METHODS

Plant Materials

An RIL population of 231 lines was developed from a cross between two homozygous upland cotton cultivars, Lumianyan28 (LMY28), a commercial transgenic cultivar with high yield potential and wide adaptability developed by the Cotton Research Center of Shandong Academy of Agricultural Sciences as a maternal line, and Xinluzao24 (XLZ24), a high fiber quality upland cotton cultivar with long-staple developed by XinJiang KangDi company as a paternal line.

The RIL development was briefed as follows: the cross between LMY28 and XLZ24 was made in the summer growing season in 2008 in Anyang, Henan Province. F_1 were planted and self-pollinated in the winter growing season in 2008 in Hainan Province. In the spring of 2009, 238 F_2 plants were grown and self-pollinated, and $F_{2:3}$ seeds were harvested in Anyang (Kong et al., 2011). Of the 238 $F_{2:3}$ lines, 231 were self-pollinated in each generation until $F_{2:6}$. Then single plant selection was made from each of the 231 $F_{2:6}$ lines to form the $F_{6:7}$ population. The $F_{6:7}$ population was planted in plant rows and self-pollinated to construct the $F_{6:8}$ RIL population. All the generations beyond $F_{6:8}$ are regarded as $F_{6:8}$ for convenience of analysis. The target traits of the $F_{6:8}$ RIL population were evaluated in Henan (Anyang, 2013, 2014, 2015, and 2016, designated as 13AY, 14AY, 15AY, and 16AY, respectively), Shandong (LinQing, 2013 and 2014, designated as 13LQ and 14LQ, respectively), Hebei (Quzhou 2013, designated as 13QZ), and Xinjiang (Kuerle 2014 and Alaer 2015, designated as 14KEL and 15ALE, respectively), and a randomized complete block design with two replications was adopted in all nine environmental evaluations. A single-row plot with 5-m row length, 0.8-m row spacing, and 0.25-m plant spacing was adopted in 13AY, 13LQ, 13QZ, 14AY, 14LQ, 15AY, and 16AY, whereas a two-narrow-row plot with 3-m row length, 0.66/0.10-m alternating row spacing, and 0.12-m plant spacing were adopted in 14KEL and 15ALE.

Phenotypic Detection and Data Analysis

Thirty naturally opened bolls from each plot were hand-harvested on the inner fruiting nods from middle to upper branches. Yield component traits, including boll weight (BW, g), lint percentage (LP, %), and seed index (SI, g), were evaluated. No less than 15 g fibers were sampled to evaluate the fiber quality traits, including fiber length (FL, mm), fiber strength (FS, cN tex^{-1}), and fiber micronaire (FM). The evaluations were conducted using HFT9000 (Premier Evolvics Pvt. Ltd., India) instruments with HVICC Calibration in the Cotton Quality Supervision, Inspection and Testing Center, Ministry of Agriculture, Anyang, Henan Province, China.

One-way analysis of variance (ANOVA) between parents and the descriptive statistics for the RIL population was conducted using Microsoft Excel 2016, and correlation analysis was performed using SPSS 20.0 (SPSS, Chicago, IL, United States). Integrated ANOVA across nine environments along with the heritability of all the traits was conducted using ANOVA function in the QTL IciMapping software.

DNA Extraction and Genotyping

Genomic DNA was extracted from fresh leaves of parents and 231 RILs with a modified cetyltrimethyl ammonium bromide (CTAB) method (Song et al., 1998). The DNA was used both for SSR screening and CottonSNP80K array hybridization.

A total of 9668 pairs of SSR primer pool, which contained a variety of sources including NAU, BNL, DPL, CGR, PGML, SWU, and CCRI, were used to screen the polymorphisms between parents. The primer information was also available at the CottonGen Database[1]. PCR amplification and product detection were conducted according to the procedures described by Zhang et al. (2005). The polymorphic primers between the parents were used to genotype the population, and the SSR markers that were codominant and had a unique physical location in the reference genome were used to construct the linkage map.

The cottonSNP80K array, which contained 77,774 SNPs (Cai et al., 2017), was used to genotype the parents and the 231 RILs. The genotyping was conducted according to the Illumina suggestions (Illumina Inc., San Diego, CA, United States) (Cai et al., 2017). After genotyping, the raw data were filtered based on the following criteria (Zhang Z. et al., 2017): first, any or both of the SNP loci of parents were missing (69,395 SNPs were remained after filtering); second, the loci had no polymorphism between parents (15,128 loci were remained); third, the loci of any of the parent were heterozygous (7480 SNPs were remained); forth, the missing rate of SNPs in the population was more than 40% (Hulse-Kemp et al., 2015) (7479 loci were remained); and finally, the segregation distortion of SNPs reached criteria of $P < 0.001$ (5202 loci were remained). Subsequently, the remaining SNP markers were applied to the genetic map construction after converting into the "ABH" data format as SSR.

Genetic Map Construction

The remaining SSR and SNP markers were divided into the 26 chromosomes based on their position on the physical map of the upland cotton (TM-1) genome database (Zhang et al., 2015). Then, the genetic linkage map was constructed using the HighMap software with multiple sorting and error-correcting functions (Liu et al., 2014). Map distances were estimated using Kosambi's mapping function (Kosambi, 1943).

The significance of segregation distortion markers (SDMs; $P < 0.05$) was detected using the chi-square test. The regions containing at least three consecutive SDMs were defined as segregation distortion regions (SDRs) (Zhang et al., 2016). The distribution of SDMs and SDRs, and the size of SDRs on the map were analyzed.

QTL Mapping and Genome-Wide Association Studies

The Windows QTL Cartographer 2.5 software (Wang et al., 2012) was employed using the CIM method with a mapping step of 1.0 cM and five control markers (Zeng, 1994) for QTL identification. The threshold value of the logarithm of odds (LOD) was calculated by 1000 permutations at the 0.05 significance level. QTLs, identified in different environments and had fully or partially overlapping confidence intervals, were regarded as the same QTL. The QTL detected in at least two environments was regarded as a stable one. Nomenclature of QTL was designated following Sun's description (Sun F.D. et al., 2012). MapChart 2.3 (Voorrips, 2002) was used to graphically represent the genetic map and QTL.

Quantitative trait nucleotides for the target traits were identified by four multilocus GWAS methods. The first one is mrMLM (Wang et al., 2016), in which calculate Kinship (K) matrix model was used, with critical *P*-value of 0.01, search radius of the candidate gene of 20 kb, and critical LOD score for significant QTN of 3. The second one is FASTmrEMMA (Wen et al., 2017), with restricted maximum likelihood, in which calculate K matrix model was used, critical *P*-value of 0.005, and critical LOD score for significant QTN of 3. The third one is ISIS EM-BLASSO (Tamba et al., 2017), with critical *P*-value of 0.01. The fourth one is pLARmEB (Zhang J. et al., 2017); each chromosome selected 50 potential associations at a critical LOD score of 2 with variable selection through LAR.

QTL Congruency Comparison With Previous Studies

Previous QTLs for the target traits were detected and downloaded in the CottonQTLdb database[2] (Said et al., 2015). The QTLs sharing similar genetic positions (spacing distance < 15 cM) were regarded as common or same QTL. The physical positions of a QTL were identified in the CottonGen database[3]. When a QTL in the current study shared the same physical region as the previous QTL, it was regarded as a repeated identification of the previous QTL; otherwise, the QTL in the current study was regarded as a new one.

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

[2] http://www.cottondb.org

[3] http://www.cottongen.org

The Candidate Genes Identification

Candidate genes harbored in the stable QTLs were searched and identified based on their confidence intervals in the following steps: The markers including the closest flanking ones in the confidence interval of a QTL were identified. The physical interval of that QTL was determined based on the physical position of its markers in the upland cotton (TM-1) genome[4] (Zhang et al., 2015). All the genes in the physical interval were identified as candidate genes.

Candidate genes associated with QTNs in the multilocus GWAS analysis were confirmed based on the location of QTNs in the upland cotton (TM-1) reference genome (Zhang et al., 2015). The gene in which the QTL was located was considered as the candidate gene. But when the physical location of a QTN was between two genes, both of the genes were considered as candidate genes.

RESULTS

Phenotypic Evaluation of the RIL Populations

The one-way ANOVA between parents in nine environments showed that a significant difference for FS at the 0.001 level and no significant differences for the other traits were observed (**Table 1**). The descriptive statistical analysis showed that all traits in the RIL population performed transgressive segregations, with approximately normal distribution in all the nine environments (**Table 1**). The integrated ANOVA of the RILs across nine environments also revealed significant variations for all traits among the RILs (**Supplementary Table S1**).

Most of the traits exhibited medium–high heritability across nine environments (**Supplementary Table S2**). Correlation analysis showed that significant or very significant positive correlations were observed between the trait pairs of FL–FS, FL–SI, FS–SI, FM–LP, FM–BW, and SI–BW; and significant negative correlations were observed between the pairs of FL–FM, FL–LP, FS–FM, FS–LP, BW–LP, and SI–LP. In addition, FL–BW showed a significant or very significant positive correlation in three environments, whereas no significant correlation was observed in the remaining six environments (**Table 2**).

Genetic Map Construction

The genetic linkage map totally covered 2477.99 cM of the upland cotton genome with an average adjacent marker interval of 0.51 cM (**Figure 1** and **Table 3**). It contained 4851 markers, including 4729 SNP and 122 SSR loci, with uneven distributions in the A_t and D_t subgenomes as well as on 26 chromosomes. A total of 3300 markers were mapped in the A_t subgenome, covering a genetic distance of 1474.63 cM with an average adjacent marker interval of 0.45 cM. On the other hand, a total of 1551 markers were mapped in the D_t subgenome, covering a genetic distance of 1003.36 cM with an average adjacent marker interval of 0.65 cM. At the chromosome level, chr08 contained the maximum number of markers (481 markers), spanning a genetic distance of 142.55 cM with an average adjacent marker interval of 0.32 cM. chr17 contained the minimum number of markers (19 markers), spanning a total genetic distance of 60.60 cM with an average adjacent marker interval of 3.56 cM. Gap analysis revealed that there were 33 gaps ($\geq$10 cM), of which 19 were in the A_t subgenome with the largest of 22.68 cM on chr07, whereas 14 were in the D_t subgenome with the largest of 42.23 cM on chr17. chr11, chr16, chr19, chr20 and chr24 had no gap larger than 10 cM.

Segregation Distortion

There were a total of 1,563 SDMs (32.22%) ($P < 0.05$), which were unevenly distributed at both subgenome and chromosome levels (**Tables 3** and **Supplementary Table S3**). One thousand and sixty-one SDMs were found in the A_t subgenome, whereas 502 in the D_t subgenome. chr08 had the maximum number of SDMs of 237 (15.16% of total SDMs). The SDMs formed 110 SDRs, of which 66 were in the A_t subgenome whereas 44 in the D_t subgenome. chr05 contained the maximum number of SDRs of 10. There was no SDR in chr03 and chr17.

Collinearity Analysis

The reliability of the genetic map was usually assessed by comparing it with the physical maps of the upland cotton (TM-1) reference genome (Zhang et al., 2015). The results of the collinear analysis are shown in **Figure 2**. The results revealed an overall good congruency between the linkage map and its physical one, while there also existed some discrepancies between the two on chr03, chr06, chr08, and chr13 in the A_t subgenome and on chr15, chr16, chr17, chr19, chr22, chr23, and chr26 in the D_t subgenome. The collinearity in subgenomes revealed that the A_t subgenome showed a better compatibility between the linkage and the physical maps than the D_t subgenome did.

QTL Mapping for Fiber Quality Traits and Yield Components

A total of 256 QTLs (**Supplementary Table S4**), 134 for fiber quality traits, and 122 for yield components, were identified across nine environments using the CIM algorithm, with 1.68–28.27% proportions of the phenotypic variance (PV) explained by each QTL. Fifty-seven stable QTLs (**Figure 3** and **Supplementary Table S4**) were identified in at least two environments, of which 32 were for fiber quality traits and 25 for yield components.

Fiber Length

A total of 36 QTLs for FL were identified on 21 chromosomes except chr02, chr04, chr09, chr10, and chr25, among which 7 were stable (**Figure 3** and **Supplementary Table S4**). In these stable QTLs, qFL-chr17-1 was identified in three environments, and could explain 3.95–5.36% proportions of the observed PV. In its marker interval of TM53503–TM53577, there harbored 88 candidate genes. The stable QTLs, qFL-chr05-1, qFL-chr06-2, qFL-chr11-1, qFL-chr16-1, qFL-chr19-1, and qFL-chr26-1, could

[4] http://mascotton.njau.edu.cn/info/1054/1118.htm

TABLE 1 | The results of the statistical analysis of the parents and the RIL population.

Trait[a]	Env.[b]	Parents				RIL population							
		XLZ24[c]	**LMY28[d]**	**Range**	***P*-value**	**Min.**	**Max.**	**Range**	**Mean**	***SD***	**Var.**	**Skew.**	**Kurt.**
FL	13*AY*	29.72	29.23	0.49	0.06	28.12	32.52	4.40	30.54	0.92	0.03	−0.13	−0.44
	13*LQ*	30.85	29.57	1.28		27.86	33.05	5.19	30.31	1.02	0.03	0.28	−0.10
	13*QZ*	29.04	28.28	0.75		27.29	33.61	6.32	29.60	1.16	0.04	0.52	0.72
	14*AY*	31.43	29.44	2.00		28.85	33.65	4.80	31.26	1.00	0.03	0.11	−0.36
	14*KEL*	29.77	29.69	0.09		28.35	34.07	5.72	31.14	1.04	0.03	0.35	0.23
	14*LQ*	30.80	30.40	0.40		28.90	34.05	5.15	31.27	1.04	0.03	0.15	−0.30
	15*AY*	29.95	26.85	3.10		26.15	31.60	5.45	29.06	1.06	0.04	−0.12	−0.35
	15*ALE*	29.10	29.20	−0.10		26.60	31.80	5.20	29.05	1.08	0.04	0.22	−0.31
	16*AY*	28.95	28.80	0.15		28.40	34.05	5.65	30.55	0.99	0.03	0.61	1.02
FS	13*AY*	33.95	30.45	3.50	< 0.001***	29.15	39.30	10.15	33.39	1.90	0.06	0.44	−0.03
	13*LQ*	30.15	27.55	2.60		27.20	36.25	9.05	32.14	1.82	0.06	0.01	−0.26
	13*QZ*	31.65	29.20	2.45		28.10	39.10	11.00	32.69	2.05	0.06	0.43	−0.16
	14*AY*	31.95	29.45	2.50		28.55	36.85	8.30	32.42	1.55	0.05	0.18	0.01
	14*KEL*	31.55	29.15	2.40		27.65	37.70	10.05	32.02	1.75	0.05	0.46	0.49
	14*LQ*	30.40	29.10	1.30		28.80	37.50	8.70	32.77	1.60	0.05	0.36	0.23
	15*AY*	33.75	28.60	5.15		27.85	39.65	11.80	33.48	2.10	0.06	0.09	0.21
	15*ALE*	31.25	27.90	3.35		26.90	35.15	8.25	30.92	1.64	0.05	0.07	−0.14
	16*AY*	32.80	27.85	4.95		28.75	38.05	9.30	32.98	1.67	0.05	0.21	0.08
FM	13*AY*	4.49	4.48	0.00	0.46	2.96	5.42	2.46	4.26	0.50	0.12	−0.08	−0.42
	13*LQ*	3.91	4.41	−0.50		2.25	5.32	3.07	3.91	0.59	0.15	−0.31	−0.15
	13*QZ*	5.09	4.70	0.40		2.74	5.58	2.84	4.29	0.61	0.14	−0.43	−0.36
	14*AY*	4.79	4.65	0.14		3.50	5.62	2.12	4.59	0.40	0.09	−0.12	−0.21
	14*KEL*	4.58	4.56	0.02		3.56	5.28	1.72	4.43	0.34	0.08	0.04	−0.17
	14*LQ*	5.30	4.85	0.45		3.60	5.50	1.90	4.74	0.38	0.08	−0.36	−0.11
	15*AY*	4.65	4.95	−0.30		3.50	5.60	2.10	4.55	0.39	0.09	−0.12	−0.20
	15*ALE*	4.60	4.30	0.30		4.00	5.55	1.55	4.86	0.32	0.07	−0.20	−0.60
	16*AY*	5.30	4.70	0.60		3.65	5.80	2.15	4.77	0.41	0.09	−0.22	−0.06
BW	13*AY*	5.47	5.79	−0.32	0.16	4.40	7.20	2.80	5.93	0.50	0.08	−0.29	0.34
	13*LQ*	5.26	5.58	−0.33		3.76	7.07	3.31	5.55	0.69	0.12	−0.27	−0.57
	13*QZ*	5.03	5.61	−0.58		3.12	6.87	3.75	5.11	0.77	0.15	−0.24	−0.40
	14*AY*	5.68	6.13	−0.46		5.24	7.88	2.64	6.39	0.50	0.08	0.16	0.00
	14*KEL*	6.04	6.16	−0.12		4.65	7.41	2.76	6.23	0.53	0.09	−0.21	0.19
	14*LQ*	6.23	6.42	−0.19		4.80	8.12	3.32	6.80	0.62	0.09	−0.49	0.32
	15*AY*	5.44	5.65	−0.21		4.21	6.69	2.48	5.41	0.42	0.08	0.05	0.65
	15*ALE*	4.91	5.00	−0.08		4.97	7.22	2.25	5.99	0.43	0.07	0.20	−0.21
	16*AY*	5.47	5.81	−0.34		4.62	7.81	3.19	6.27	0.54	0.09	−0.08	0.28
LP	13*AY*	41.87	39.01	2.86	0.15	29.56	43.56	14.00	37.82	2.33	0.06	−0.45	0.88
	13*LQ*	38.53	36.95	1.59		29.02	42.59	13.57	36.36	2.38	0.07	−0.02	0.13
	13*QZ*	37.93	36.52	1.41		27.09	42.97	15.88	35.61	3.06	0.09	−0.22	0.09
	14*AY*	44.48	42.13	2.35		35.70	46.73	11.03	41.16	1.99	0.05	−0.15	0.00
	14*KEL*	43.07	39.10	3.97		33.34	46.03	12.69	39.62	2.21	0.06	−0.10	0.08
	14*LQ*	42.74	42.42	0.32		33.96	48.51	14.55	41.14	2.39	0.06	−0.14	0.34
	15*AY*	47.08	44.53	2.55		31.56	47.15	15.59	40.80	2.42	0.06	−0.58	1.69
	15*ALE*	43.42	41.96	1.46		38.10	48.66	10.56	44.33	1.92	0.04	−0.49	0.53
	16*AY*	41.59	40.24	1.36		34.09	46.83	12.74	39.49	2.30	0.06	0.29	0.04
SI	13*AY*	10.60	11.00	−0.40	0.68	9.39	14.44	5.05	11.80	1.03	0.09	0.11	−0.31
	13*LQ*	10.66	11.63	−0.97		9.23	14.18	4.95	11.84	1.05	0.09	0.22	−0.43
	13*QZ*	12.2	11.48	0.76		9.31	15.20	5.89	12.21	1.17	0.10	0.11	−0.05
	14*AY*	–	–	–		–	–	–	–	–	–	–	–

(Continued)

TABLE 1 | Continued

Trait[a]	Env.[b]	Parents				RIL population							
		XLZ24[c]	LMY28[d]	Range	*P*-value	Min.	Max.	Range	Mean	*SD*	Var.	Skew.	Kurt.
	14*KEL*	11.53	11.97	−0.45		10.09	14.80	4.71	12.28	1.08	0.09	0.21	−0.76
	14*LQ*	–	–	–		–	–	–	–	–	–	–	–
	15*AY*	9.55	9.85	−0.30		8.25	12.55	4.30	10.26	0.84	0.08	0.19	−0.24
	15*ALE*	9.92	9.77	0.15		8.73	13.40	4.67	10.47	0.89	0.09	0.45	0.06
	16*AY*	12.55	12.13	0.43		10.13	15.35	5.22	12.55	1.01	0.08	0.24	0.03

[a] *FL, fiber length; FS, fibers strength; FM, fiber micronaire; BW, weight; LP, lint percent; SI, seed index.* [b] *13AY, Anyang in 2013; 13LQ, Linqing in 2013; 13QZ, Quzhou in 2013; 14AY, Anyang in 2014; 14KEL, Kerle in 2014; 14LQ, Linqing in 2014; 15AY, Anyang in 2015; 15ALE, Alaer in 2015; 16AY, Anyang in 2016.* [c] *Xinluzao24.* [d] *Lumianyan28.*

explain 12.13–13.83, 6.35–6.62, 5.15–9.41, 5.24–6.23, 4.65–5.07, and 4.56–5.59% proportions of the observed PVs, respectively. In their marker intervals of CICR0262, TM18200–TM18321, TM39956–TM39953, TM66757–NAU3563, TM57055–TM57082, and TM77259–TM77261, there harbored 2, 141, 15, 309, 65, and 1 candidate genes, respectively.

Fiber Strength

Forty-six QTLs for FS were identified on 19 chromosomes except chr02, chr03, chr14, chr17, chr18, chr22, and chr23, among which 10 were stable (**Figure 3** and **Supplementary Table S4**). In these stable QTLs, qFS-chr07-2 was identified in all nine environments, and could explain 5.81–19.47% proportions of the observed PV. In its marker interval of DPL0852–DPL0757, eight candidate genes were harbored. qFS-chr16-3 was identified in five environments, and could explain 4.28–6.45% proportions of the observed PV. In its marker interval of SWU2707–DPL0492, 342 candidate genes were harbored. qFS-chr01-2 and qFS-chr20-5 were identified in three environments, and could explain 5.32–8.86 and 4.50–5.90% proportions of the observed PVs, respectively. In their marker intervals of TM379–TM404 and NAU4989–TM73152, 20 and 7 candidate genes, respectively, were harbored. qFS-chr07-1, qFS-chr11-1, qFS-chr11-2, qFS-chr13-1, qFS-chr20-1, and qFS-chr24-1 were identified in two environments, and could explain 5.97–6.21, 4.87–5.59, 5.26–7.21, 5.74–10.69, 2.91–8.18, and 5.11–5.43% proportions of the observed PVs, respectively. In their marker intervals of TM19848–TM19875, TM37826–TM37828, TM37897–TM37935, TM43230–TM43229, TM75088–TM75100, and TM67152–TM67146, 4, 1, 29, 1, 8, and 6 candidate genes, respectively, were harbored.

Fiber Micronaire

Fifty-two QTLs for FM were identified on 21 chromosomes except chr02, chr12, chr17, chr23, and chr26, among which 15 were stable (**Figure 3** and **Supplementary Table S4**). In these stable QTLs, qFM-chr07-1 and qFM-chr13-1 were identified in six environments, and could explain 5.51–24.45 and 4.73–8.88% proportion of the observed PV, respectively. In their marker intervals of DPL0852–DPL0757 and TM43230–TM43241, 8 and 15 candidate genes, respectively, were harbored. qFM-chr01-2 was identified in five environments, and could explain 3.94–6.17% proportions of the observed PVs. In its marker interval, one marker of TM3451 was exclusively contained and two candidate genes were harbored. qFM-chr19-1 and qFM-chr19-2 were identified in four environments, and could explain 4.57–8.54 and 5.19–8.20% proportions of the observed PVs, respectively. In their marker intervals of TM57055–TM57057 and TM56813–TM56753, 4 and 161 candidate genes, respectively, were harbored. qFM-chr14-1, qFM-chr15-1, and qFM-chr24-2 were identified in three environments, and could explain 4.18–6.53, 4.45–5.35, and 4.25–4.69% proportions of the observed PVs, respectively. In their marker intervals of TM50241–TM50231, CGR5709–TM50087, and TM67152–TM67125, 13, 1, and 18 candidate genes, respectively, were harbored. qFM-chr03-1, qFM-chr05-1, qFM-chr10-1, qFM-chr11-4, qFM-chr14-3, qFM-chr15-2, and qFM-chr20-2 were identified in two environments, and could explain 3.89–5.18, 4.13–4.42, 4.66–5.16, 4.22–4.30, 4.52–4.54, 3.75–5.95, and 4.47–5.02% proportions of the observed PVs, respectively. In their marker intervals of TM7008–TM7102, TM10798–TM10805, TM33784–TM33813, TM39510–TM39490, TM52033–TM52031, TM50087–TM50082, and TM75041–TM75030, 125, 6, 100, 8, 1, 5, and 44 candidate genes, respectively, were harbored.

Boll Weight

A total of 53 QTLs for BW were identified on 25 chromosomes except chr15, among which 7 were stable (**Figure 3** and **Supplementary Table S4**). In these stable QTLs, qBW-chr24-1 was identified in three environments, and could explain 4.13–6.99% proportions of the observed PVs. In its marker interval of TM67152–TM67127, 18 candidate genes were harbored. qBW-chr04-2, qBW-chr05-5, qBW-chr06-3, qBW-chr07-4, qBW-chr20-1, and qBW-chr21-4 were identified in two environments, and could explain 3.77–5.74, 4.28–6.42, 3.87–4.07, 7.62–8.08, 5.56–8.12, and 6.05–7.26% proportions of the observed PVs, respectively. In their marker intervals of TM9831–TM9827, TM10953–TM10979, TM14514–TM14509, DPL0852, NAU4989–CICR0002, and TM76018–TM75887, 6, 59, 23, 2, 7, and 119 candidate genes were harbored, respectively.

Lint Percentage

A total of 39 QTLs for LP were identified on 20 chromosomes except chr02, chr12, chr15, chr17, chr23, and chr24, among

TABLE 2 | Correlation analysis between fiber quality and yield component traits in the RIL population.

Trait[a]	Environment[b]	FL	FS	FM	LP	BW
FS	13AY	0.271**				
	13QZ	0.510**				
	13LQ	0.139*				
	14AY	0.600**				
	14KEL	0.642**				
	14LQ	0.513**				
	15ALE	0.660**				
	15AY	0.466**				
	16AY	0.534**				
FM	13AY	−0.209**	−0.587**			
	13QZ	−0.226**	−0.576**			
	13LQ	−0.051	−0.348**			
	14AY	−0.417**	−0.486**			
	14KEL	−0.449**	−0.378**			
	14LQ	−0.413**	−0.463**			
	15ALE	−0.539**	−0.353**			
	15AY	−0.383**	−0.425**			
	16AY	−0.369**	−0.566**			
LP	13AY	−0.285**	−0.204**	0.366**		
	13QZ	−0.225**	−0.271**	0.454**		
	13LQ	−0.169*	−0.088	0.430**		
	14AY	−0.259**	−0.179**	0.220**		
	14KEL	−0.424**	−0.421**	0.351**		
	14LQ	−0.311**	−0.278**	0.360**		
	15ALE	−0.373**	−0.226**	0.352**		
	15AY	−0.149*	−0.088	0.044		
	16AY	−0.189**	−0.202**	0.336**		
BW	13AY	0.178**	−0.294**	0.352**	−0.197**	
	13QZ	0.006	−0.343**	0.543**	0.123	
	13LQ	0.127	−0.373**	0.544**	0.052	
	14AY	0.122	0.085	0.240**	−0.137*	
	14KEL	0.064	0.276**	−0.089	−0.352**	
	14LQ	0.060	−0.046	0.300**	−0.061	
	15ALE	0.237**	0.225**	−0.078	−0.146*	
	15AY	0.134*	−0.037	0.320**	−0.259**	
	16AY	−0.050	−0.199**	0.379**	−0.142*	
SI	13AY	0.284**	0.275**	−0.299**	−0.619**	0.324**
	13QZ	0.245**	0.250**	−0.159*	−0.467**	0.294**
	13LQ	0.187**	0.202**	−0.267**	−0.541**	0.094
	14AY	–	–	–	–	–
	14KEL	0.307**	0.453**	−0.102	−0.597**	0.565**
	14LQ	–	–	–	–	–
	15ALE	0.402**	0.418**	−0.179**	−0.470**	0.670**
	15AY	0.299**	0.363**	−0.182**	−0.309**	0.332**
	16AY	0.158*	0.336**	−0.192**	−0.256**	0.107

*[a]FL, fiber length; FS, fibers strength; FM, fiber micronaire; BW, weight; LP, lint percent; SI,seed index. [b]13AY, Anyang in 2013; 13LQ, Linqing in 2013; 13QZ, Quzhou in 2013; 14AY, Anyang in 2014; 14KEL, Kerle in 2014; 14LQ, Linqing in 2014; 15AY, Anyang in 2015; 15ALE, Alaer in 2015; 16AY, Anyang in 2016. *Indicate significance at the 0.05 level. **Indicate significance at the 0.01 level.*

which nine were stable (**Figure 3** and **Supplementary Table S4**). In these stable QTLs, qLP-chr10-1 was identified in five environments, and could explain 4.44–8.80% proportions of the observed PVs. In its marker interval of DPL0468–CGR5624, 148 candidate genes were harbored. qLP-chr04-1 was identified in four environments, and could explain 3.81–4.50% proportions of the observed PVs. In its marker interval of TM9862–TM9831, 217 candidate genes were harbored. qLP-chr26-2 was identified in three environments, and could explain 3.98–5.34% proportions of the observed PVs. In its marker interval of TM77259–TM77267, 3 candidate genes were harbored. qLP-chr03-1, qLP-chr06-2, qLP-chr08-1, qLP-chr11-1, qLP-chr22-1, and qLP-chr25-3 were identified in two environments, and could explain 2.69–2.83, 3.76–6.32, 4.43–6.02, 3.91–4.75, 3.61–4.26, and 4.77–7.64% proportions of the observed PVs, respectively. In their marker intervals of TM6006–TM6010, TM18161–TM18322, TM29470–TM29463, TM39443–TM39427, TM55461–TM55466, and TM63143–TM63142, 1, 141, 26, 12, 16, and 1 candidate genes, respectively, were harbored.

Seed Index

A total of 30 QTLs for SI were identified on 16 chromosomes except chr01, chr14, chr15, chr18, chr21, chr22, chr23, chr24, chr25, and chr26, among which nine were stable (**Figure 3** and **Supplementary Table S4**). In these stable QTLs, qSI-chr07-2 was identified in five environments, which could explain 4.83–28.27% of the observed PVs. In its confidence interval of DPL0852–DPL0757, there harbored 8 candidate genes. qSI-chr16-1 was identified in four environments, which could explain 4.24–6.91% of the observed PVs. In its confidence interval of TM66717–TM66737, there harbored 19 candidate genes. qSI-chr10-1, qSI-chr10-2, and qSI-chr11-2 were identified in three environments, which could explain 6.67–7.83%, 4.28–6.50%, and 4.35–6.01% of the observed PVs, respectively. In their confidence intervals of DPL0468, TM36374–TM36487, and TM37826–TM37828, there harbored 2, 87, and 1 candidate genes, respectively. qSI-chr04-2, qSI-chr07-1, qSI-chr11-3, and qSI-chr13-2 were identified in two environments, which could explain 4.57–5.23%, 5.59–8.50%, 5.52–5.66%, and 3.37–5.29% of the observed PVs, respectively. In their confidence intervals of TM9702–TM9697, TM19691–TM19898, TM37970–TM39953, and TM43247–TM43263, there harbored 8, 39, 73, and 11 candidate genes, respectively.

GWAS for Fiber Quality Traits and Yield Components

A total of 209 and 139 QTNs were identified by four multilocus GWAS methods to be associated with fiber quality and yield component traits, respectively, in the current study (**Supplementary Table S6**). Among these QTNs, 74 were simultaneously found by at least two algorithms or in two environments (**Supplementary Table S6**), each with 0.15–47.17% proportions of the observed PVs explained, and a total of 104 candidate genes were mined.

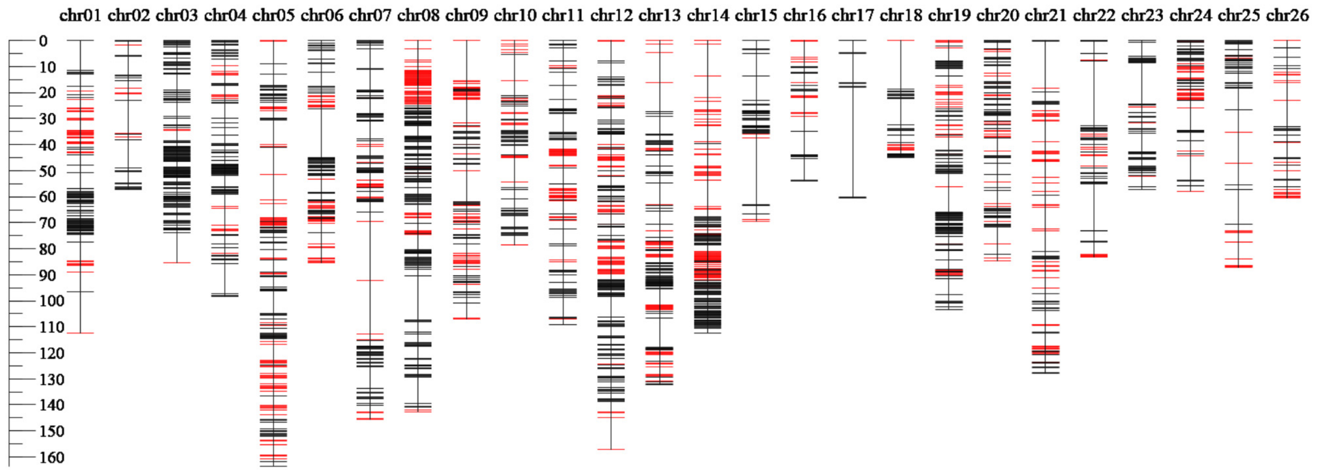


FIGURE 1 | The genetic linkage map constructed by the SNP marker and SSR marker.

TABLE 3 | Detailed information of the genetic map.

Chr.	Number of SSRs	Number of SNPs	Total markers	Total distance (cM)	Average distance (cM)	Largest gap (cM)	Number of gaps (>10 cM)	Number of SDMs[a]	Percentage of SDMs (%)	SDR[b] number	χ^2-value	*P*-value
Chr01	2	234	236	112.38	0.49	15.83	2	90	38.14	3	2.17	0.46
Chr02	1	60	61	57.22	0.97	12.51	2	11	18.03	1	1.56	0.51
Chr03	6	304	310	85.26	0.28	11.59	1	3	0.97	0	0.75	0.56
Chr04	1	225	226	98.49	0.44	11.56	1	22	9.73	3	0.83	0.51
Chr05	15	279	294	163.46	0.57	10.44	1	138	46.94	10	3.93	0.22
Chr06	2	185	187	85.28	0.46	18.77	1	80	42.78	6	2.63	0.36
Chr07	11	220	231	145.52	0.66	22.68	2	44	19.05	5	2.09	0.38
Chr08	9	472	481	142.55	0.32	17.20	2	237	49.27	5	3.69	0.28
Chr09	3	222	225	107.03	0.49	15.63	2	96	42.67	8	2.83	0.25
Chr10	3	100	103	78.66	0.78	10.30	1	36	34.95	6	2.66	0.31
Chr11	2	203	205	109.13	0.57	7.68	0	97	47.32	6	3.06	0.22
Chr12	10	387	397	157.24	0.41	12.57	1	107	26.95	6	2.44	0.42
Chr13	7	337	344	132.40	0.40	11.47	3	100	29.07	7	2.70	0.45
A_t	72	3228	3300	1474.63	0.45	22.68	19	1061	32.15	66	–	–
Chr14	4	315	319	112.26	0.36	12.16	2	161	50.47	9	3.37	0.25
Chr15	1	123	124	69.36	0.64	25.46	1	5	4.03	1	0.79	0.44
Chr16	4	51	55	53.91	0.98	8.83	0	26	47.27	2	2.70	0.26
Chr17	0	19	19	60.60	3.56	42.23	2	0	0.00	0	1.00	0.46
Chr18	1	57	58	45.05	0.79	18.45	2	9	15.52	1	1.39	0.47
Chr19	5	235	240	103.53	0.44	9.79	0	44	18.33	7	1.59	0.40
Chr20	12	133	145	84.74	0.59	6.37	0	32	22.07	4	1.70	0.36
Chr21	7	143	150	127.93	0.87	18.09	1	97	64.67	8	4.15	0.18
Chr22	1	92	93	83.17	0.92	24.89	2	32	34.41	5	2.43	0.31
Chr23	1	93	94	57.14	0.63	15.88	1	6	6.38	1	0.84	0.53
Chr24	7	106	113	57.89	0.52	9.26	0	40	35.40	3	2.46	0.28
Chr25	3	79	82	87.29	1.06	13.29	2	20	24.39	1	2.17	0.38
Chr26	4	55	59	60.49	1.06	10.09	1	30	50.85	2	3.64	0.23
D_t	50	1501	1551	1003.36	0.65	42.23	14	502	32.37	44	–	–
Total	122	4729	4851	2477.99	0.51	42.23	33	1563	32.22	110	–	–

[a] *SDM, segregation distortion marker.* [b] *SDR, segregation distortion region.*

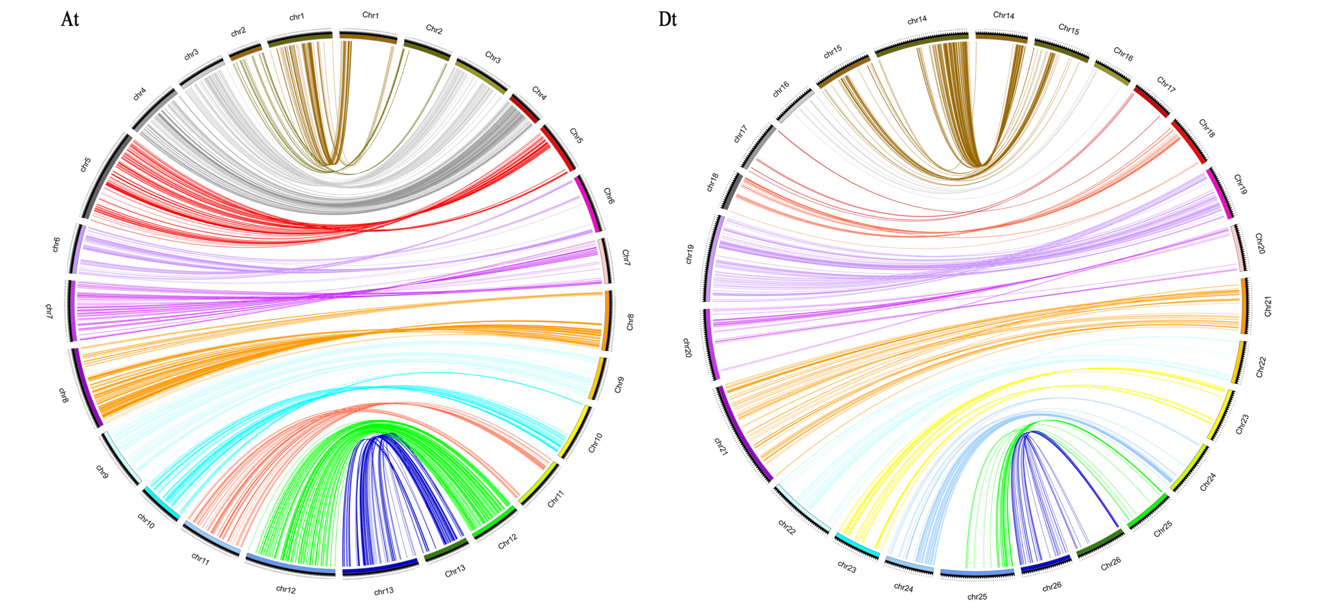


FIGURE 2 | Collinearity between the genetic map (left) and the physical map (right). A_t, collinearity of the A_t subgenome; D_t, collinearity of the D_t subgenome.

Fiber Quality Traits

A total of 68, 65, and 76 QTNs were found to be associated with FL, FS, and FM, respectively, and the corresponding 110, 99, and 126 candidate genes were identified. In these QTNs, 11 for FL, 17 for FS, and 22 for FM were simultaneously associated by at least two algorithms or in two environments, and each could explain 0.15–29.10, 1.43–47.17, and 2.54–41.39% proportions of the observed PVs, respectively.

Yield Components

A total of 51, 50, and 38 QTNs were found to be associated with BW, LP, and SI, respectively, and the corresponding 82, 83, and 65 candidate genes were identified. In these QTNs, 9 for BW, 5 for LP, and 10 for SI were simultaneously associated by at least two algorithms or in two environments, and each could explain 3.41–28.76, 3.00–22.49, and 1.21–38.73% proportions of the observed PVs, respectively.

Candidate Genes Annotation

A total of 2133 candidate genes, among which 621 were for FL, 426 for FS, 510 for FM, 234 for BW, 565 for LP, and 323 for SI, were identified from stable QTL (**Supplementary Table S5**), and 506 candidate genes, among which 110 for FL, 99 for FS, 126 for FM, 82 for BW, 83 for LP, and 65 for SI, were identified from GWAS (**Supplementary Table S6**). Annotation analysis of the 35 common genes from these two candidate gene pools revealed that 33 of them had annotation information, whereas 8 had unknown function (**Supplementary Table S7**). In the gene ontology (GO) analysis of the candidate gene for fiber quality (**Supplementary Table S8**), 24, 17, and 29 candidate genes were identified in the cellular component, molecular function, and biological process category, respectively. In the cellular component category, three main brackets of cell (six genes), cell part (six genes), and organelle (five genes) were enriched, whereas in the molecular function category, two main brackets of binding (eight genes) and catalytic activity (six genes), and in biological process category, four main brackets of metabolic process (seven genes), single-organism process (seven genes), cellular process (five genes), and response to stimulus (five genes) were, respectively, enriched (**Figure 4A**). In gene ontology (GO) analysis of the candidate gene for yield components (**Supplementary Table S10**), 19, 13, and 27 candidate genes were identified in the cellular component, molecular function, and biological process category, respectively. In the cellular component category, three main brackets of cell (five genes), organelle (five genes), and cell part (five genes) were enriched, whereas in the molecular function category, two main brackets of binding (six genes) and catalytic activity (five genes), and in the biological process category, four main brackets of single-organism process (seven genes), metabolic process (five genes), cellular process (five genes), and localization (four genes) were, respectively, enriched (**Figure 4B**). Kyoto encyclopedia of genes and genomes (KEGG) analysis indicated that six candidate genes for fiber quality were involved in 10 pathways and two candidate genes for yield were involved in six pathways (**Supplementary Tables S9**, **S11**).

DISCUSSION

The High-Density Genetic Map Construction and Its Reliability

The development of high-throughput sequencing technology enabled its applications in genotyping the accessions of both natural populations for GWAS and segregating ones for map

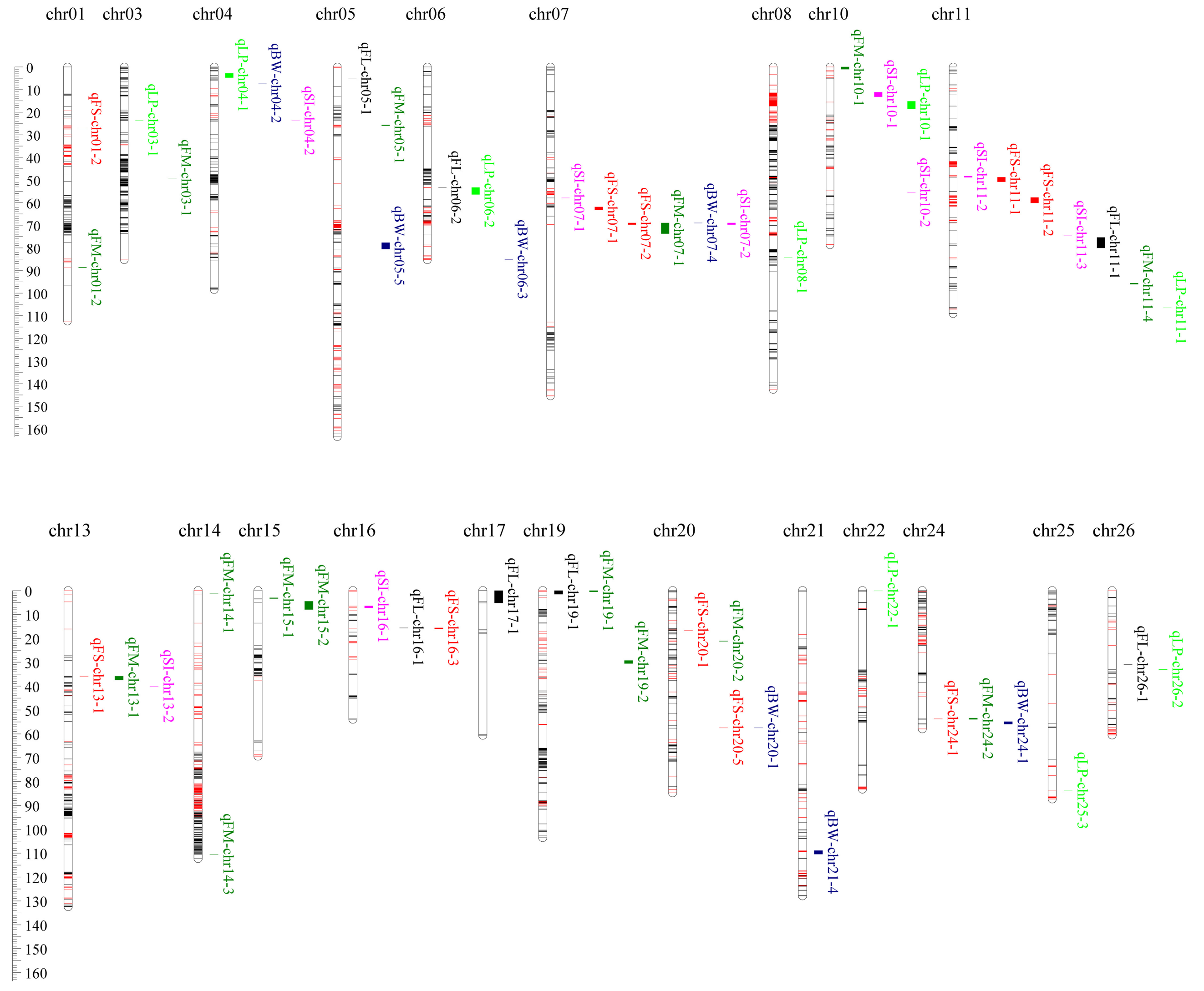


FIGURE 3 | The annotation of the common candidate genes in GO analysis. **(A)** Fiber quality traits. **(B)** Yield components.

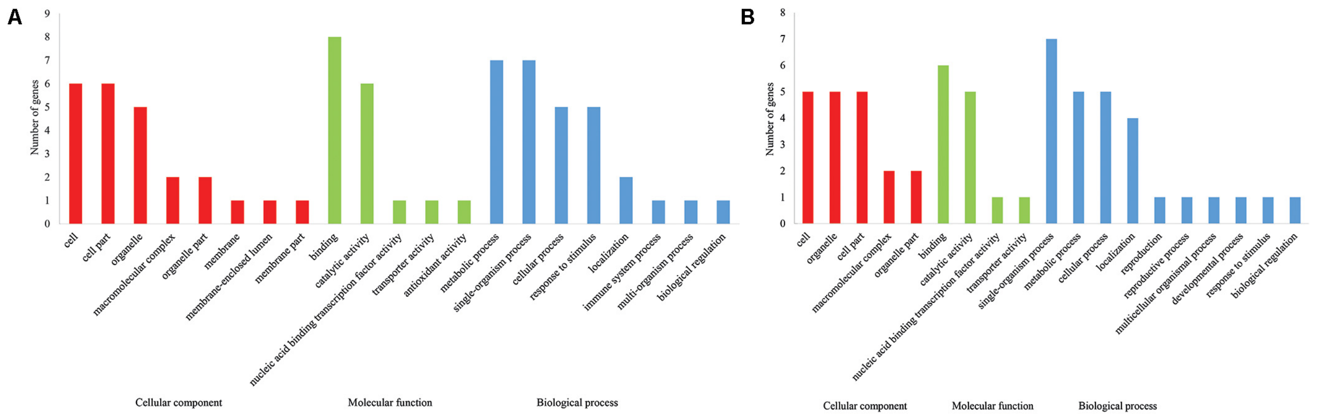


FIGURE 4 | The chromosome-wise distribution of stable QTL for fiber quality traits and yield components.

construction and QTL identification to be accumulated to agricultural important crops (Huang et al., 2010; Kump et al., 2011; Dhanapal et al., 2015; Zeng et al., 2017). SNPs provided abundant genetic variation loci at the genome level and much improved the genome coverage and marker saturation when they were applied to genetic map construction (Agarwal et al., 2008; Hulse-Kemp et al., 2015; Cai et al., 2017). At present, two sets of SNP arrays were developed for *Gossypium* (Hulse-Kemp et al., 2015; Cai et al., 2017). Different from the first set of CottonSNP63K arrays (Hulse-Kemp et al., 2015; Zhang Z. et al., 2017), which was developed by international consortium of several different studies (Hulse-Kemp et al., 2015), the CottonSNP80K array (Cai et al., 2017) was developed from the re-sequencing of 100 upland cotton cultivars and the TM-1 genome database (Zhang et al., 2015). Even though both sets were successfully applied in upland cotton linkage map construction and QTL identifications (Hulse-Kemp et al., 2015; Cai et al., 2017; Zhang Z. et al., 2017; Tan et al., 2018), the second set could have a higher genotyping accuracy, better coverage, and representative of *hirsutum* genome (Cai et al., 2017; Tan et al., 2018). In the current study, a linkage map was constructed mainly using SNP markers from the CottonSNP80K array in combination with SSR ones. The map spanned a total genetic length of 2477.99 cM, containing 122 SSR and 4729 SNP markers, with an average marker interval of 0.51 cM between adjacent markers. Compared with previous SSR maps (Shappley et al., 1998; Shen et al., 2005; Sun F.D. et al., 2012; Wang X. et al., 2015), the current map contained more markers and were more effective in map construction (Liu et al., 2015; Li et al., 2016; Zhang et al., 2016; Zhang Z. et al., 2017; Tan et al., 2018), and exhibited a high consistency with the genomic distribution of the SNP array, which demonstrated its representativeness in map construction (**Figure 2**; Cai et al., 2017).

The reliability of the genetic map is also estimated by gap size, collinearity, and segregation distortion analyses (**Figure 2** and **Table 3**). Although the development of SNP markers was based on the CottonSNP80K array, a few chromosomes still had a large gap or uneven distribution of makers (Li et al., 2016; Zhang Z. et al., 2017). Totally, there were 33 gaps larger than 10 cM, of which the largest one was of 42.23 cM on chr17 and there were only 19 markers mapped on it. The result of collinearity between the genetic map and the *G. hirsutum* (TM-1) reference genome indicated accuracy and quality of the map.

The segregation distortion is recognized as strong evolutionary force in the process of biological evolution (Taylor and Ingvarsson, 2003), which was also a common phenomenon in the study of genetic mapping (Shappley et al., 1998; Ulloa et al., 2002; Jamshed et al., 2016; Zhang et al., 2016; Tan et al., 2018). The current study observed that 32.22% of the total mapping markers were SDMs ($P < 0.05$). The maximum SDMs were on chr08, where there were 237 SDMs of the total 481 markers, forming five SDRs (**Figure 1**). This was in consistency with the SSR map constructed from the F_2 population of the same parents of the current study (Kong et al., 2011). However, some studies observed an increase of the SDM ratio from F_2 generation to the completion of RILs (Tan et al., 2018). This phenomenon was influenced by plenty of factors, including genetic drift (Shen et al., 2007) of mapping population, pollen tube competition, preferential fertilization of particular gametic genotypes, and others (Zhang et al., 2016; Zhang Z. et al., 2017; Tan et al., 2018). In the current study, some chromosomal uneven distribution of QTLs in SDR versus normal regions was also observed in chr01, chr06, chr07, chr10, chr16, chr19, and chr20. These facts implied an impact of the selections being imposed during the construction of the RIL population.

Linkage and Association Analyses for Fiber Quality Traits and Yield Components

The QTLs detected in this study were compared with those in previous studies. As a result, 22 QTLs for FL, 25 QTLs for FS, 31 QTLs for FM, 36 QTLs for BW, 22 QTLs for LP, and 19 QTLs for SI in this study were coincided in the same physical regions of QTLs identified in previous studies (indicated with asterisks in **Supplementary Table S4**). The remaining could possibly be novel QTLs, of which 21 were stable ones, namely qFL-chr11-1, qFL-chr16-1, qFL-chr19-1, qFL-chr26-1, qFS-chr01-2, qFS-chr16-3, qFS-chr20-1, qFS-chr20-5, qFM-chr01-2, qFM-chr03-1, qFM-chr10-1, qFM-chr14-1, qFM-chr19-1, qBW-chr20-1, qLP-chr03-1, qLP-chr22-1, qLP-chr25-3, qLP-chr26-2, qSI-chr07-1, qSI-chr10-1, and qSI-chr16-1. Even though in the phenotypic evaluations of the population, the phenotypic differences between the two parents did not reach the significant level except that of FS, transgressive segregation in the RILs and significant differences among RILs indicated that the parents might harbor different favorable alleles for the target traits. QTL identification results well illustrated such presuppositions as these different favorable alleles contributed greatly to the similarity or nonsignificant differences between the two parents. These alleles could be addressed through map construction and detected in QTL identification. The high heritability of the target traits also enhanced the reliability of the QTL identification.

In addition, four multilocus GWAS algorithms were applied to the association of QTNs with the target traits, and their results were compared with the previous identified QTLs (Said et al., 2015). The results confirmed that quite a ratio of QTNs were coincided in the physical regions of the confidence intervals of reported QTLs in the database, namely 43 QTNs for FL, 44 QTNs for FS, 51 QTNs for FM, 40 QTNs for BW, 34 QTNs for LP, and 25 QTNs for SI (indicated with asterisks in **Supplementary Table S6**). The remaining QTNs could possibly be novel QTNs, of which 27 were associated by at least two algorithms or in two environments. These loci could be of great significance for cotton molecular-assisted breeding, particularly the loci of TM9941 and TM54893, which were identified both by multiple algorithms and in multiple environments for more than one target trait.

Based on linkage disequilibrium, GWAS is an effective genetic analysis method to dissect the genetic foundation of complex traits in plants in natural populations. The four multilocus

GWAS algorithms provided promising alternatives in GWAS. Usually, GWAS needed a large panel size with sufficient marker polymorphism (Bodmer and Bonilla, 2008; Manolio et al., 2009), and was effective to identify major loci while ineffective to rare or polygenes (Asimit and Zeggini, 2010; Gibson, 2012) in the population. Linkage analysis in segregating populations could effectively eliminate the false-positive results, which was a built-in defect of GWAS in natural populations. But linkage analysis usually identified large DNA fragments, which made it difficult to further study the initial identification results. In the current study, both GWAS and linkage analysis were applied in the segregating RILs to study the correlations between genotypes and phenotypes. When comparing the results of GWAS to the QTLs of both previous studies (Said et al., 2015) and current study, common loci (genes) (**Supplementary Table S7**) demonstrated the effectiveness and feasibility of multilocus GWAS methods to address the correlation between genotypes and phenotypes in segregating RILs. Especially under the condition of increased marker density and improved genome coverage, the accuracy of QTN identification in GWAS would also increase. The increased accuracy probably rendered the application of GWAS in segregating population to have a higher effect on the observed PVs, sometimes even higher than that of QTL on the PVs in linkage analysis, which was usually low in natural populations.

Congruency and Function Analysis of Candidate Genes

In this study, candidate genes were identified independently both from the physical region in the marker intervals of the QTLs, which were identified by CIM (Zeng, 1994) in WinQTL Cartographer 2.5 (Wang et al., 2012), and from the physical position of the QTNs, which were associated by multilocus GWAS algorithms. As the CIM algorithm gave not only the QTL position where the highest LOD value located, but also a marker interval of that QTL, the physical regions where the marker interval resided by QTL/QTN were used to search the candidate genes around the QTLs. To avoid redundant genes, the markers, which resided far away from the physical positions of the rest in the same confidence interval, were discarded for consideration of candidate gene searching. This increased the accuracy of the functional analysis of the candidate genes harbored in the confidence intervals of QTLs.

When comparing both candidate gene lists, even if they were not completely consistent, they still revealed a good congruency of candidate gene identification from both algorithms of QTL/QTN; namely, three congruent candidate genes for FL, seven for FS, nine for FM, five for BW, eight for LP, and nine for SI were identified (**Supplementary Table S7**). Further analysis of these candidate genes indicated that 1 for FL, 17 for FS, and 2 for FM (indicated with asterisks in **Supplementary Table S6**) were congruent with some previous reports (Huang et al., 2017; Sun et al., 2017). Two candidate genes, *Gh_D102255* (a protein kinase superfamily gene) and *Gh_A13G0187* (*actin 1* gene), which were for fiber quality, were also reported to participate in fiber elongation (Li et al., 2005; Huang et al., 2008). *Gh_A07G1730* and *Gh_D03G0236* belonged to a WD40 protein superfamily were mainly involved in yield formation in the current study, and might be related to a series of functions (Sun Q. et al., 2012; Gachomo et al., 2014). *Gh_D11G1653* (myb domain protein 6) functioned in BW formation, whereas reports indicated that several members of MYB family were involved in fiber development (Suo et al., 2003; Machado et al., 2009; Sun et al., 2015; Huang et al., 2016). Findings in the current study also indicated that some candidate genes could possibly be "pleiotropic," namely *Gh_A07G1744* for FS, FM, and SI; *Gh_A07G1745* for FS and FM; *Gh_A07G1743* for BW and SI; and *Gh_D08G0430* for FM and BW. These candidate genes could be of great significance for further studies including functional gene cloning as well as cultivar development.

CONCLUSION

The enriched high-density genetic map, which contained 4729 SNP and 122 SSR markers, spanned 2477.99 cM with a marker density of 0.51 cM between adjacent markers. A total of 134 QTLs for fiber quality traits and 122 for yield components were identified by the CIM, of which 57 are stable. A total of 209 and 139 QTNs for fiber quality traits and yield components were, respectively, associated by four multilocus GWAS algorithms, of which 74 QTNs were detected by at least two algorithms or in two environments. Comparing the candidate genes harbored in 57 stable QTLs with those associated with the QTN, 35 were found to be congruent, 4 of which were possibly "pleiotropic." Results in the study could be promising for future breeding practices through MAS and candidate gene functional studies.

AUTHOR CONTRIBUTIONS

WG and YY initiated the research. WG, RL, and QC designed the experiments. RL, XX, and ZZ performed the molecular experiments. JG, JL, AL, HS, YS, QG, QL, MI, XD, SL, JP, LD, QZ, XJ, XZ, and AH conducted the phenotypic evaluations and collected the data from the field. RL, WG, YY, and HG performed the analysis. RL drafted the manuscript. YY and WG finalized the manuscript. All authors contributed in the interpretation of results and approved the final manuscript.

FUNDING

This work was funded by the National Key R&D Program of China (2016YFD0100500), the Fundamental Research Funds for Central Research Institutes (Y2017JC48), the National Key R&D Program of China (2017YFD0101600 and 2016YFD0101401), the Natural Science Foundation of China (31371668, 31471538), the National High Technology Research and Development Program of China (2012AA101108), and the National Agricultural Science and Technology Innovation Project for CAAS and the Henan province foundation with cutting-edge technology research projects (14230041 3202).

ACKNOWLEDGMENTS

The authors thank the Biomarker Technologies Corporation (Beijing, China) for providing the software Highmap and help in the genetic map construction with Highmap.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01067/full#supplementary-material

REFERENCES

Abdurakhmonov, I. Y., Kohel, R. J., Yu, J. Z., Pepper, A. E., Abdullaev, A. A., Kushanov, F. N., et al. (2008). Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum L.* germplasm. *Genomics* 92, 478–487. doi: 10.1016/j.ygeno.2008.07.013

Agarwal, M., Shrivastava, N., and Padh, H. (2008). Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep.* 27, 617–631. doi: 10.1007/s00299-008-0507-z

Asimit, J., and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* 44, 293–308. doi: 10.1146/annurev-genet-102209-163421

Atwell, S., Huang, Y., Vilhjálmsson, B., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465, 627–631. doi: 10.1038/nature08800

Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* 40, 695–701. doi: 10.1038/ng.f.136

Cai, C., Ye, W., Zhang, T., and Guo, W. (2014). Association analysis of fiber quality traits and exploration of elite alleles in upland cotton cultivars/accessions (*Gossypium hirsutum* L.). *J. Integr. Plant Biol.* 56, 51–62. doi: 10.1111/jipb.12124

Cai, C., Zhu, G., Zhang, T., and Guo, W. (2017). High-density 80K SNP array is a powerful tool for genotyping *G. hirsutum,* accessions and genome analysis. *BMC Genomics* 18:654. doi: 10.1186/s12864-017-4062-2

Chen, Z. J., Scheffler, B. E., Dennis, E., Triplett, B. A., Zhang, T., Guo, W., et al. (2007). Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* 145, 1303–1310. doi: 10.1104/pp.107.107672

Dhanapal, A. P., Ray, J. D., Singh, S. K., Hoyos-Villegas, V., Smith, J. R., Purcell, L. C., et al. (2015). Genome-wide association study (GWAS) of carbon isotope ratio (δ13C) in diverse soybean [Glycine max (L.) Merr.]. genotypes. *Theor. Appl. Genet.* 128, 73–91. doi: 10.1007/s00122-014-2413-9

Eizenga, G. C., Agrama, H. A., Lee, F. N., Yan, W., and Jia, Y. (2006). Identifying novel resistance genes in newly introduced blast resistant rice germplasm. *Crop Sci.* 46, 1870–1878. doi: 10.2135/cropsci2006.0143

Fang, D. D., Jenkins, J. N., Deng, D. D., McCarty, J. C., Li, P., and Wu, J. (2014). Quantitative trait loci analysis of fiber quality traits using a random-mated recombinant inbred population in upland cotton (*Gossypium hirsutum*, L.). *BMC Genomics* 15:397. doi: 10.1186/14712164-15-397

Fernandes, E. G., Lombardi, A., Solaro, R., and Chiellini, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* 44, 841–847. doi: 10.1038/ng.2355

Flint-Garcia, S. A., Thuillet, A. C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., et al. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* 44, 1054–1064. doi: 10.1111/j.1365-313X.2005.02591.x

Gachomo, E. W., Jimenez-Lopez, J. C., Baptiste, L. J., and Kotchoni, S. O. (2014). GIGANTUS1 (GTS1), a member of Transducin/WD40 protein superfamily, controls seed germination, growth and biomass accumulation through ribosome-biogenesis protein interactions in *Arabidopsis thaliana*. *BMC Plant Biol.* 14:37. doi: 10.1186/1471-2229-14-37

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145. doi: 10.1038/nrg3118

Horton, M. W., Hancock, A. M., Huang, Y. S., Toomajian, C., Atwell, S., Auton, A., et al. (2012). Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the regmap panel. *Nat. Genet.* 44, 212–216. doi: 10.1038/ng.1042

Huang, C., Nie, X., Shen, C., You, C., Li, W., Zhao, W., et al. (2017). Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. *Plant Biotechnol. J.* 15, 1374–1386. doi: 10.1111/pbi.12722

Huang, J., Chen, F., Wu, S., Li, J., and Xu, W. (2016). Cotton GhMYB7 is predominantly expressed in developing fibers and regulates secondary cell wall biosynthesis in transgenic *Arabidopsis*. *Sci. China Life Sci.* 59, 194–205. doi: 10.1007/s11427-015-4991-4

Huang, Q. S., Wang, H. Y., Gao, P., Wang, G. Y., and Xia, G. X. (2008). Cloning and characterization of a calcium dependent protein kinase gene associated with cotton fiber development. *Plant Cell Rep.* 27, 1869–1875. doi: 10.1007/s00299-008-0603-0

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967. doi: 10.1038/ng.695

Hulse-Kemp, A. M., Jana, L., Joerg, P., Ashrafi, H., Buyyarapu, R., Fang, D. D., et al. (2015). Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium* spp. *G3* 5, 1187–1209. doi: 10.1534/g3.115.018416

Jamshed, M., Jia, F., Gong, J., Palanga, K. K., Shi, Y., Li, J., et al. (2016). Identification of stable quantitative trait loci (QTLs) for fiber quality traits across multiple environments in *Gossypium hirsutum* recombinant inbred line population. *BMC Genomics* 17:197. doi: 10.1186/s12864-016-2560-2

Jia, Y., Sun, X., Sun, J., Pan, Z., Wang, X., He, S., et al. (2014). Association mapping for epistasis and environmental interaction of yield traits in 323 cotton cultivars under 9 different environments. *PLoS One* 9:e95882. doi: 10.1371/journal.pone.0095882

Jiang, F., Zhao, J., Zhou, L., Guo, W. Z., and Zhang, T. Z. (2009). Molecular mapping of *Verticillium, wilt* resistance QTL clustered on chromosomes D7 and D9 in upland cotton. *Sci. China* 52, 872–884. doi: 10.1007/s11427-009-0110-8

Kantartzi, S. K., and Stewart, J. M. (2008). Association analysis of fibre traits in *Gossypium arboreum*, accessions. *Plant Breed.* 127, 173–179. doi: 10.1111/j.1439-0523.2008.01490.x

Kong, F., Li, J., Gong, J., Shi, Y., Liu, R., Shang, H., et al. (2011). QTL mapping for lint percentage and seed index in upland cotton (*Gossypium hirsutum* L.) of different genetic backgrounds. *Chin. Agric. Sci. Bull.* 27, 104–109. doi: 10.1007/s00438-015-1027-5

Kosambi, D. D. (1943). The estimation of map distances from recombination values. *Ann. Hum. Genet.* 12, 172–175. doi: 10.1111/j.1469-1809.1943.tb02321.x

Kump, K. L., Bradbury, P. J., Wisser, R. J., Buckler, E. S., Belcher, A. R., and Oropeza-Rosas, M. A. (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* 43, 163–168. doi: 10.1038/ng.747

Lacape, J. M., Jacobs, J., Arioli, T., Derijcker, R., Forestier-Chiron, N., Llewellyn, D., et al. (2009). A new interspecific Gossypium hirsutum × G. barbadense, RIL population: towards a unified consensus linkage map of tetraploid cotton. *Theor. Appl. Genet.* 119, 281–292. doi: 10.1007/s00122-009-1037-y

Levi, A., Paterson, A. H., Cakmak, I., and Saranga, Y. (2011). Metabolite and mineral analyses of cotton near-isogenic lines introgressed with QTLs for productivity and drought-related traits. *Physiol. Plant.* 141, 265–275. doi: 10.1111/j.1399-3054.2010.01438.x

Li, C., Dong, Y., Zhao, T., Li, L., Li, C., Yu, E., et al. (2016). Genome-wide SNP linkage mapping and QTL analysis for fiber quality and yield traits in the upland cotton recombinant inbred lines population. *Front. Plant Sci.* 7:1356. doi: 10.3389/fpls.2016.01356

Li, C., Wang, C., Dong, N., Wang, X., Zhao, H., Converse, R., et al. (2012). QTL detection for node of first fruiting branch and its height in upland cotton (*Gossypium hirsutum* L.). *Euphytica* 188, 441–451. doi: 10.1007/s10681-012-0720-2

Li, C., Wang, X., Dong, N., Zhao, H., Xia, Z., Wang, R., et al. (2013). QTL analysis for early-maturing traits in cotton using two upland cotton (*Gossypium hirsutum* L.) crosses. *Breed. Sci.* 63, 154–163. doi: 10.1270/jsbbs.63.154

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R. J., et al. (2015). Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* 33, 524–530. doi: 10.1038/nbt.3208

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., et al. (2014). Genome sequence of the cultivated cotton *Gossypium arboreum*. *Nat. Genet.* 46, 567–572. doi: 10.1038/ng.2987

Li, X. B., Fan, X. P., Wang, X. L., Cai, L., and Yang, W. C. (2005). The cotton ACTIN1 gene is functionally expressed in fibers and participates in fiber elongation. *Plant Cell* 17, 859–875. doi: 10.1105/tpc.104.029629

Liu, D., Liu, F., Shan, X., Zhang, J., Tang, S., Fang, X., et al. (2015). Construction of a high-density genetic map and lint percentage and cottonseed nutrient trait QTL identification in Upland cotton (*Gossypium hirsutum* L.). *Mol. Genet. Genomics* 290, 1683–1700. doi: 10.1007/s00438-015-1027-5

Liu, D., Ma, C., Hong, W., Huang, L., Liu, M., and Liu, H. (2014). Construction and analysis of high-density linkage map using high-throughput sequencing data. *PLoS One* 9:e98855. doi: 10.1371/journal.pone.0098855

Liu, X., Teng, Z., Wang, J., Wu, T., Zhang, Z., Deng, X., et al. (2017). Enriching an intraspecific genetic map and identifying QTL for fiber quality and yield component traits across multiple environments in Upland cotton (*Gossypium hirsutum*, L.). *Mol. Genet. Genomics* 292, 1281–1306. doi: 10.1007/s00438-017-1347-8

Maccaferri, M., Sanguineti, M. C., Noli, E., and Tuberosa, R. (2005). Population structure and long-range linkage disequilibrium in a durum wheat elite collection. *Mol. Breed.* 15, 271–290. doi: 10.1007/s11032-004-7012-z

Machado, A., Wu, Y., Yang, Y., Llewellyn, D. J., and Dennis, E. S. (2009). The MYB transcription factor GhMYB25 regulates early fiber and trichome development. *Plant J.* 59, 52–62. doi: 10.1111/j.1365-313X.2009.03847.x

Manolio, T. A., Collins, F. S., Cox, N. J., Goldetein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Mei, H., Zhu, X., and Zhang, T. (2013). Favorable QTL alleles for yield and its components identified by association mapping in Chinese upland cotton cultivars. *PLoS One* 8:e82193. doi: 10.1371/journal.pone.0082193

Nie, X., Huang, C., You, C., Wu, L., Zhao, W., Shen, C., et al. (2016). Genome-wide SSR-based association mapping for fiber quality in nation-wide upland cotton inbreed cultivars in China. *BMC Genomics* 17:352. doi: 10.1186/s12864-016-2662-x

Palanga, K. K., Jamshed, M., Rashid, M. H., Gong, J., Li, J., Iqbal, M. S., et al. (2017). Quantitative trait locus mapping for *Verticillium wilt* resistance in an upland cotton recombinant inbred line using SNP-based high density genetic map. *Front. Plant Sci.* 8:382. doi: 10.3389/fpls.2017.00382

Qi, H., Wang, N., Qiao, W., Xu, Q., Zhou, H., Shi, J., et al. (2017). Construction of a high-density genetic map using genotyping by sequencing (GBS) for quantitative trait loci (QTL) analysis of three plant morphological traits in upland cotton (*Gossypium hirsutum*, L.). *Euphytica* 213:83. doi: 10.1007/s10681-017-1867-7

Said, J. I., Knapka, J. A., Song, M., and Zhang, J. (2015). Cotton QTLdb: a cotton QTL database for QTL analysis, visualization, and comparison between *Gossypium hirsutum* and *G. hirsutum× G. barbadense* populations. *Mol. Genet. Genomics* 290, 1615–1625. doi: 10.1007/s00438-015-1021-y

Said, J. I., Lin, Z., Zhang, X., Song, M., and Zhang, J. (2013). A comprehensive meta QTL analysis for fiber quality, yield, yield related and morphological traits, drought tolerance, and disease resistance in tetraploid cotton. *BMC Genomics* 14:776. doi: 10.1186/1471-2164-14-776

Samayoa, L., Malvar, R., Olukolu, B. A., Holland, J. B., and Butrón, A. (2015). Genome-wide association study reveals a set of genes associated with resistance to the Mediterranean corn borer (*Sesamia nonagrioides* L.) in a maize diversity panel. *BMC Plant Biol.* 15:35. doi: 10.1186/s12870-014-0403-3

Segura, V., Vilhjalmsson, B. J., Platt, A., Korte, A., Seren, Ü, Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

Shappley, Z. W., Jenkins, J. N., Meredith, W. R., and McCarty, J. C. Jr. (1998). An RFLP linkage map of upland cotton, *Gossypium hirsutum* L. *Theor. Appl. Genet.* 97, 756–761. doi: 10.1007/s001220050952

Shen, X., Guo, W., Lu, Q., Zhu, X., Yuan, Y., and Zhang, T. (2007). Genetic mapping of quantitative trait loci for fiber quality and yield trait by RIL approach in upland cotton. *Euphytica* 155, 371–380. doi: 10.1007/s10681-006-9338-6

Shen, X., Guo, W., Zhu, X., Yuan, Y., Yu, J. Z., Kohel, R. J., et al. (2005). Molecular mapping of QTLs for fiber qualities in three diverse lines in Upland cotton using SSR markers. *Mol. Breed.* 15, 169–181. doi: 10.1007/s11032-004-4731-0

Song, G. L., Cui, R. X., Wang, K. B., Guo, L. P., Li, S. H., Wang, C. Y., et al. (1998). A rapid improved CTAB method for extraction of cotton genomic DNA. *Acta Gossypii Sin.* 1998, 50–52.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., et al. (2015). Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genet.* 11:e1004982. doi: 10.1371/journal.pgen.1004982

Stiller, W. N., Reid, P. E., and Constable, G. A. (2004). Maturity and leaf shape as traits influencing cotton cultivar adaptation to dryland conditions. *Agron. J.* 96, 656–664. doi: 10.2134/agronj2004.0656

Su, J., Pang, C., Wei, H., Li, L., Liang, B., Wang, C., et al. (2016). Identification of favorable SNP alleles and candidate genes for traits related to early maturity via GWAS in upland cotton. *BMC Genomics* 17:687. doi: 10.1186/s12864-016-2875-z

Sun, F. D., Zhang, J. H., Wang, S. F., Gong, W. K., Shi, Y. Z., Liu, A. Y., et al. (2012). QTL mapping for fiber quality traits across multiple generations and environments in upland cotton. *Mol. Breed.* 30, 569–582. doi: 10.1007/s11032-011-9645-z

Sun, Q., Cai, Y., Zhu, X., He, X., Jiang, H., and He, G. (2012). Molecular cloning and expression analysis of a new WD40 repeat protein gene in upland cotton. *Biologia* 67, 1112–1118. doi: 10.2478/s11756-012-0103-0

Sun, X., Gong, S. Y., Nie, X. Y., Li, Y., Li, W., Huang, G. Q., et al. (2015). A R2R3-MYB transcription factor that is specifically expressed in cotton (*Gossypium hirsutum*) fibers affects secondary cell wall biosynthesis and deposition in transgenic *Arabidopsis*. *Physiol. Plant.* 154, 420–432. doi: 10.1111/ppl.12317

Sun, Z., Wang, X., Liu, Z., Gu, Q., Zhang, Y., Li, Z., et al. (2017). Genome-wide association study discovered genetic variation and candidate genes of fibre quality traits in *Gossypium hirsutum* L. *Plant Biotechnol. J.* 15, 982–996. doi: 10.1111/pbi.12693

Suo, J., Liang, X., Pu, L., Zhang, Y., and Xue, Y. (2003). Identification of GhMYB109 encoding a R2R3 MYB transcription factor that expressed specifically in fiber initials and elongating fibers of cotton (*Gossypium hirsutum* L.). *Biochim. Biophys. Acta* 1630, 25–34. doi: 10.1016/j.bbaexp.2003.08.009

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tan, Z., Fang, X., Tang, S., Zhang, J., Liu, D., Teng, Z., et al. (2015). Genetic map and QTL controlling fiber quality traits in upland cotton (*Gossypium hirsutum*, L.). *Euphytica* 203, 615–628. doi: 10.1007/s10681-014-1288-9

Tan, Z., Zhang, Z., Sun, X., Li, Q., Sun, Y., Yang, P., et al. (2018). Genetic map construction and fiber quality QTL mapping using the CottonSNP80K array in upland cotton. *Front. Plant Sci.* 9:225. doi: 10.3389/fpls.2018.00225

Tang, F., and Xiao, W. (2014). Genetic association of within-boll yield components and boll morphological traits with fibre properties in upland cotton (*Gossypium hirsutum* L.). *Plant. Breed.* 133, 521–529. doi: 10.1111/pbr.12176

Taylor, D. R., and Ingvarsson, P. K. (2003). Common features of segregation distortion in plants and animals. *Genetica* 117, 27–35. doi: 10.1023/A:1022308414864

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler, E. S. IV (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nat. Genet.* 28, 286–289. doi: 10.1038/90135

Ulloa, M., Hutmacher, R. B., Roberts, P. A., Wright, S. D., Nichols, R. L., and Michael Davis, R. (2013). Inheritance and QTL mapping of *Fusarium wilt* race 4 resistance in cotton. *Theor. Appl. Genet.* 126, 1405–1418. doi: 10.1007/s00122-013-2061-5

Ulloa, M., Meredith, W. R. Jr., Shappley, Z. W., and Kahler, A. L. (2002). RFLP genetic linkage maps from four F2:3 populations and a joinmap of *Gossypium hirsutum* L. *Theor. Appl. Genet.* 104, 200–208. doi: 10.1007/s001220100739

Voorrips, R. E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93, 77–78. doi: 10.1093/jhered/93.1.77

Wang, H., Huang, C., Guo, H., Li, X., Zhao, W., Dai, B., et al. (2015). QTL mapping for fiber and yield traits in upland cotton under multiple environments. *PLoS One* 10:e0130742. doi: 10.1371/journal.pone.0130742

Wang, S., Basten, C. J., and Zeng, Z. B. (2012). *Windows QTL Cartographer 2.5. Raleigh: Department of Statistics, North Carolina State University*. Available at: http://statgen.ncsu.edu/qtlcart/WQTLCart.htm

Wang, S. B., Feng, J. Y., Ren, W. W., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wang, X., Yu, K., Li, H., Peng, Q., Chen, F., Zhang, W., et al. (2015). High-density SNP map construction and QTL identification for the apetalous character in *Brassica napus* L. *Front. Plant Sci.* 6:1164. doi: 10.3389/fpls.2015.01164

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* doi: 10.1093/bib/bbw145 [Epub ahead of print].

Wendel, J. F., and Grover, C. E. (2015). "Taxonomy and evolution of the cotton genus. *Gossypium*," *Cotton*, eds D. D. Fang.and R. G. Percy (Madison, WI: American Society of Agronomy Inc.), 25–44. doi: 10.2134/agronmonogr57.2013.0020

Xia, Z., Zhang, X., Liu, Y. Y., Jia, Z. F., Zhao, H. H., Li, C. Q., et al. (2014). Major gene identiifcation and quantitative trait locus mapping for yield-related traits in upland cotton (*Gossypium hirsutum* L.). *J. Integr. Agric.* 13, 299–309. doi: 10.1016/S2095-3119(13)60508-0

Xu, P., Cao, Z., Zhang, X., Gao, J., Zhang, X., and Shen, X. (2014). Identifcation of quantitative trait loci for fiber quality properties on homoeologous chromosomes 13 and 18 of *Gossypium klotzschianum*. *Crop Sci.* 54, 484–491. doi: 10.2135/cropsci2013.01.0013

Yang, X., Wang, Y., Zhang, G., Wang, X., Wu, L., Ke, H., et al. (2016). Detection and validation of one stable fiber strength QTL on c9 in tetraploid cotton. *Mol. Genet. Genomics* 291, 1625–1638. doi: 10.1007/s00438-016-1206-z

Zeng, A., Chen, P., Korth, K., Hancock, F., Pereira, A., Brye, K., et al. (2017). Genome-wide association study (GWAS) of salt tolerance in worldwide soybean germplasm lines. *Mol. Breed.* 37:30. doi: 10.1007/s11032-017-0634-8

Zeng, L., Meredith, W. R. Jr., Gutiérrez, O. A., and Boykin, D. L. (2009). Identification of associations between SSR markers and fiber traits in an exotic germplasm derived from multiple crosses among Gossypium tetraploid species. *Theor. Appl. Genet.* 119, 93–103. doi: 10.1007/s00122-009-1020-7

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* 136, 1457–1468.

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical bayes for multilocus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537. doi: 10.1038/nbt.3207

Zhang, T., Qian, N., Zhu, X., Chen, H., Wang, S., Mei, H., et al. (2013). Variations and transmission of QTL alleles for yield and fiber qualities in Upland cotton cultivars developed in China. *PLoS One* 8:e57220. doi: 10.1371/journal.pone.0057220

Zhang, Z., Ge, Q., Liu, A., Li, J., Gong, J., Shang, H., et al. (2017). Construction of a high-density genetic map and its application to QTL identification for fiber strength in Upland cotton. *Crop Sci.* 57, 774–788. doi: 10.2135/cropsci2016.06.0544

Zhang, Z., Shang, H., Shi, Y., Huang, L., Li, J., Ge, Q., et al. (2016). Construction of a high-density genetic map by specific locus amplified fragment sequencing (SLAF-seq) and its application to quantitative trait loci (QTL) analysis for boll weight in upland cotton (*Gossypium hirsutum*). *BMC Plant Biol.* 16:79. doi: 10.1186/s12870-016-0741-4

Zhang, Z. S., Xiao, Y. H., Ming, L., Li, X. B., Luo, X. Y., Hou, L., et al. (2005). Construction of a genetic linkage map and QTL analysis of fiber-related traits in upland cotton (*Gossypium hirsutum* L.). *Euphytica* 144, 91–99. doi: 10.1007/s10681-005-4629-x

Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* 2:467. doi: 10.1038/ncomms1467

Zhao, Y., Wang, H., Chen, W., and Li, Y. (2014). Genetic structure, linkage disequilibrium and association mapping of *Verticillium wilt* resistance in elite cotton (*Gossypium hirsutum* L.) germplasm population. *PLoS One* 9:e86308. doi: 10.1371/journal.pone.0086308

Zhu, C., Gore, M. A., Buckler, E., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome* 1, 5–20. doi: 10.3835/plantgenome2008.02.0089

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 02 October 2018
doi: 10.3389/fpls.2018.01405

Deciphering the Genetic Architecture of Cooked Rice Texture

*Gopal Misra, Saurabh Badoni, Cyril John Domingo, Rosa Paula O. Cuevas, Cindy Llorente, Edwige Gaby Nkouaya Mbanjo and Nese Sreenivasulu**

International Rice Research Institute, Metro Manila, Philippines

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Umesh K. Reddy,
West Virginia State University,
United States
Tian Qing Zheng,
Institute of Crop Sciences (CAAS),
China
Anca Macovei,
University of Pavia, Italy

****Correspondence:***
Nese Sreenivasulu
n.sreenivasulu@irri.org

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *20 June 2018*
Accepted: *05 September 2018*
Published: *02 October 2018*

Citation:
Misra G, Badoni S, Domingo CJ, Cuevas RPO, Llorente C, Mbanjo EGN and Sreenivasulu N (2018) Deciphering the Genetic Architecture of Cooked Rice Texture. Front. Plant Sci. 9:1405. doi: 10.3389/fpls.2018.01405

The textural attributes of cooked rice determine palatability and consumer acceptance. Henceforth, understanding the underlying genetic basis is pivotal for the genetic improvement of preferred textural attributes in breeding programs. We characterized diverse set of 236 *Indica* accessions from 37 countries for textural attributes, which includes adhesiveness (ADH), hardness (HRD), springiness (SPR), and cohesiveness (COH) as well as amylose content (AC). A set of 147,692 high quality SNPs resulting from genotyping data of 700K high Density Rice Array (HDRA) derived from the *Indica* diversity panels of 218 lines were retained for marker-trait associations of textural attributes using single-locus (SL) genome wide association studies (GWAS) which resulted in identifying hotspot on chromosome 6 for AC and ADH attributes. Four independent multi-locus approaches (ML-GWAS) including FASTmrEMMA, pLARmEB, mrMLM, and ISIS_EM-BLASSO were implemented to dissect additional loci of major/minor effects influencing the rice texture and to overcome limitations of SL-based GWAS approach. In total 224 significant quantitative trait nucleotide (QTNs) were identified using ML-GWAS, of which 97 were validated with at least two out of the four multi-locus methods. The GWAS results were in accordance with the very significant negative correlation ($r = -0.83$) observed between AC and ADH, and the significant correlation exhibited by AC ($r < 0.4$) with HRD, SPR, and COH. The novel haplotypes and putative candidate genes influencing textural properties beyond AC will be a useful resource for deployment into the marker assisted program to capture consumer preferences influencing rice texture and palatability.

Keywords: rice texture, multi-locus GWAS, quantitative trait nucleotide, amylose content, adhesiveness, cohesiveness, springiness, hardness

INTRODUCTION

Texture is an important attribute of consumer's acceptance criteria and thus quality evaluation of texture is a critical step in breeding. High range of phenotypic textural variability exist in rice which are described as sticky, soft, dry, firm, and hard. These textural attributes appeals to be of interest to different segment of rice consumers across the globe (Kaosa-Ard and Juliano, 1991). For instance, Japanese and Chinese like the soft or sticky short rice grain, while consumers from Middle East, United States and the Latin America prefer non-sticky firm rice. While the widely preferred target of consumers of India and Pakistan favor soft and fluffy rice varieties (Lyon et al., 2000; Champagne et al., 2010), consumers from Bangladesh, Sri Lanka, Indonesia and Myanmar prefer hard textured rice varieties.

The texture of rice is primarily influenced by the structure and composition of starch in the rice grain. Cooking and eating quality are among the most important components of grain quality routinely assessed in rice breeding programs by three main physicochemical characteristics such as amylose content (AC), gel consistency (GC), and gelatinization temperature (GT) (Ramesh et al., 2000; Zhao and Fitzgerald, 2013; Kong et al., 2015; Cuevas et al., 2016). Primarily, AC is considered as key determinant of rice eating quality (Juliano, 1992). Rice varieties with no or low AC is being linked to sticky and soft texture, respectively (He et al., 1999). Rice cultivars within high AC range showed variable textural attributes (Champagne et al., 1999, 2010). Thus additional parameters such as GC and GT were considered to unravel the degree of hardness (Bhattacharya and Juliano, 1985; Juliano, 1992; Shi et al., 1997; He et al., 1999; Lyon et al., 2000). Though, routine quality parameters found to be useful in predicting cooking quality, these predictive methods do not give sufficient information about the totality of rice textural attributes (Anacleto et al., 2015). Hence, there is a need to explore or develop other rice grain quality metrics which can be used to further differentiate rice texture (Juliano, 1985; Reddy et al., 1994).

Rice texture is regarded as a multidimensional sensory property that perceived by mouth feel characteristic features due to mechanical chewing, rheological, and surface attributes of a product perceptible by means of auditory receptors (Lawless and Heymann, 2010). Mechanical textural attributes of cooked rice such as hardness, cohesiveness, stickiness, and springiness can be characterized by a trained descriptive sensory panel. However, texture assessment pipeline through sensory is not routinely applied during selections of breeding lines because of the low throughput, nature of subjectivity, high cost of training and requirement of maintaining a descriptive panel (Sesmat and Meullenet, 2001). This is particularly true for breeding line selection at the early stages of a rice varietal improvement program, where thousands of lines were subjected for selection. Hence, understanding grain texture has focused on semi-throughput methodologies such as instrumental methods that correlate well with scores reported by sensory panels for the different textural characters (Meullenet et al., 1998; Champagne et al., 1999; Ramesh et al., 2000; Bett-Garber et al., 2001; Mestres et al., 2011). Texture Profile Analysis (TPA) is an instrumental method in which cooked rice grains undergo two compression cycles, mimicking the first and second bites on a food sample and thereby providing information on the mechanical responses such as hardness (HRD), adhesiveness (ADH), springiness (SPR), and cohesiveness (COH) (Stokes et al., 2013). Force-related textural properties of cooked rice can be measured using Texture Analyzer instrument which generate quantitative data, inexpensive, and the results are reproducible and reliable (Ramesh et al., 2000; Mestres et al., 2011).

High diversity in textural properties of rice has been reported (Bao et al., 2006). Like most grain quality traits, phenotypic variation in rice texture is quantitatively inherited (Hori et al., 2016). The genetic complexity of rice texture has been unraveled using classical QTL mapping. Through conventional QTL mapping, TPA parameters have been associated with quantitative trait loci (QTLs) on chromosomes 4 and 5 (HRD), 1 and 7 (ADH), and 8 (SPR) in a recombinant inbred mapping population whose individuals have low AC (Cho et al., 2010). So far these reported QTLs influencing texture have not been fine mapped and candidate genes not identified yet. These findings are notable because AC, coded by the *Waxy* gene in chromosome 6, is known to correlate positively with HRD and negatively with ADH (Suwannaporn et al., 2007); yet, no significant associations were reported by Cho et al. (2010), indicating that other genes are contributing to these textural attributes in cooked rice. This warrants the need of precise and robust statistical approaches for efficient capturing major with minor effect loci influencing the rice texture, which would pave the way to better understand the underlying genetic architecture.

More recently, genome wide association studies (GWAS) has become the state-of-art method to link genotypic variation to corresponding differences in phenotype, with the aim of dissecting the genetic basis of complex trait in various crops (Ingvarsson and Street, 2011; Xiao et al., 2017). GWAS offers high resolution-mapping by utilizing the historical recombination events, which leverages it with identification of key allelic variants and haplotypes in the underlying candidate genes. Nevertheless, single locus approach fails to consider the integrated effect of multiple markers under specific loci (Wang et al., 2016; Tamba et al., 2017). Moreover, using too conservative Bonferroni correction minimizes the likelihood to detect many important small effect loci (Wen et al., 2018). These issues have been addressed by efficiently utilizing the multi-locus GWAS approach in recent studies (Segura et al., 2012; Liu et al., 2016; Wang et al., 2016; Tamba et al., 2017; Wen et al., 2018; Zhang et al., 2018).

In addition to single locus genome wide association, the multi-locus GWAS methods were performed in the present study to overcome the limitations of single locus–based GWAS and to define the genetic basis of cooked rice texture and grain amylose content traits in *Indica* diversity lines. Furthermore, we conducted targeted gene-based association study using the available SNPs in the neighborhood region, which led to the construction of haplotypes showing phenotypic variation for the texture component traits.

MATERIALS AND METHODS

Plant Materials

A total of 236 diverse *Indica* accessions were selected from the rice diversity panels (RDP) (McCouch et al., 2016) by ensuring that the days to maturity is close between all entries, which did not exceed 140 days. These germplasm lines have been grown at the Robert S. Zeigler Experiment Station (ZES) of the International Rice Research Institute (IRRI), Laguna, Philippines (14°N, 121°E) during the 2014 dry season and wet season in randomized block design in three replications. Standard uniform field and crop management procedures have been adopted based on IRRI standard procedure across all of the replicates. Harvesting was done in the month of May/June depending upon their maturity time. After harvesting, standard IRRI drying method was followed in order to attain 12–14% seed moisture content. Subsequently, seeds were stored in the brown double-layer seed paper bags inside the seed storage room maintained at 18°C with optimum relative humidity.

Sample Processing

Paddy rice samples were dehulled using THU-35A test dehusker (Satake Corp., Japan) and brown rice was milled through Grainman 60-230-60-2AT instrument, (Grain Machinery Mfg. Corp., USA) to produce white milled rice samples. A portion from each sample corresponding to 100 unbroken grains was used for texture profile analyses (TPA) and the rest of samples were ground to a fine powder using Cyclone Sample Mill 3010-030, Udy Corp, USA. The resulted homogenized rice flour was further used for estimating amylose content (AC).

Amylose Content Measurement

AC determination was based on iodine colorimetric reaction using method of ISO 6647 (International Organization for Standardization, 2007) on milled rice flour. Briefly, gelatinized flour suspension was injected into the glass transition lines of a San++ Segmented Flow Analyser (SFA) system (Skalar Analytical B.V., The Netherlands) and allowed to react with iodine to form amylose-iodine complex (K-I_2). The absorbance of the sample's containing K-I_2 complex was estimated at 620 nm wavelength and subsequently, AC was quantified with standards by plotted against the standard curve.

Texture Profile Analysis (TPA)

Twenty-five whole polished rice grains per sample of accession were washed thrice, soaked for 30 min in Milli-Q water (1 mL) for 15 min in a test tube. The samples were heated to boiling point for 20 min and kept at 50°C prior to avoid retrogradation. Textural parameters of the cooked rice (hardness, cohesiveness, springiness, and adhesiveness) were analyzed according to the method described by Lyon et al. (2000) with modifications. The Ta.XT-Plus Texture Analyzer (Stable Micro Systems Ltd., Surrey, UK), equipped with a 35-mm aluminum cylinder probe with a 5-kg load cell, was used. The probe was positioned 15 mm above the base. Three intact cooked rice kernels were placed parallel with each other on the aluminum plate base under the center of the probe and compressed to 90% of their original height. The TPA force-deformation curve was obtained using a two-cycle compression test. The instrument is set with a test and post-test speed of 0.5 mm s^{-1}. Values of HRD (peak force of the first compression by the height of first curve), ADH (Negative force area under the first bite), COH (A_2/A_1), and SPR (T_2/T_1) were obtained and processed using Exponent Lite Software (version 3.0.5.0). ADH was recorded as negative numbers to indicate the direction of the probe's movement. Hence, adhesiveness values were reported in absolute values. Texture experiments were conducted in triplicate with three biological replications. For further details, refer **Supplementary Note 1**.

Genotyping Dataset

A 700K high Density Rice Array (HDRA) SNP genotyping set developed by an Affymetrix Custom Gene Chip Array from a SNP discovery dataset (McCouch et al., 2016) was used to develop genotyping information from the panel of 236 cultivars. A total of distinct 218 diverse germplasm lines were selected from the panel of 236 cultivars after following the standard filtering criterion with a missing rate of not more than 10% (mind 10%, geno 10%) and a minor allele frequency of at least 5%. This resulted into the consideration of final set of 147,692 high quality SNPs for conducting GWAS.

Single-Locus and Multi-Locus Genome Wide Association Studies (GWAS)

Mixed linear model based EMMAx (Kang et al., 2010) was carried out to conduct single locus (SL)-GWAS pipeline (Butardo et al., 2017). WarpedLMM (Fusi et al., 2014) was used to transform the phenotype to fulfill the normally distributed phenotype data for conducting the mixed linear model based approach. EMMAx-kin function was used to create the kinship matrix. Furthermore, Manhattan plot and Q-Q plot were created using the R package qqman (Turner, 2018). The Bonferroni corrected p-value $[-\log_{10}(P) = 6.47;\ P = 0.05/147692]$ was used as a threshold p-value. Nevertheless, since few loci have surpassed this threshold, significant SNPs above suggestive line at p-value of utmost 1e-5 were extracted as set of significant SNPs identified from SL-GWAS approach. Linkage Disequilibrium (LD)-plot and beta-effects of SNPs were plotted using combination of Haploview (Barrett et al., 2005) and Rscript. Targeted associations were done for the selected genes based on LD block and defined significant level. Annotations of candidate genes are based on MSUv7 annotation. Genetic regions identified from SL-GWAS approach were validated using at least two independent methods of ML-GWAS.

Four different methods namely FASTmrEMMA (Wen et al., 2018), pLARmEB (Zhang et al., 2017), mrMLM (Wang et al., 2016), and ISIS_EM-BLASSO (Tamba et al., 2017) were used to conduct multi-locus GWAS on 218 diverse germplasm with 147,692 high quality SNPs. All parameters used at their default values of respective method (Misra et al., 2017). A SearchRadius parameter (bin) (https://cran.r-project.org/web/packages/mrMLM/mrMLM.pdf) size of 20 SNPs was used as parameter to run the multi-locus association using algorithms mrMLM and FASTmrMLM. A threshold criterion of LOD of 3 and above was used to get the final set of significant SNPs. In house perl script was used to identify the overlapping genes with significant LOD score SNPs. SNPEff (Cingolani et al., 2012) was used for identifying the functional annotation of the respective SNP. Through implication of multi-locus GWAS and targeted associations, the boxplots were created to visualize the phenotype distribution among constructed haplotypes. Using R program, boxplots depicting phenotype distributing ranges were plotted for the respective textural traits. Multiple-t-test (pair-wise comparison) was implemented to identify the boxplot with significant phenotypic value at the significant level of $P \leq 0.05$.

RESULTS

Correlations Between AC and Texture Affecting Traits

Texture profile of 236 *Indica* diversity lines was evaluated. However, owing to follow high quality genotypic information upon filtering, a total of 218 diversity lines originated from 35 countries were selected to study AC and texture associated traits encompassing ADH, COH, HRD, and SPR. The pattern of phenotypic values for AC and ADH was skewed as many *Indica* germplasm found to be enriched for intermediate and high AC. Conversely, broad range of phenotypic values was observed for HRD, COH, and SPR attributes (**Figure S1**). Cor function with Pearson's method was used to detect the correlation and corrplot was used to create the plot.

AC exhibited a high negative correlation with ADH ($r = -0.83$, $P = 2.2E^{-16}$), which supports the fact that low amylose rice accessions tends to be stickier (**Figure 1**). Although, AC was considered as a key selection criterion for predicting cooked rice texture in normal breeding practices, HRD showed weak positive correlation with AC ($r = 0.39$, $P = 3.05E^{-09}$). Moreover, we did not observe significant correlation between AC with other textural attributes viz. COH and SPR. Instead, we observed a positive correlation within 3 textural attributes such as COH, HRD, and SPR ($r > 0.59$, $P = 2.2E^{-16}$; **Figure 1**).

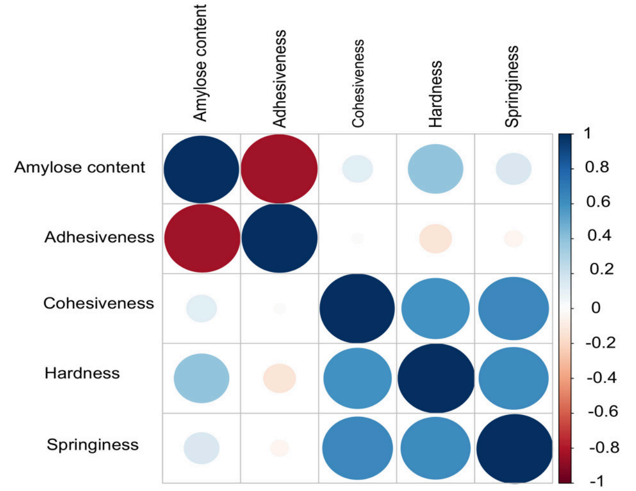


FIGURE 1 | Correlations among different grain quality attributes contributing to texture including the amylose content. Pearson method was used to calculate the correlation. Adhesiveness (ADH) and amylose content (AC) inversely correlated. No correlation was observed between AC other textural parameters cohesiveness (COH), hardness (HRD) and springiness (SPR).

Genetic Dissection of Texture Related Traits Using Single-Locus Genome Wide Association Studies (SL-GWAS)

To identify the large effect genetic regions of textural attributes in cooked rice and to delineate its interrelationship with AC, single-locus (SL) GWAS was performed using EMMAx (Kang et al., 2010) for marker trait associations in 218 germplasm lines using high quality 147,692 SNPs from the high density rice array (HDRA) panel of 700 K SNPs (McCouch et al., 2016). A total of 131 loci were associated [above suggestive lines; $-\log_{10}(P) \geq 5$] with texture-related traits. Among them the prominent peak on chromosome 6 associated with highly heritable traits AC ($h^2 = 0.86$) and ADH ($h^2 = 0.86$) (**Figure 2**), as indicated with the threshold value of significance by a red horizontal line in the Manhattan plot at $-\log_{10}(P\text{-value}) = 6.39$. HRD is another highly heritable trait ($h^2 = 0.82$), a significant region [$-\log_{10}(P) > 5$; blue suggestive line] was detected on chromosome 8 (**Figure 3**). Genetic association of SPR found for loci on chromosome 2, 6, and 9 were statistically significant (**Figure 4**). No QTN significantly associated with cohesiveness was detected utilizing SL-GWAS (**Figure S2**). Therefore, we furthermore employed multi-locus GWAS to reveal the significant loci including the small-effect loci (**Figure 5**).

Genetic Dissection of Textural Attributes Employing Multi-Locus -Genome Wide Association Studies (ML-GWAS)

For identifying novel associations and validation of loci detected using SL-GWAS, the multi-locus (ML) model approach was followed utilizing four independent methods namely FASTmrEMMA (Wen et al., 2018), pLARmEB (Zhang et al., 2017), mrMLM (Wang et al., 2016), and ISIS_EM-BLASSO (Tamba et al., 2017). The multi-locus method pLARmEB identified the highest number (90) of associated SNPs, followed by ISIS-EM-BLASSO (57). The lowest number of SNPs linked to texture related-traits was obtained with the FASTmrEMMA method (**Tables S1**, **S2**). A total of 173 quantitative trait nucleotides (QTNs) associated with texture attributes including AC, ADH, COH, HRD, and SPR were identified on all 12 chromosomes (**Tables S1**, **S2**). Notably, highly associated 10 QTNs identified with SL-GWAS method were validated using ML-GWAS methods (**Table 1**). These candidate genes includes QTNs on chromosome 2 regulating SPR (LOC_Os02g39630) and chromosome 4 QTN (LOC_Os04g55780) associated with ADH. An 87kb fine-mapped region (25.38–25.46 Mb) on chromosome eight identified QTN (LOC_Os08g40080) linked with HRD. Interestingly, hot spot QTLs on chromosome 6 co-located for two traits linking ADH textural trait (LOC_Os06g04169, LOC_Os06g04200, LOC_Os06g04530, intergenic region covering LOC_Os06g38564-LOC_Os06g38580) with AC (LOC_Os06g04000, LOC_Os06g04200, LOC_Os06g04290) validated using SL-GWAS and ML-GWAS methods (**Table 1**).

In addition, we observed a number of unique 38 QTNs associated with texture identified by two or more independent methods of ML-GWAS (**Table 2**) indicated with green dots in the Manhattan plot, which could not be otherwise detected by SL-GWAS due to adoption of higher threshold criterion (**Figure 2**). ML-GWAS enabled detection of 3 QTNs regulating both AC and ADH on chromosome 1 (LOC_Os01g09680), chromosome 4 (LOC_Os04g55780), and major effect genetic loci LOC_Os06g04200 from chromosome six encompassing the *GBSS1* (Waxy) gene (**Table 2**, **Figures 2A,B**). Additionally, a highly significant SNP found on chromosome 12 (LOC_Os12g22020, intronic splice variant) being associated with COH (LOD score of 6.35 with r^2 value of 12.05) and SPR (LOD score of 4.96 with r^2 value of 2.89). Additional SNPs associated with ADH was identified on chromosome 1 (LOC_Os01g17402) with a highly significant SNP of splice variant in Cyclin B1-3

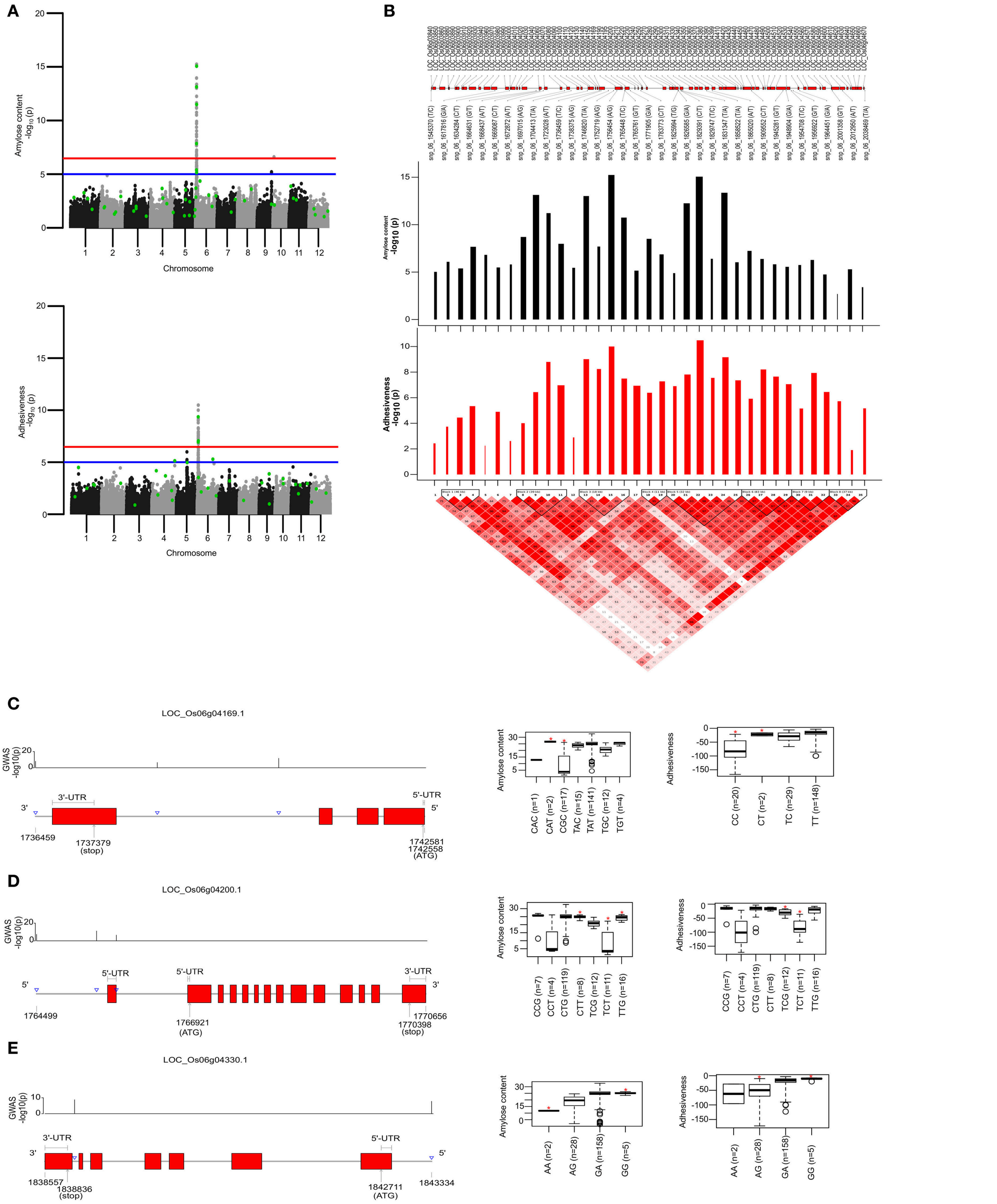


FIGURE 2 | Single-locus (SL) and multi-locus (ML)-GWAS for AC and ADH. **(A)** Manhattan plot showing QTNs identified from SL-GWAS showed in gray/black and ML-GWAS QTNs highlighted in green dot within the Manhattan plot. Genome-wide significant threshold line [$-\log_{10}(P) = 6.47$] is drawn as red, whereas suggestive line is represented by blue line at $-\log_{10}(P)$ of 5. **(B)** Linkage Disequilibrium (LD) plot with tagged SNPs from the pool of significant SNPs over suggestive line were plotted. A total of 8 blocks were identified based on D′ threshold criterion equal to 0.8. The $-\log_{10}(P)$ values was plotted as bar plot with positive effect as black bars and negative effect with red bars where width of bars represent the phenotypic effect size termed as beta effect. The overlapping genes were plotted in the top most lane **(C–E)**. Targeted gene associations for LOC_Os06g04169, LOC_Os06g04200, and LOC_Os06g04330 present in second, third and fifth LD-block, respectively. Gene structure with significant SNPs and phenotype distribution as boxplot were presented. An asterisks (*) represented the haplotype with significant phenotypic value (at significance level of $P \leq 0.05$) using pair wise *t*-test.

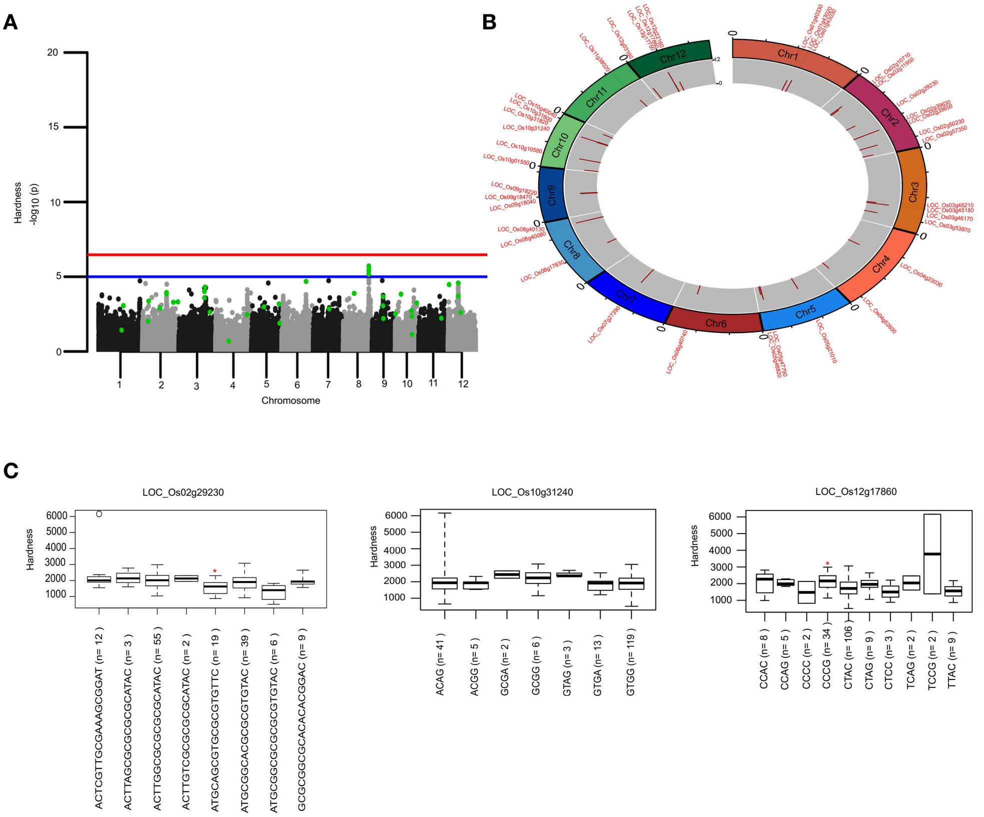


FIGURE 3 | Genetic regions identified through ML-GWAS and SL-GWAS for hardness (HRD). **(A)** Manhattan plot showing the multi-locus associations (QTNs with LOD ≥3; highlighted as green dots) overlaid with SL-GWAS QTNs (black/gray) for HRD. **(B)** Circos representing the physical positioning of 12 chromosomes with locus IDs of significant QTNs identified in ML-GWAS, followed by depiction of LOD score in the innermost circle. **(C)** Phenotypic distribution of haplotypes shown as boxplot for selected genes identified from ML-GWAS method. Haplotypes showing significant HRD values were highlighted with an asterisks (*) (at the significance level of $P \leq 0.05$) using pair wise *t*-test.

and on chromosome 5 (LOC_Os05g26850, promoter region of unknown gene), identified by all four multi-locus methods (**Table 2**).

Major Genomic Region Determines the Adhesiveness (ADH) and Amylose Content (AC)

AC being the starch component, it is one of the key determinants of the cooking and eating quality. In cooked form, AC negatively influences the ADH (**Figure 1**). Major genetic region of ~490 kb candidate genomic region (1.54–2.0 Mb) possessing 8 LD-blocks on chromosome 6 has been mapped for both ADH and AC, confirmed using both SL-GWAS and ML-GWAS methods (**Table 1**, **Figures 2A,B**, **Figure 3**). Interestingly, this fine mapped genetic region on chromosome 6 was consistently identified when GWAS has been conducted across wet and dry seasons (**Figure S3**). Notably, multi-loci GWAS detected moderate to high effect significant SNPs from LD blocks 2, 3, and 5 defining the variations for AC and ADH (**Figure 2B**). Furthermore, haplotypes identified from LD-block 2 and 3 contributed in distinguishing samples from high/intermediate amylose classes with low AC (**Figures 2C,D**), whereas allelic variants from LD-block 5 differentiates lines possessing intermediate AC with low AC (**Figure 2E**). Most haplotypes showing the high AC (>25%) were observed as less adhesive (low magnitude) as reflected in the haplotypes from LD-block 2 and 3. Likewise, haplotypes fixed for intermediate AC, showed moderate variations for ADH. Contrastingly, low AC samples showed high ADH values, as observed in LD-block 5. Targeted-gene association study of the potential candidate genes (LOC_Os06g04169 encoding beta-hydrolase, LOC_Os06g04200 identified as *GBSS1* and LOC_Os06g04330 annotated as Phosphotransferase) distinguished different AC and ADH classes of phenotypes (**Figures 2C,E**, **Table S2**), were validated by two independent methods such as ML-GWAS and SL-GWAS. For the candidate gene LOC_Os06g04169, haplotype CGC were found to be correlated to low AC and high ADH phenotypes and its alternative haplotype (TAC/TAT/TGC/TGT) containing lines were correlated to possess high AC and low ADH (**Figure 2C**). Haplotype (CCT/TCT) identified from the LOC_Os06g04200 were correlated to low AC and high ADH phenotypes and its alternative haplotypes (CTG/CTT/TTG) containing lines found to possess high AC and low ADH

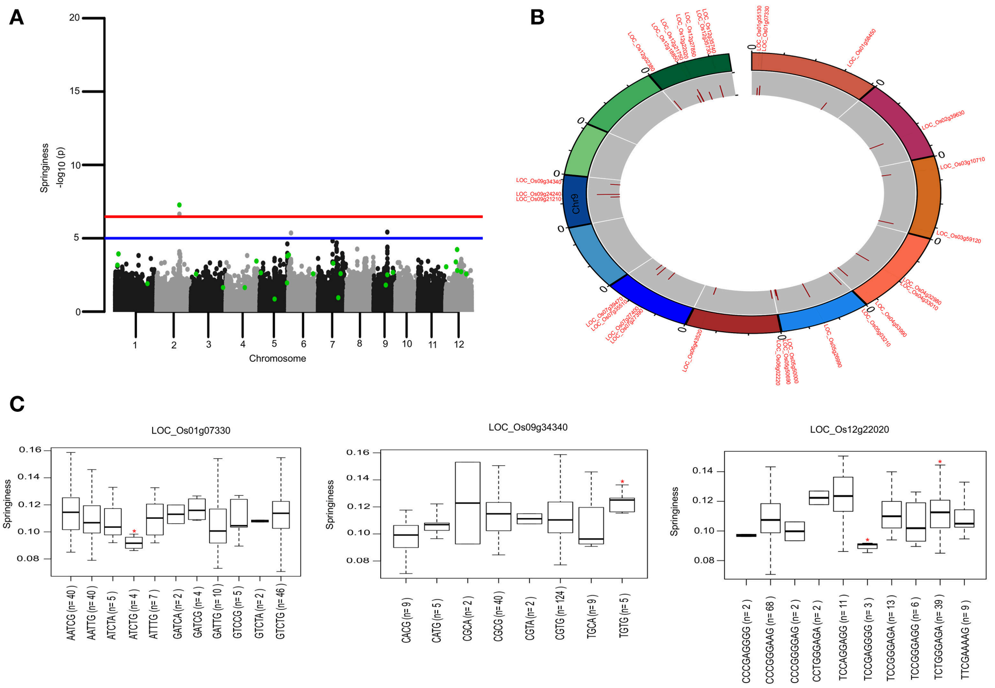


FIGURE 4 | Genetic regions identified through ML-GWAS and SL-GWAS for springiness (SPR). **(A)** Manhattan plot representing the multi-locus associations (QTNs with LOD ≥3; highlighted as green dots) overlaid with SL-GWAS QTNs (black/gray) for SPR. **(B)** Circos representing the physical positioning of 12 chromosomes with locus IDs of significant QTNs identified in ML-GWAS, followed by depiction of LOD score in the innermost circle. **(C)** Phenotypic distribution of haplotypes shown as boxplot for selected genes identified from ML-GWAS method. Haplotypes showing significant SPR values were highlighted with an asterisks (*) (at the significance level of $P \leq 0.05$) using pair wise *t*-test.

(**Figure 2D**).Within LOC_Os06g04200 (granule bound starch synthase I), the high effect QTN at position 1765761 (LOD > 11) lying at splice junction of exon 1 detected using FASTmrEMMA and pLARmEB significantly affected both AC and ADH values (**Table S2**). Additionally, splice junction QTNs along with other two QTNs identified in *GBSS1* showed variable level of allele frequencies (**Table S3**). Conversely, haplotypes from candidate gene LOC_Os06g04330 explained the variations among intermediate and low AC classes (**Figure 2E**).

Utilizing the SL and ML-GWAS, prominent association signals were identified for ADH alone from chromosome 4 (LOC_Os04g55780), which is linked negatively with ADH trait with a beta value of −0.42 identified using SL-GWAS and confirmed using two independent methods of ML-GWAS with LOD scores of 3.2 and 6.4. In addition, a total 11 QTNs were identified for ADH but not linked to AC, which are identified from at least two independent methods of ML-GWAS (**Figure 5**). These candidate loci of major effect QTNs affecting ADH trait located on chromosome 1 (LOC_Os01g09680 and LOC_Os01g55070 with r^2 value of 2.53 and 1.65, respectively), chromosome 4 (LOC_Os04g16260 with r^2 value of 1.38), chromosome 5 (LOC_Os05g26850 with r^2 value of 3.05), chromosome 7 (LOC_Os07g26990 with r^2 value of 1.29), chromosome 9 (LOC_Os09g07890 with r^2 value of 1.32), chromosome 10 (LOC_Os10g22590 with r^2 value of 1.27), and chromosome 11 (LOC_Os11g39680 with r^2 value of 1.10) (**Table 2**). Using four different methods, multi-locus GWAS yielded highly significant (LOD > 10) SNP with splice variant in candidate gene LOC_Os01g17402 and in the promoter region of LOC_Os05g26850, explaining high heritable phenotypic variation for traits AC and ADH (**Table S2**). Among them highly significant snp_05_15589585 (c.-2404T>A) present in the upstream of LOC_Os05g26850 (unclassified) on chromosome 5 showed prominent association with ADH using different ML-GWAS methods, while it was not significant under SL-GWAS (**Table S2**).

Additional 11 QTNs were identified to influence AC but not linked to ADH confirmed from two or more independent methods of ML-GWAS. With the exception of 3 QTNs (LOC_Os05g24190, LOC_Os06g33360, and LOC_Os08g32520), many of them were turned out to be minor effect QTNs with low r^2 value. The major effect QTN affecting AC namely snp_06_19437296 has a non-synonymous nucleotide change (C>T) in the candidate gene LOC_Os06g33360 lead to amino acid change (Ala>val) (**Table 2**).

Genetic Dissection of Hardness (HRD), Cohesiveness (COH), and Springiness (SPR) as Components of Textural Attributes

The key components determining the cooked rice texture include the HRD, COH, and SPR, which showed higher correlations among each other (**Figure 1**), but not linked to

TABLE 1 | Significant SNPs associated with amylose content and texture-related traits identified using single and multi-locus GWAS methods.

SNP_ID	Chr	Method	LOD	r^2 (%)	Trait	Beta	$-\log_{10}(p)$	Gene	Nucleotide change	SNP location/Amino acid change	SWISSPROT
snp_02_23917456	2	mrMLM	5.68	10.27	SPR	0.45	7.27	LOC_Os02g39630	c.*4252G>A	Downstream gene	No hits
snp_04_33203256	4	mrMLM, pLARmEB	3.21; 6.38	0.84; 0.16	ADH	−0.42	5.13	LOC_Os04g55780	c.-3159G>A	Promoter region	Arogenate dehydratase/prephenate dehydratase

snp_06_1634284	6	ISIS-EM-BLASSO
		mrMLM,
		pLARmEB
			3.08;
			12.48;
			22.33
				4.11;
				9.03;
				1.50
					AC	0.59	5.39	LOC_Os06g04000	c.-895C>T	Promoter region	Peptidyl-prolyl cis-trans isomerase

snp_06_1744863	6	mrMLM, pLARmEB	7.88; 29.90	3.25; 1.37	ADH	−0.61	9.36	LOC_Os06g04169	c.-2305C>T	Promoter region	No hits
snp_06_1765761	6	FASTmrEMMA pLARmEB	11.69; 12.18	5.64; 1.11	AC	0.54	5.15	LOC_Os06g04200	c.-33+2G>T	Splice variant	Granule-bound starch synthase1
snp_06_1765761	6	FASTmrEMMA	12.32	7.77	ADH	−0.63	6.94	LOC_Os06g04200	c.-33+2G>T	Splice variant	Granule-bound starch synthase1
snp_06_1826635	6	ISIS-EM-BLASSO pLARmEB	12.36; 23.59	3.69; 0.84	AC	0.64	11.49	LOC_Os06g04290	c.-4704G>T	Promoter region	40S ribosomal protein S20

snp_06_1948904	6	mrMLM,
		FASTmrEMMA,
		ISIS_EM-BLASSO
			5.57;
			14.91;
			16.38
				1.87;
				8.05;
				5.86
					ADH	−0.53	7.06	LOC_Os06g04530	c.-601G>A	Promoter region	ATP-dependent Clp protease

snp_06_22848516	6	pLARmEB,
		ISIS_EM-BLASSO,
		mrMLM
			3.33;
			3.37;
			5.55
				0.08;
				1.26;
				1.69
					ADH	−0.56	5.29	LOC_Os06g38564-
								LOC_Os06g38580
									n.22848516G>C	Intergenic region	No hits

snp_08_25375457	8	pLARmEB, mrMLM	4.27; 6.51	1.68; 4.56	HRD	0.36	5.17	LOC_Os08g40080	c.-181C>T	Promoter region	No hits

TABLE 2 | Significant SNPs associated with amylose content and texture-related traits identified using several multi-locus GWAS method.

SNP_ID	Chr	Method(s) used	LOD	r^2 (%)	Trait	Gene	Nucleotide change	SNP location/Amino acid change	SWISSPROT
snp_01_3464018	1	pLARmEB; FASTmrEMMA	3.38; 4.63	0.76; 4.04	SPR	LOC_Os01g07330	c.-187A>G	Promoter region	No hits
snp_01_10014553	1	pLARmEB; ISIS_EM-BLASSO; FASTmrEMMA; mrMLM	4.50; 5.36; 7.97; 13.09	0.33; 3.96; 5.55; 6.54	ADH	LOC_Os01g17402	c.-3673C>A	p.Val140Ile	No hits
snp_01_31665007	1	FASTmrEMMA; mrMLM; pLARmEB	3.76; 8.18; 17.48	1.65; 3.04; 0.74	ADH	LOC_Os01g55070	c.-3673C>A	Downstream gene	No hits
snp_01_4991874	1	ISIS_EM-BLASSO; FASTmrEMMA	3.51; 10.29	0.76; 2.43	AC	LOC_Os01g09680	c.212-4G>A	Promoter region	Cyclin-B1-3
snp_01_4991874	1	FASTmrEMMA	5.98	2.53	ADH	LOC_Os01g09680	c.-926C>T	Promoter region	Pentatricopeptide repeat containing protein
snp_02_23917456	2	pLARmEB	10.44	3.08	HRD	LOC_Os02g39630	c.*4252G>A	Intergenic region	No hits
snp_02_29507254	2	pLARmEB; mrMLM	3.82; 6.21	0.08; 2.26	AC	LOC_Os02g48184	c.-5090C>T	Promoter region	No hits
snp_02_30672463	2	mrMLM; pLARmEB	6.89; 6.90	7.41; 2.15	HRD	LOC_Os02g50230	c.1227C>T	p.Ala399Val	No hits
snp_03_30485102	3	pLARmEB; FASTmrEMMA	3.60; 4.42	0.08; 1.36	AC	LOC_Os03g53150	c.-1973G>A		Auxin-responsive protein_IAA13
snp_04_33203256	4	pLARmEB; ISIS_EM-BLASSO	5.82; 8.86	0.13; 2.15	AC	LOC_Os04g55780	c.-2831T>C	Intergenic region	No hits
snp_04_8836677	4	mrMLM; pLARmEB	4.57; 13.14	1.38; 0.48	ADH	LOC_Os04g16260	c.-3159G>A	Promoter region	Arogenate dehydratase/prephenate dehydratase
snp_05_12361761	5	pLARmEB; ISIS_EM-BLASSO	4.36; 5.47	3.07; 10.70	HRD	LOC_Os05g21010	c.418G>A	p.Val140Ile	No hits
snp_05_13977160	5	ISIS_EM-BLASSO; pLARmEB	3.42; 6.05	1.26; 0.27	AC	LOC_Os05g24190	c.*3646T>C	Downstream gene	polyprotein
snp_05_15589585	5	ISIS_EM-BLASSO; FASTmrEMMA; mrMLM; pLARmEB	6.11; 11.45; 13.72; 26.36	3.05; 5.03; 4.72; 1.22	ADH	LOC_Os05g26850	c.-2404T>A	Promoter region	No hits
snp_05_21765243	5	pLARmEB; FASTmrEMMA	4.66; 13.75	0.11; 4.42	AC	LOC_Os05g37225	c.-2713C>T	Promoter region	No hits
snp_05_27396039	5	mrMLM; ISIS_EM-BLASSO	3.82; 6.36	2.29; 3.77	HRD	LOC_Os05g47790- LOC_Os05g47810	n.27396039A>G	Intergenic region	No hits
snp_06_19437296	6	mrMLM; ISIS_EM-BLASSO; pLARmEB	3.92; 4.15; 6.69	1.47; 1.15; 0.17	AC	LOC_Os06g33360	c.-1942C>T	Promoter region	No hits
snp_06_27648676	6	pLARmEB; mrMLM	6.72; 7.82	0.32; 3.00	ADH	LOC_Os06g45650- LOC_Os06g45660	c.1196C>T	p.Ala399Val	NB-ARC_domain-containing protein

(Continued)

TABLE 2 | Continued

SNP_ID	Chr	Method(s) used	LOD	r^2 (%)	Trait	Gene	Nucleotide change	SNP location/Amino acid change	SWISSPROT
snp_06_6710536	6	pLARmEB; mrMLM	4.56; 7.11	0.12; 2.00	AC	LOC_Os06g12380	n.27648676A>G	Intergenic region	No hits
snp_07_15625737	7	mrMLM; pLARmEB	3.38; 6.96	1.29; 0.21	ADH	LOC_Os07g26990	c.52G>A	p.Thr18Thr	No hits
snp_07_15950275	7	pLARmEB	4.01	0.7	SPR	LOC_Os07g27390-LOC_Os07g27400	n.15950275C>T	Intergenic region	No hits
snp_07_15950275	7	ISIS_EM-BLASSO; pLARmEB	4.61; 7.83	3.56; 2.23	HRD	LOC_Os07g27390-LOC_Os07g27400	n.15950275C>T	Intergenic region	No hits
snp_07_23649340	7	ISIS_EM-BLASSO	3.06	1.5	SPR	LOC_Os07g39470	c.-2786T>G	Promoter region	Chitin-inducible gibberellin responsive gene
snp_07_26549549	7	pLARmEB	5.59	0.06	AC	LOC_Os07g44430	c.513C>T	p.Asp171Asp	phospholipase
snp_08_20145633	8	FASTmrEMMA; pLARmEB	5.90; 6.40	1.03; 0.09	AC	LOC_Os08g32520	c.*441G>A	Downstream gene	No hits
snp_09_14649274	9	pLARmEB; ISIS_EM-BLASSO; mrMLM	3.72; 3.80; 4.7	0.15; 1.82; 2.42	ADH	LOC_Os09g24590	c.-3712C>T	Promoter region	polyprotein
snp_09_4018179	9	ISIS_EM-BLASSO; pLARmEB; mrMLM	3.60; 4.18; 6.67	1.32; 0.11; 1.91	ADH	LOC_Os09g07890	c.*523A>T	3 prime UTR	No hits
snp_10_11723831	10	FASTmrEMMA; mrMLM	4.07; 6.25	1.27; 1.43	ADH	LOC_Os10g22590	c.-2300C>T	Promoter region	No hits
snp_10_13693521	10	ISIS_EM-BLASSO; FASTmrEMMA	3.69; 5.76	10.69; 0.00	COH	LOC_Os10g26370	c.187+112C>A	Intron	allergen and extensin family protein
snp_10_5839378	10	mrMLM; pLARmEB	4.12; 7.89	8.67; 5.85	HRD	LOC_Os10g10580	c.*4639A>T	Downstream gene	No hits
snp_11_10872055	11	pLARmEB; ISIS_EM-BLASSO	4.29; 5.27	0.16; 2.56	AC	LOC_Os11g19060	c.*3782C>T	Downstream gene	AP2-like ethylene-responsive transcription famylose contenttor
snp_11_22559123	11	ISIS_EM-BLASSO; pLARmEB	3.35; 3.93	1.74; 0.80	HRD	LOC_Os11g38020	c.1110+4C>T	Splice variant	small GTP-binding protein
snp_11_23638170	11	ISIS_EM-BLASSO; FASTmrEMMA	3.71; 4.50	1.10; 1.49	ADH	LOC_Os11g39680	c.*2157G>A	Downstream gene	No hits
snp_12_10170782	12	pLARmEB; ISIS_EM-BLASSO; mrMLM	6.29; 8.71; 10.25	2.26; 7.76; 10.68	HRD	LOC_Os12g17750	c.830G>A	p.Gly277Asp	No hits
snp_12_1204320	12	FASTmrEMMA; mrMLM	3.10; 5.90	3.24; 5.81	HRD	LOC_Os12g03160	c.-4154C>T	Promoter region	No hits
snp_12_12232293	12	mrMLM; pLARmEB	3.62; 6.64	11.00; 0.27	SPR	LOC_Os12g21750	c.-4282G>C	Promoter region	No hits
snp_12_12385384	12	ISIS_EM-BLASSO	4.96	2.89	SPR	LOC_Os12g22020	c.175-375G>A	Intron	No hits
snp_12_12385384	12	ISIS_EM-BLASSO	6.35	12.05	COH	LOC_Os12g22020	c.175-375G>A	Intron	No hits

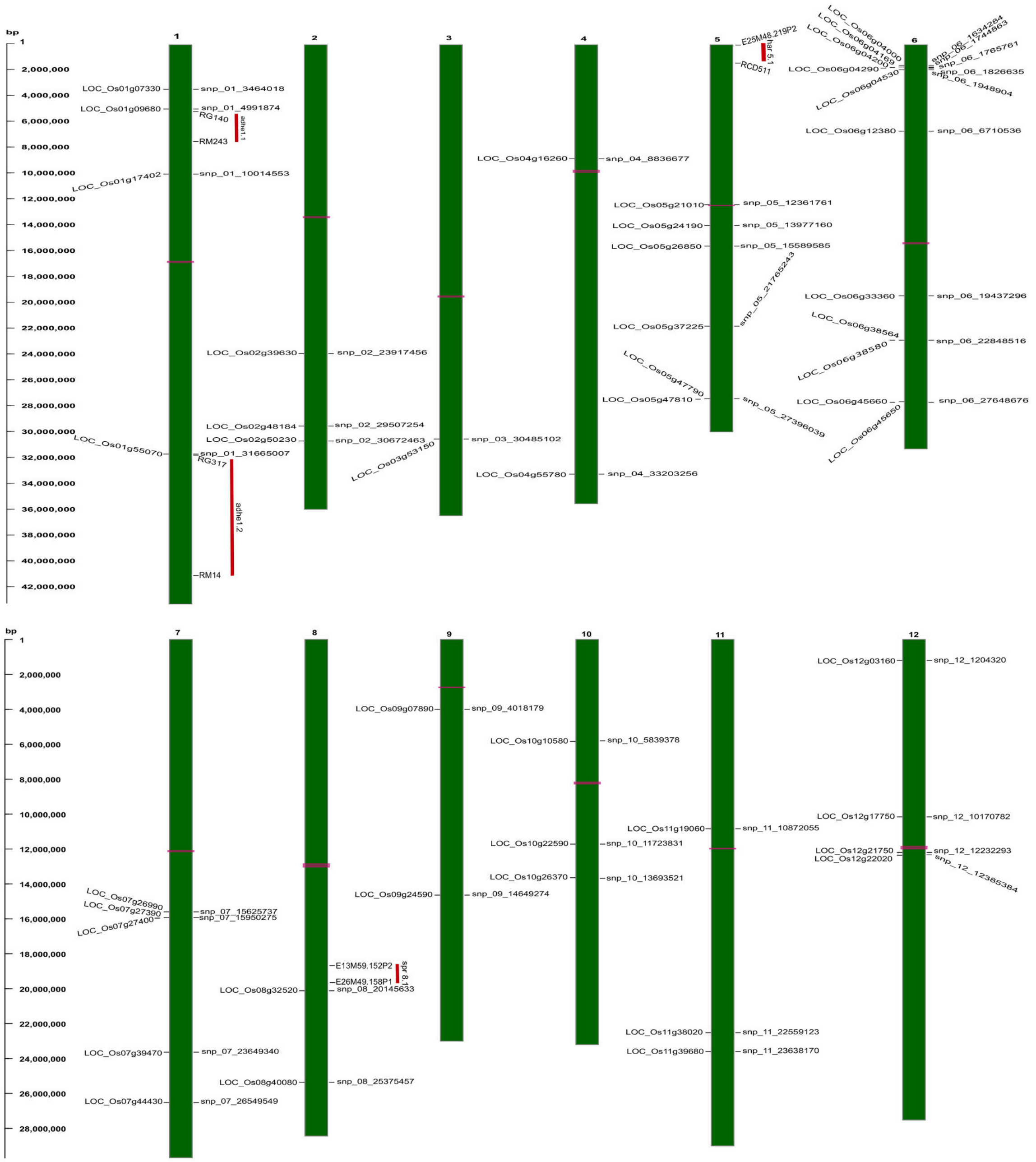


FIGURE 5 | Physical map of strongly associated QTNs on rice chromosomal maps identified using ML-GWAS methods. The represented QTNs on map includes either QTNs jointly identified using SL- and ML-GWAS, or spotted in at least two independent ML-GWAS methods, mapped at their physical positions on 12 chromosomes. On the left, scale indicates the base pair (bp) distance. Publicly available texture QTL information from Cho et al. (2010) have been mapped, and were highlighted as vertical red colored bars aligning with respective physical positions on the map. Horizontal bars (in pink color) on chromosomal maps represent the position of centromere.

AC variation in the grain. Significant QTN on chromosome 8 with SNP (C181>T) found in the promoter region of LOC_Os08g40080 influence HRD, validated by both SL-GWAS and ML-GWAS methods (**Table 1**). For HRD, multi-locus GWAS yielded 9 significant QTNs identified on chromosomes 2 (LOC_Os02g39630, LOC_Os02g50230), chromosome 5 (LOC_Os05g21010, intergenic region interval LOC_Os05g47790-LOC_Os05g47810), chromosome 7 (intergenic interval LOC_Os07g27390-LOC_Os07g27400), chromosome 10 (LOC_Os10g10580), chromosome 11 (LOC_Os11g38020) and chromosome 12 (LOC_Os12g03160, LOC_Os12g17750) (**Table 2**, **Figure 3B**). These QTNs were identified as major QTNs with higher genetic heritability, which are validated by two or more independent methods namely, FASTmrEMMA, ISIS_EM-BLASSO, mrMLM, and pLARmEB (**Table 2**, **Figure 3B**). Of the 9 QTNs associated with hardness, none of the QTNs was identified to affect both hardness and AC. A prominent QTN snp_12_10170782 (C>T) was identified in the promoter region of LOC_Os12g17750 (unknown function) identified for influencing hardness using three independent methods pLARmEB, ISIS_EM-BLASSO, mrMLM with LOD score of 6.3, 8.7, and 10.24 (**Table 2** and **Table S2**). For the candidate gene LOC_Os12g17750, reference haplotype CTAC being abundant in major *Indica* germplasm with intermediate hardness and its alternative haplotype TCCG representing lines depicted higher hardness value (**Figure 3C**). Likewise, additional QTNs representing missense mutations such as snp_05_12361761 leading to amino acid change (Val>Ile) in candidate LOC_Os05g21010 and snp_12_1204320 (Gly>Asp) was detected in LOC_Os12g03160, associated with HRD trait (**Table 2**).

Genetic basis of springiness textural trait was defined through SL-GWAS on chromosome 2 (LOC_Os02g39630) validated using ML-GWAS approach (**Table 1**, **Figure 4**). Additional QTN identified through SL-GWAS resulted in identifying a locus LOC_Os09g34340 with contrasting haplotypes (CGCA with higher value of SPR and CACG with lower value of SPR). Employing ML-GWAS approach 5 additional QTNs were identified and among them LOC_Os12g22020 locus was defined by contrasting haplotype TCCAGGAGG with higher SPR value and alternative haplotype TCCGAGGGG containing lines possess lower SPR (**Figure 4**). We also identified snp_01_3464018 causing premature termination at start codon of the candidate locus LOC_Os01g07330 (unclassified), which exhibited the distinct haplotypes showing variation for the SPR (**Figure 4C**). Moreover, the extreme haplotypes detected for ADH, HRD, and SPR also showed the consistent phenotype across wet and dry seasons (**Figure S4**).

For COH, we observed significant snp_10_13693521 located downstream of LOC_Os10g26370 validated through ISIS_EM-BLASSO, FASTmrEMMA multi-locus GWAS methods with LOD score value of 3.7 and 5.7 (**Table 2**). Additional QTN snp_12_12385384 identified from ISIS_EM-BLASSO was found to associate very significantly with COH (LOD 6.35 with r^2 12.05) and SPR (LOD 4.95 with r^2 2.88).

Kyota Encyclopedia of Genes and Genomes (KEGG) analysis was conducted to identify the functional categories across all the QTNs identified. A total of 40 candidate genes with functional information were mined, of which 17 were involved in genetic information processing, whereas the rest of the candidate genes were involved in other cellular and metabolic processes (**Figure S6**).

DISCUSSION

Interlinking Amylose Content Variation With Textural Attributes

Texture of cooked rice plays a pivotal role in consumer acceptability; henceforth researchers continuously develop strategies to predict texture of cooked rice. AC is widely explored to capture the diversity of rice quality (Anacleto et al., 2015) in rice breeding programs. The challenge lies when rice varieties within similar AC quality class are easily differentiated by consumers (Champagne et al., 1999, 2010), and thus secondary traits derived based on AC versus GT or AC versus GC will be prioritized. This situation highlights the importance of secondary assays that could further differentiate rice varieties into distinct quality classes. It is assumed that rice varieties within the same AC and GC ranges are distinguished from each other for textural attributes. Mega varieties which have been developed during 1965–1990 were found to possess unique textural attributes, which cannot be distinguished alone by AC. Hence revealing textural attributes is a crucial step in fine-tuning product profiles for capturing major rice markets tend to distinguish rice within intermediate and high AC. Multi-modal descriptive sensory description is the ultimate reference to distinguish textural features of rice varieties (Anacleto et al., 2015). However, it is difficult to routinely implement descriptive sensory methods to capture textural preferences for selecting breeding material from rice improvement programs due to large number of samples, low throughput methodology. Thus to develop and deliver selection tools to breeders, there is a need to bridge proxy grain quality selection tools (AC) with instrumental based TPA analysis and predict textural ideotypes from diverse germplasm (Champagne et al., 1999).

TPA is an semi-throughput approach used to measure the mechanical response during a double compression, which mimics first and second bite of a food sample (Stokes et al., 2013). Until now, attempts have been made to correlate instrumental texture profiles with various rice quality predictors (Ohtsubo et al., 1990; Champagne et al., 1999, 2004). Among textural attributes, HRD is considered as an important attribute of cooked rice texture strongly affect the consumer acceptance (Perez et al., 1979; Li et al., 2016). We employed highly diverse *Indica* germplasm enriched with intermediate to high AC to generate phenotyping data of various textural attributes using TPA (**Figure 1**). These results suggest a week positive correlation between AC and HRD (**Figure 1**). This is differed with the outcome of previous studies where HRD showed very significant positive correlation with AC (Perez et al., 1979; Bao et al., 2004; Li et al., 2016, 2017). These results have proved inconsistent and at times coincidental, most likely because of utilization of small sample sets to establish associations. In addition, the samples are most likely not representative of covering the entire breadth and depth of the diversity of *Indica* germplasm.

Our results delineated that AC was negatively correlated with the ADH ($r = -0.83$). Previous studies reported that low AC value with higher amylopectin content increase the stickiness in cooked rice (Juliano, 1992; Reddy et al., 1993; Windham et al., 1997; Li et al., 2016). Additionally, AC displayed non-significant correlation with COH, SPR, which is in agreement with the Cho et al. (2010). These results suggested that AC is not the sole determinant of cooked rice texture, and it is important to equally consider other component textural attributes viz. HRD, COH, and SPR. Moderate to high correlations exist among HRD, COH, and SPR traits, which suggest the existence of interdependence among these three textural traits. To support future breeding programs, we need to dissect genetics and leverage the gene discovery attempts toward crop improvement of textural attributes.

Dissecting the Genetic Basis of Texture Related Traits

Genetic dissection of the textural attributes has been explored in previous studies using *Japonica*/*Indica* biparental mapping population (Bao et al., 2002, 2004; Cho et al., 2010). Nevertheless, studies involving high resolution dissection of genetic basis of important texture related traits are very limited. To estimate the heritability, we have measured amylose content from two independent years and used this phenotyping data to identify the genetic regions. We identified the same genomic regions on chromosome 6, regulating amylose content across two season data with higher heritability's values ($h^2 = 0.85$ and 0.86), reflecting the consistency across both seasons (**Figure S3**). As the selected diversity panel showed high phenotypic variation for amylose, which influences texture, we have used the data for textural parameters from 2014 dry season (in randomized complete block design) for conducting the GWAS. The texture based phenotyping data were generated from nine technical replicates. In this approach we have considered the ample replications to ensure the phenotype consistency and explored the diversity population covering vast range of phenotypic variation for textural attributes to define its genetic basis. In present study, AC and ADH have reflected the high narrow sense heritability ($h^2 = 0.86$). Thus the defined genetic region will be an added value. Furthermore, notably AC and ADH showed high degree of correlation and overlapping genetic region influenced by both traits. Likewise, HRD ($h^2 = 0.82$) and COH ($h^2 = 0.68$), showed higher heritable values from the RDP than that observed in previous studies dealing with biparental mapping populations (Bao et al., 2002; Cho et al., 2010). In addition, we selected the accessions possessing the haplotypes exhibiting the extreme phenotypes for adhesiveness, hardness and springiness, and phenotyped for different textural attributes in the seed lots collected from two different seasons. As a result, we identified the values from both of the seasons close and comparable to each other (**Figure S4**). These results further confirmed that genetic component majorly regulating the textural trait is highly heritable and less affected by the environmental effects. In addition, we have utilized days to maturing data as covariate and re-run single locus GWAS for AC and four textural attributes. All the genetic regions identified for textural attributes using SL-GWAS peaks for AC, ADH, HRD, and SPR remains significant, when we run with days to maturity as covariate (refer **Figure S5**). Through this approach we rule out any influence of days to maturity on texture in the currently studied core collection panel.

EMMA algorithm has been extensively used to dissect the complex traits due to its robustness and reliability. Furthermore EMMA model corrects for confounding effects of subpopulation structure and relatedness between individuals (Kang et al., 2008; Kumar et al., 2015; Campbell et al., 2017). Application of single-locus scan approach under polygenic background with diverse population structure controls do not facilitate the detection of small effect QTNs, as the model fails to consider the integrated effect of multiple markers under specific loci (Zhang et al., 2018). Alternative and more powerful approaches for marker-trait association have been developed to address the shortcomings of one dimensional scan. Hence in the present study we used four multi-locus model approaches FASTmrEMMA (Wen et al., 2018), pLARmEB (polygenic-background-control-based least angle regression plus empirical Bayes) (Zhang et al., 2017), mrMLM (Wang et al., 2016), and ISIS_EM-BLASSO (Tamba et al., 2017) to conduct GWAS analysis in order to capture minor QTNs related to texture traits. A total of 224 SNPs associated with AC and textural attributes such as ADH, COH, HDR, and SPR were defined using ML-GWAS method. When comparing the four multi-locus methods, a high number of 97 SNPs were validated with at least 2 out of the four multi-locus methods. Among them 48 novel loci were defined to influence texture attributes. Among the implemented ML-GWAS methods, pLARmEB and ISIS EM-BLASSO detected the higher number of significant QTNs. The same observation was made in the recent study (Sant'ana et al., 2018), who reported a high number of trait-associated SNPs using two methods. On the other hand (Ma et al., 2018), acknowledge the robustness of ISIS EM-BLASSO than the other three methods (FASTmrEMMA, pLARmEB, mrMLM). More than half of the QTN detected were specific to one of the four methods used. Thus validation accounted by combinatory approaches of ML-GWAS methods has been considered in our study for interpreting biological inferences of rice texture.

Using the efficient mixed-model association we identified 10 robust QTNs with major effect QTNs significantly associated with texture-related traits validated using SL-GWAS and ML-GWAS approaches. To discover medium and small effect QTNs, we used confirmatory validation through at least two independent methods of ML-GWAS approaches. Unlike SL-associations, ML-association approaches are considered as effective in taking the joint effects of multiple genetic markers into account and avoid any stringent criterion leading to likelihood of missing out functionally relevant genomic loci (Wang et al., 2016; Tamba et al., 2017; Wen et al., 2018). Indeed, although, no SNP significantly associated QTNs were identified for cohesiveness using SL-GWAS and mrMLM methods, we detected putative genomic regions underlying COH using three other ML-GWAS methods. Multi-locus GWAS has gained popularity with a growing number of studies reporting the use of this approach (Misra et al., 2017; Ma et al., 2018; Sant'ana et al., 2018; Zhang et al., 2018) to perform marker-trait association. Some of the distinct advantages of multi-locus GWAS over single-locus GWAS are their power, accuracy in QTN effect estimation, reduced rate of false positives (Wang et al., 2016; Tamba et al., 2017; Ma et al., 2018).

We identified a well characterized major effect QTN affecting AC and ADH within *GBSS1* region confirmed through SL- and ML-GWAS, particularly loci at the 5′-splice site of first intron (**Figure 2**), which is in agreement with previous reports (Hsu et al., 2014; Yang et al., 2014; Wang et al., 2017). Furthermore, two additional SNPs were detected for regulating AC, earlier identified by Butardo et al. (2017) for determining AC. In addition to *GBSS1, w*ithin the hot spot QTL of chromosome 6 linking AC with ADH textural attribute we identified additional loci influencing texture traits. For instance, haplotypes derived from candidate genes encoding alpha/beta-hydrolases, which belongs to largest group of structurally related enzymes (Holmquist, 2000) and uncharacterized phosphotransferase, explicitly showed the phenotypic variation with AC and ADH traits. Notably, adjacent region of *GBSSI* reflected the variations for ADH has been reported previously but the candidate loci were not unraveled (Wang et al., 1995; Isshiki et al., 1998). This may be attributed to low resolution owing to limited recombination events observed in case of bi-parental population.

Besides *GBSSI,* we did not detect the QTNs from candidates involved in starch biosynthesis, namely starch synthase IIa (*SSIIa*) (Umemoto et al., 2002; Nakamura et al., 2005), starch branching enzyme (*SBE IIb*) (Nishi et al., 2001; Tanaka et al., 2004; Nakata et al., 2018), as the RDP panel employed in the present study belongs to *indica* population. From the past studies, variation in *SSIIa* alleles were identified by exploring inter species genetic variation between *indica* compared to *japonica*, due to enrichment of the intermediate amylopectin chains (DP 12–24) (Umemoto et al., 2002, 2004). Similarly, SBEIIb revealed to possess different alleles in two subgroups *indica* and *japonica* (Luo et al., 2015). Allelic variants of both of the genes can markedly distinguish respective favored allele in *indica* vs. *japonica* germplasm. In the present study, underlying large effect candidate genes influencing ADH alone, but not AC, related to the identification of candidate gene metallic protease involved in the protein degradation and the allergenic protein, which might potentially involve in regulation of the structural proteins determining texture of cooked rice. In previous studies, protein content was correlated negatively with the adhesiveness (Lyon et al., 2000) and other rice texture attributes (Champagne et al., 1999; Martin and Fitzgerald, 2002). Besides the textural traits being highly correlated among each other, we also detected common QTNs between AC and ADH (**Tables 1**, **2**). Additionally, several significant SNPs involved in textural attributes found to influence alternative splicing were identified in this study, which suggests the importance of post-transcriptional regulation. Since most of the genes involved in the core regulatory pathways, the functional characterization of novel candidate genes with non-synonymous amino acid alteration influencing various textural attributes of ADH, HRD, and COH traits will be worth exploring its functional validation through transgenic studies. These novel haplotypes defined in the present study will serve as important genetic resource for future breeding strategies to capture textural attributes.

In summary, our findings addressed the underlying genetic basis of rice texture attributes. We found considerable phenotypic variations in texture attributes (adhesiveness, hardness, springiness, cohesiveness, and amylose content) among the 218 *indica* accessions. Highly negative correlation between amylose content and adhesiveness was observed, which could explain the fact that rice with low amylose is stickier. We identified multiple major/minor QTNs linked with rice cooking properties using SL-GWAS, followed by ML-GWAS using 4 independent methods (FASTmrEMMA, ISIS_EM-BLASSO, mrMLM, and pLARmEB). An important hot spot QTLs on chromosome 6 where QTNs for ADH co-localized with QTNs for AC were identified. Furthermore, a fine mapped genetic region on chromosome 8 affecting HRD was identified. The use of different models increased the number of variations captured across the diverse germplasm lines. Multi-locus model using different methods could overcome the limitations of single-locus analysis. Furthermore, this integrative approach has enabled the identification of novel small and large effect putative potential candidate genes and diagnostic haplotypes, which subsequently on validation, potentially be deployed in breeding to improve rice texture.

AUTHOR CONTRIBUTIONS

NS conceptualized the research work, designed and supervised all experiments, wrote and edited the manuscript. GM conducted GWAS and haplotype analyses, generated figures, and created an in-house functional annotation pipeline. SB and EM interpreted the data, prepared excel sheets, and developed the first draft of the manuscript. CD and RC conducted texture profile experiments, developed TPA methodology and processed the raw data. CL conducted texture profile experiments from different seasons for the contrasting haplotype lines.

FUNDING

This work was funded by the Global Rice Science Partnership (GRISP), the CGIAR Research Program on Rice Agri-Food Systems (RICE), Stress-Tolerant Rice for Africa and South Asia (STRASA) Phase III of BMGF funding.

ACKNOWLEDGMENTS

The authors acknowledge support of IRRI staff, Lucena Samadio for technical assistance in conducting texture profile analyses; Lilia Molina and the staff of the Grain Quality and Nutrition Service Laboratory (GQNSL) for performing amylose content and Roslen Anacleto, Roldan Ilagan, Artemio Madrid, Jr., and Fernando Salisi for growing the core collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01405/full#supplementary-material

REFERENCES

Anacleto, R., Cuevas, R. P., Jimenez, R., Llorente, C., Nissila, E., Henry, R., et al. (2015). Prospects of breeding high-quality rice using post-genomic tools. *Theor. Appl. Genet.* 128, 1449–1466. doi: 10.1007/s00122-015-2537-6

Bao, J., Sun, M., and Corke, H. (2002). Analysis of genetic behavior of some starch properties in *indica* rice (*Oryza sativa* L.): thermal properties, gel texture, swelling volume. *Theor. Appl. Genet.* 104, 408–413. doi: 10.1007/s001220100688

Bao, J.-S., Kong, X., Xie, J., and Xu, L. (2004). Analysis of genotypic and environmental effects on rice starch. 1. Apparent amylose content, pasting viscosity, and gel texture. *J. Agric. Food Chem.* 52, 6010–6016. doi: 10.1021/jf049234i

Bao, J.-S., Shen, S., Sun, M., and Corke, H. (2006). Analysis of genotypic diversity in the starch physicochemical properties of nonwaxy rice: apparent amylose content, pasting viscosity and gel texture. *Starch Stärke* 58, 259–267. doi: 10.1002/star.200500469

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

Bett-Garber, K. L., Champagne, E. T., McClung, A. M., Moldenhauer, K. A., Linscombe, S. D., and McKenzie, K. S. (2001). Categorizing rice cultivars based on cluster analysis of amylose content, protein content and sensory attributes. *Cereal Chem.* 78, 551–558. doi: 10.1094/CCHEM.2001.78.5.551

Bhattacharya, K., and Juliano, B. (1985). *Rice: Chemistry and Technology*. St Paul, MN: AACC.

Butardo, V. M., Anacleto, R., Parween, S., Samson, I., De Guzman, K., Alhambra, C. M., et al. (2017). Systems genetics identifies a novel regulatory domain of amylose synthesis. *Plant Physiol.* 173, 887–906. doi: 10.1104/pp.16.01248

Campbell, M. T., Du, Q., Liu, K., Brien, C. J., Berger, B., Zhang, C., et al. (2017). A comprehensive image-based phenomic analysis reveals the complex genetic architecture of shoot growth dynamics in rice (*Oryza sativa*). *Plant Genome* 10, 1–14. doi: 10.3835/plantgenome2016.07.0064

Champagne, E. T., Bett, K. L., Vinyard, B. T., Mcclung, A. M., Barton, F. E., Moldenhauer, K., et al. (1999). Correlation between cooked rice texture and rapid visco analyser measurements. *Cereal Chem.* 76, 764–771. doi: 10.1094/CCHEM.1999.76.5.764

Champagne, E. T., Bett-Garber, K. L., Fitzgerald, M. A., Grimm, C. C., Lea, J., Ohtsubo, K. I., et al. (2010). Important sensory properties differentiating premium rice varieties. *Rice* 3, 270–281. doi: 10.1007/s12284-010-9057-4

Champagne, E. T., Bett-Garber, K. L., McClung, A. M., and Bergman, C. J. (2004). Sensory characteristics of diverse rice cultivars as influenced by genetic and environmental factors. *Cereal Chem.* 81, 237–243. doi: 10.1094/CCHEM.2004.81.2.237

Cho, Y. G., Kang, H. J., Lee, Y. T., Jong, S. K., Eun, M. Y., and McCouch, S. R. (2010). Identification of quantitative trait loci for physical and chemical properties of rice grain. *Rice* 4, 61–73. doi: 10.1007/s11816-009-0120-9

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Cuevas, R. P., Pede, V. O., McKinley, J., Velarde, O., and Demont, M. (2016). Rice grain quality and consumer preferences: a case study of two rural towns in the Philippines. *PLoS ONE* 11:e0150345. doi: 10.1371/journal.pone.0150345

Fusi, N., Lippert, C., Lawrence, N. D., and Stegle, O. (2014). Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nat. Commun.* 5:4890. doi: 10.1038/ncomms5890

He, P., Li, S., Qian, Q., Ma, Y., Li, J., Wang, W., et al. (1999). Genetic analysis of rice grain quality. *Theor. Appl. Genet.* 98, 502–508. doi: 10.1007/s001220051098

Holmquist, M. (2000). Alpha beta-hydrolase fold enzymes structures, functions and mechanisms. *Curr. Protein Peptide Sci.* 1, 209–235. doi: 10.2174/1389203003381405

Hori, K., Suzuki, K., Iijima, K., and Ebana, K. (2016). Variation in cooking and eating quality traits in Japanese rice germplasm accessions. *Breed. Sci.* 66, 309–318. doi: 10.1270/jsbbs.66.309

Hsu, Y. C., Tseng, M. C., Wu, Y. P., Lin, M. Y., Wei, F. J., Hwu, K. K., et al. (2014). Genetic factors responsible for eating and cooking qualities of rice grains in a recombinant inbred population of an inter-subspecific cross. *Mol. Breed.* 34, 655–673. doi: 10.1007/s11032-014-0065-8

Ingvarsson, P. K., and Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytol.* 189, 909–922. doi: 10.1111/j.1469-8137.2010.03593.x

International Organization for Standardization (2007). *ISO 6647-1: 2007–Rice—Determination of amylose content—Part 1: Reference Method.*

Isshiki, M., Morino, K., Nakajima, M., Okagaki, R. J., Wessler, S. R., Izawa, T., et al. (1998). A naturally occurring functional allele of the rice waxy locus has a GT to TT mutation at the 5′ splice site of the first intron. *Plant J.* 15, 133–138. doi: 10.1046/j.1365-313X.1998.00189.x

Juliano, B. O. (1985). "Criteria and tests for rice grain qualities," in *Rice Chemistry and Technology, 2nd Edn*, ed B. O. Juliano (St. Paul, MN: American Association of Cereal Chemists, Inc.), 443–524.

Juliano, B. O. (1992). Structure, chemistry, and function of the rice grain and its fractions. *Cereal Foods World* 37, 772–774.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Kaosa-Ard, M., and Juliano, B. O. (1991). "Assessing rice quality characteristics and prices in selected international markets," in *Rice Grain Marketing and Quality Issues,* eds B. O. Juliano, L. R. Pollard, and G. Argosino (Manila: International Rice Research Institute), 23–36.

Kong, X., Chen, Y., Zhu, P., Sui, Z., Corke, H., and Bao, J.-S. (2015). Relationships among genetic, structural, and functional properties of rice starch. *J. Agric. Food Chem.* 63, 6241–6248. doi: 10.1021/acs.jafc.5b02143

Kumar, V., Singh, A., Mithra, S. A., Krishnamurthy, S., Parida, S. K., Jain, S., et al. (2015). Genome-wide association mapping of salinity tolerance in rice (*Oryza sativa*). *DNA Res.* 22, 133–145. doi: 10.1093/dnares/dsu046

Lawless, H. T., and Heymann, H. (2010). "Sensory evaluation of food: principles and practices," in *Food Science Text Series*, *2nd Edn*, eds H. T. Lawless and H. Heymann (New York, NY: Springer-Verlag), 596.

Li, H., Fitzgerald, M. A., Prakash, S., Nicholson, T. M., and Gilbert, R. G. (2017). The molecular structural features controlling stickiness in cooked rice, a major palatability determinant. *Sci. Rep.* 7:43713. doi: 10.1038/srep43713

Li, H., Prakash, S., Nicholson, T. M., Fitzgerald, M. A., and Gilbert, R. G. (2016). Instrumental measurement of cooked rice texture by dynamic rheological testing and its relation to the fine structure of rice starch. *Carbohydr. Polym.* 146, 253–263. doi: 10.1016/j.carbpol.2016.03.045

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12:e1005767. doi: 10.1371/journal.pgen.1005767

Luo, J., Jobling, S. A., Millar, A., Morell, M. K., and Li, Z. (2015). Allelic effects on starch structure and properties of six starch biosynthetic genes in a rice recombinant inbred line population. *Rice* 8:15. doi: 10.1186/s12284-015-0046-5

Lyon, B. G., Champagne, E. T., Vinyard, B. T., and Windham, W. R. (2000). Sensory and instrumental relationships of texture of cooked rice from selected cultivars and postharvest handling practices. *Cereal Chem.* 77, 64–69. doi: 10.1094/CCHEM.2000.77.1.64

Ma, L., Liu, M., Yan, Y., Qing, C., Zhang, X., Zhang, Y., et al. (2018). Genetic dissection of maize embryonic callus regenerative capacity using multi-locus genome-wide association studies. *Front. Plant Sci.* 9:561. doi: 10.3389/fpls.2018.00561

Martin, M., and Fitzgerald, M. (2002). Proteins in rice grains influence cooking properties! *J. Cereal Sci.* 36, 285–294. doi: 10.1006/jcrs.2001.0465

McCouch, S. R., Wright, M. H., Tung, C. W., Maron, L. G., McNally, K. L., Fitzgerald, M., et al. (2016). Open access resources for genome-wide association mapping in rice. *Nat. Commun.* 7:10532. doi: 10.1038/ncomms10532

Mestres, C., Ribeyre, F., Pons, B., Fallet, V., and Matencio, F. (2011). Sensory texture of cooked rice is rather linked to chemical than to physical characteristics of raw grain. *J. Cereal Sci.* 53, 81–89. doi: 10.1016/j.jcs.2010.10.001

Meullenet, J. F. C., Gross, J., Marks, B. P., and Daniels, M. (1998). Sensory descriptive texture analyses of cooked rice and its correlation to instrumental parameters using an extrusion cell. *Cereal Chem.* 75, 714–720. doi: 10.1094/CCHEM.1998.75.5.714

Misra, G., Badoni, S., Anacleto, R., Graner, A., Alexandrov, N., and Sreenivasulu, N. (2017). Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice. *Sci. Rep.* 7:12478. doi: 10.1038/s41598-017-12778-6

Nakamura, Y., Francisco, P. B., Hosaka, Y., Sato, A., Sawada, T., Kubo, A., et al. (2005). Essential amino acids of starch synthase IIa differentiate amylopectin structure and starch quality between japonica and indica rice varieties. *Plant Mol. Biol.* 58, 213–227. doi: 10.1007/s11103-005-6507-2

Nakata, M., Miyashita, T., Kimura, R., Nakata, Y., Takagi, H., Kuroda, M., et al. (2018). MutMapPlus identified novel mutant alleles of a rice starch branching enzyme II b gene for fine-tuning of cooked rice texture. *Plant Biotechnol. J.* 16, 111–123. doi: 10.1111/pbi.12753

Nishi, A., Nakamura, Y., Tanaka, N., and Satoh, H. (2001). Biochemical and genetic analysis of the effects ofamylose-extender mutation in rice endosperm. *Plant Physiol.* 127, 459–472. doi: 10.1104/pp.010127

Ohtsubo, K., Siscar, J. J. H., Juliano, B. O., Iwasaki, T., and Yakoo, M. (1990). Comparative study of texturometer and Instron texture measurements on cooked Japanese milled rices. *Rep. Natl. Food Res. Inst.* 54, 1–9.

Perez, C. M., Pascual, C. G., and Juliano, B. O. (1979). Eating quality indicators for waxy rices. *Food Chem.* 4, 179–184. doi: 10.1016/0308-8146(79)90002-5

Ramesh, M., Bhattacharya, K., and Mitchell, J. (2000). Developments in understanding the basis of cooked-rice texture. *Crit. Rev. Food Sci. Nutr.* 40, 449–460. doi: 10.1080/10408690091189220

Reddy, K. R., Ali, S. Z., and Bhattacharya, K. R. (1993). The fine structure of rice-starch amylopectine and its relation to the texture of cooked rice. *Carbohydr. Polym.* 22, 267–275. doi: 10.1016/0144-8617(93)90130-V

Reddy, K. R., Subramanian, R., Ali, S. Z., and Bhattacharya, K. R. (1994). Viscoelastic properties of rice-flour pastes and their relationship to amylose content and rice quality. *Cereal Chem.* 71, 548–552.

Sant'ana, G. C., Pereira, L. F., Pot, D., Ivamoto, S. T., Domingues, D. S., Ferreira, R. V., et al. (2018). Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in *Coffea arabica* L. *Sci. Rep.* 8:465. doi: 10.1038/s41598-017-18800-1

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

Sesmat, A., and Meullenet, J. F. (2001). Prediction of rice sensory texture attributes from a single compression test, multivariate regression, and a stepwise model optimization method. *J. Food Sci.* 66, 124–131. doi: 10.1111/j.1365-2621.2001.tb15593.x

Shi, C., Zhu, J., Zang, R., and Chen, G. (1997). Genetic and heterosis analysis for cooking quality traits of indica rice in different environments. *Theor. Appl. Genet.* 95, 294–300. doi: 10.1007/s001220050562

Stokes, J. R., Boehm, M. W., and Baier, S. K. (2013). Oral processing, texture and mouthfeel: from rheology to tribology and beyond. *Curr. Opin. Colloid Interface Sci.* 18, 349–359. doi: 10.1016/j.cocis.2013.04.010

Suwannaporn, P., Pitiphunpong, S., and Champangern, S. (2007). Classification of rice amylose content by discriminant analysis of physicochemical properties. *Starch Stärke* 59, 171–177. doi: 10.1002/star.200600565

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tanaka, N., Fujita, N., Nishi, A., Satoh, H., Hosaka, Y., Ugaki, M., et al. (2004). The structure of starch can be manipulated by changing the expression levels of starch branching enzyme IIb in rice endosperm. *Plant Biotechnol. J.* 2, 507–516. doi: 10.1111/j.1467-7652.2004.00097.x

Turner, (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Sour. Softw.* 3:731. doi: 10.21105/joss.00731

Umemoto, T., Aoki, N., Lin, H., Nakamura, Y., Inouchi, N., Sato, Y., et al. (2004). Natural variation in rice starch synthase IIa affects enzyme and starch properties. *Funct. Plant Biol.* 31, 671–684. doi: 10.1071/FP04009

Umemoto, T., Yano, M., Satoh, H., Shomura, A., and Nakamura, Y. (2002). Mapping of a gene responsible for the difference in amylopectin structure between japonica-type and indica-type rice varieties. *Theor. Appl. Genet.* 104, 1–8. doi: 10.1007/s001220200000

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wang, X., Pang, Y., Zhang, J., Wu, Z., Chen, K., Ali, J., et al. (2017). Genome-wide and gene-based association mapping for rice eating and cooking characteristics and protein content. *Sci. Rep.* 7:17203. doi: 10.1038/s41598-017-17347-5

Wang, Z. Y., Zheng, F. Q., Shen, G. Z., Gao, J. P., Snustad, D. P., Li, M. G., et al. (1995). The amylose content in rice endosperm is related to the post-transcriptional regulation of the *Waxy* gene. *Plant J.* 7, 613–622. doi: 10.1046/j.1365-313X.1995.7040613.x

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinformatics* 19, 700–712. doi: 10.1093/bib/bbw145

Windham, W. R., Lyon, B. G., Champagne, E. T., Barton, F. E., Webb, B. D., McClung, A. M., et al. (1997). Prediction of cooked rice texture quality using near-infrared reflectance analysis of whole-grain milled samples. *Cereal Chem.* 74, 626–632. doi: 10.1094/CCHEM.1997.74.5.626

Xiao, Y., Liu, H., Wu, L., Warburton, M., and Yan, J. (2017). Genome-wide association studies in maize: praise and stargaze. *Mol. Plant* 10, 359–374. doi: 10.1016/j.molp.2016.12.008

Yang, F., Chen, Y., Tong, C., Huang, Y., Xu, F., Li, K., et al. (2014). Association mapping of starch physicochemical properties with starch synthesis-related gene markers in nonwaxy rice (*Oryza sativa* L.). *Mol. Breed.* 34, 1747–1763. doi: 10.1007/s11032-014-0135-y

Zhang, J., Feng, J., Ni, Y., Wen, Y., Niu, Y., Tamba, C., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y., Liu, P., Zhang, X., Zheng, Q., Chen, M., Ge, F., et al. (2018). Multi-locus genome-wide association study reveals the genetic architecture of stalk lodging resistance-related traits in maize. *Front. Plant Sci.* 9:611. doi: 10.3389/fpls.2018.00611

Zhao, X., and Fitzgerald, M. (2013). Climate change: implications for the yield of edible rice. *PLoS ONE* 8:e66218. doi: 10.1371/journal.pone.0066218

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 04 October 2018
doi: 10.3389/fpls.2018.01464

The Application of Multi-Locus GWAS for the Detection of Salt-Tolerance Loci in Rice

Yanru Cui, Fan Zhang and Yongli Zhou**

Institute of Crop Sciences/National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, Beijing, China

Improving the salt-tolerance of direct-seeding rice at the seed germination stage is a major goal of breeders. Efficiently identifying salt tolerance loci will help researchers develop effective rice breeding strategies. In this study, six multi-locus genome-wide association studies (GWAS) methods (mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, pKWmEB, and ISIS EM-BLASSO) were applied to identify quantitative trait nucleotides (QTNs) for the salt tolerance traits of 478 rice accessions with 162,529 SNPs at the seed germination stage. Among the 371 QTNs detected by the six methods, 56 were identified by at least three methods. Among these 56 QTNs, 12, 6, 7, 4, 13, 12, and 12 were found to be associated with SSI-GI, SSI-VI, SSI-MGT, SSI-IR-24h, SSI-IR-48h, SSI-GR-5d, and SSI-GR-10d, respectively. Additionally, 66 candidate genes were identified in the vicinity of the 56 QTNs, and two of these genes (LOC_Os01g45760 and LOC_Os10g04860) are involved in auxin biosynthesis according to the enriched GO terms and KEGG pathways. This information will be useful for identifying the genes responsible for rice salt tolerance. A comparison of the six methods revealed that ISIS EM-BLASSO identified the most co-detected QTNs and performed best, with the smallest residual errors and highest computing speed, followed by FASTmrMLM, pLARmEB, mrMLM, pKWmEB, and FASTmrEMMA. Although multi-locus GWAS methods are superior to single-locus GWAS methods, their utility for identifying QTNs may be enhanced by adding a bin analysis to the models or by developing a hybrid method that merges the results from different methods.

Keywords: multi-locus, GWAS, QTNs, salt tolerance, rice

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Qishan Wang,
Shanghai Jiao Tong University, China
Chenwu Xu,
Yangzhou University, China
Gao Huijiang,
Chinese Academy of Agricultural
Sciences, China

****Correspondence:***
Fan Zhang
zhangfan03@caas.cn
Yongli Zhou
zhouyongli@caas.cn

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *31 July 2018*
Accepted: *14 September 2018*
Published: *04 October 2018*

Citation:
Cui Y, Zhang F and Zhou Y (2018)
The Application of Multi-Locus GWAS
for the Detection of Salt-Tolerance
Loci in Rice. Front. Plant Sci. 9:1464.
doi: 10.3389/fpls.2018.01464

INTRODUCTION

A genome-wide association studies (GWAS) represents a powerful option for the genetic characterization of quantitative traits, and has been widely used for analyzing agronomic traits related to plants. Numerous genetic variants for complex traits have been identified based on single-locus GWAS methods, such as empirical Bayes, efficient mixed model association (EMMA), genome-wide efficient mixed linear model association (GEMMA), settlement of mixed linear

Abbreviations: GI, germination index; GR-10d, germination rate at the 10th day; GR-5d, germination rate at 5th day; IR-24h, imbibition rate at 24 h; IR-48h, imbibition rate at 48 h; MGT, mean germination time; N, normal condition; S, salt stress condition; SSI, stress-susceptibility index; VI, vigor index.

model under progressively exclusive relationship (SUPER), and mixed linear model (MLM) (Kang et al., 2008; Zhou and Stephens, 2012; Wang et al., 2014, 2016a). Although the statistical power of quantitative trait nucleotide (QTN) detection improves after controlling the polygenic background, most of the small effects associated with complex traits are still not captured by single-locus GWAS methods.

In a single-locus GWAS model, markers are tested individually in a one-dimensional genome scan. Moreover, the multiple test correction for the critical value of a significance test should be considered. Bonferroni correction is widely used to modify the threshold value to control the false positive rate (FPR). However, this type of correction method is so conservative that true QTNs may be eliminated. Therefore, the best way to solve this problem is to develop a multi-locus GWAS method that does not require a multiple test correction. Multi-locus GWAS methods involve a multi-dimensional genome scan, in which the effects of all markers are simultaneously estimated. Many penalized multi-locus GWAS methods have been developed, including the least absolute shrinkage and selection operator (LASSO), empirical Bayes LASSO, and adaptive mixed LASSO (Yi and Xu, 2008; Cho et al., 2009, 2010; Wu et al., 2009; Ayers and Cordell, 2010; Wang et al., 2010; Giglio and Brown, 2018). These methods can minimize some marker effects to zero when the number of single nucleotide polymorphisms (SNPs) is not much larger than the sample size. However, the rapid development of sequencing technologies has enabled the detection of many SNPs (i.e., the number of SNPs is hundreds of times larger than the sample size). Thus, the available methods for minimizing marker effects are ineffective. One option for addressing this issue involves decreasing the number of SNPs. Dr. Zhang' lab developed an R package called mrMLM, which includes the following six multi-locus GWAS methods: mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, pKWmEB, and ISIS EM-BLASSO. All of these methods involve two-step algorithms. During the first step, a single-locus GWAS method is applied to scan the entire genome, and putative QTNs are detected according to a less stringent critical value, such as $P < 0.005$ or $P < 1/m$, where m is the number of markers. During the second step, all selected putative QTNs are examined by a multi-locus GWAS model to detect true QTNs (Wang et al., 2016a,b; Tamba et al., 2017; Zhang et al., 2017; Ren et al., 2018; Wen et al., 2018a,b; Zhang and Tamba, 2018). The mrMLM package solves the problem associated with co-factor selection in the multi-locus GWAS model when there are many markers.

Rice (*Oryza sativa* L.), which is one of the most important cereal crops worldwide, is sensitive to salt stress. With the increasing salinization of soils, salt stress is becoming a key abiotic factor limiting rice production that rice breeders must overcome (Hu et al., 2012). Developing salt-tolerant rice cultivars is an efficient way to minimize crop loss. Over the past several years, high density SNPs have been used to detect variants with GWAS methods to improve rice varieties (Han and Huang, 2013; Chen et al., 2014; Yang et al., 2014; Wei et al., 2017). However, most traits related to abiotic stress tolerance are controlled by several polygenes that are undetectable in single-locus GWAS models (Lee et al., 2003; Cui et al., 2015). Therefore, we should apply multi-locus GWAS methods to identify loci related to salt tolerance. In this study, 478 rice accessions, each with seven salt stress susceptibility index (SSI)-related traits, and 162,529 SNPs were used to conduct a multi-locus GWAS. Our objectives were to identify the significant QTNs related to salt tolerance and provide recommendations regarding the selection of a multi-locus GWAS method by comparing the differences among the six multi-locus methods included in the mrMLM package.

MATERIALS AND METHODS

Rice Phenotypic Data Related to Salt Tolerance

We analyzed 478 rice accessions from 46 countries and regions regarding seven salt tolerance-related traits at the seed germination stage in a multi-locus GWAS. Phenotypic data were collected for control and stress-treated plants incubated in a growth chamber, with two independent experiments conducted for the control and stress treatments. Each independent experiment involved a randomized block design with two replicates. The dataset was published by Shi et al. (2017), and the seven salt tolerance-related traits were VI, GI, germination rate (GR) at days 5 and 10, MGT, and imbibition rate (IR) at 24 and 48 h. All salt tolerance-related traits were measured for plants treated with 60 mM NaCl or water (control) as follows: IR (mg/g) was calculated as $IR = (W_2 - W_1)/W_1 \times 1000$ at 24 and 48 h after starting the incubation, where W_1 represents the dry seed weight and W_2 represents the imbibed seed weight; GR was calculated as $GR = N_t/N_0 \times 100\%$ at days 5 and 10, where N_t is the number of germinated seeds at day t and N_0 is the total number of seeds; GI was calculated as $GI = \sum(G_{\mathrm{t}}/T_{\mathrm{t}})$, where G_t is the accumulated number of germinated seeds at day t and T_t is the time (in days); MGT was calculated as $MGT = \sum T_{\mathrm{i}}N_{\mathrm{i}}/\sum N_i$, where N_i is the number of newly germinated seeds at day t and T_i is the time (in days); VI was calculated as $VI = GI \times SL$, where SL is the average shoot length of 10 germinated seeds at day 10. The salt tolerance level of rice during the germination stage was estimated with the following equation: $SSI = (1 - Y_{\mathrm{s}}/Y_{\mathrm{p}})/D$, where Y_s is the performance of an individual under the stress condition, Y_p is the performance of an individual under the normal condition, and D is the stress intensity, which was calculated as $D = 1 - (\sum Y_{\mathrm{s}}/\sum Y_{\mathrm{p}})$. Finally, 21 traits were included in this study. The abbreviated names of these 21 traits are provided in the abbreviations list.

Genotyping and Multi-Locus GWAS

The 478 rice accessions analyzed in this study were from the 3K rice genome project. The 3K rice genome project 404K coreSNP dataset from the Rice-Seek Database was downloaded from http://snp-seek.irri.org/_download.zul (Alexandrov et al., 2015). We used the PLINK program (version 1.9) (Purcell et al., 2007) to obtain a subset of 162,529 SNPs with a minor allele frequency > 5% and a missing data ratio < 0.1 for association analyses. The kinship matrix (K matrix) was calculated based on the genotype marker information described by Xu (2013). The mrMLM package, including six multi-locus GWAS methods, was downloaded from http://cran.r-project.org/web/packages/mrMLM/index.html. Default values were used for all parameters.

Annotation of Candidate Genes and Pathway Enrichment Analysis

Synonymous and non-synonymous SNPs and SNPs associated with large-effect changes were annotated using the snpEff program (version 4.0) (Cingolani et al., 2012) based on the gene models of the annotated Nipponbare reference genome (IRGSP 1.0) (Kawahara et al., 2013). All putative SNPs located within genes and annotation details have been published (Kawahara et al., 2013). Enriched gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were identified using the agriGO (version 2.0) (Tian et al., 2017) and EXPath 2.0 (Chien et al., 2015) programs, respectively.

RESULTS

Heritability and Variance

The heritability and residual errors estimated by the six multi-locus GWAS methods are presented in **Table 1**. The narrow sense heritability ranged from 0.17 for S_MGT and 0.57 for S_IR_48h. A comparison of the residual errors among the six multi-locus GWAS models revealed that the residual error estimated by FASTmrEMMA was the largest under the normal condition when the phenotypic variation was larger than 10. Under the salt stress condition, the largest residual errors for traits S_IR_24h and S_IR_48h were observed from FASTmrEMMA. Regarding the SSI-related traits, the largest residual error was estimated by FASTmrEMMA. The salt tolerance level was evaluated according to the SSI-related traits. Lower SSI values indicated a higher tolerance to salt stress. The results of the correlation analyses of the seven SSI-related traits are presented in **Figure 1A**. There were significant positive correlations among SSI_VI, SSI_GR_5d, SSI _GR_10d, and SSI_GI. The correlation coefficients between SSI-VI and the other three SSI-related traits, namely SSI_GR_5d, SSI_GR_10d, and SSI_GI, were 0.91, 0.91, and 0.96, respectively. Meanwhile, the correlation coefficients for SSI_GR_5d, SSI_GR_10d, and SSI_GI were 0.89, 0.95, and 0.96, respectively. The high correlation among the four SSI-related traits implied that some novel loci might be simultaneously detected for different traits.

QTNs Associated With Salt Tolerance at the Germination Stage Identified by a Multi-Locus GWAS

Using the six multi-locus GWAS methods in the mrMLM package (**Supplementary Table S1**), we identified 371 significant QTNs for the salt tolerance-related traits (SSI-VI, SSI-GR, SSI-IR, SSI-MGT, and SSI-GI) based on a logarithm of odds (LOD) threshold of ≥3. Of these QTNs, 41, 41, 27, 63, 56, 41, and 151 were found to be associated with SSI-GI, SSI-VI, SSI-MGT, SSI-IR-24h, SSI-IR-48h, SSI-GR-5d, and SSI-GR-10d, respectively, with the QTNs explaining 0.57 ∼ 9.80, 0.54 ∼ 8.97, 0.64 ∼ 8.21, 0.01 ∼ 4.94, 0.37 ∼ 8.93, 0.9 ∼ 6.72, and 0.7 ∼ 6.08 (%) of the phenotypic variations, respectively [i.e., phenotypic variation explained (PVE) values] (**Supplementary Table S1** and **Supplementary Figure S1**). Additionally, 3, 9, and 22 QTNs were associated with four, three, and two salt tolerance-related traits, respectively, which explained the high correlation among SSI_VI, SSI_GR_5d, SSI _GR_10d, and SSI_GI (**Figure 1B**).

In this study, 110 and 56 QTNs were co-detected by at least two and three methods, respectively (**Supplementary Table S2** and **Table 2**). Among the 56 QTNs, 12 that were located on chromosomes 1, 2, 3, 6, 8, 9, 11, and 12 were identified to be associated with SSI-GI, of which 11 were identified by ISIS EM-BLASSO, while 10, 9, 8, 7, and 3 were detected by FASTmrMLM, mrMLM, pKWmEB, pLARmEB, and FASTmrEMMA, respectively. Four of the 12 QTNs were simultaneously detected by five methods. Of these four QTNs, rs3_29294598, rs6_30827714, and rs8_24915626, were simultaneously detected by mrMLM, FASTmrMLM, pLARmEB, pKWmEB, and ISIS EM-BLASSO, with PVE values of 2.45 ∼ 5.01, 1.19 ∼ 2.82, and 1.44 ∼ 4.48 (%), respectively. Meanwhile, rs8_27233581 was simultaneously detected by mrMLM, FASTmrMLM, FASTmrEMMA, pKWmEB, and ISIS EM-BLASSO, with a PVE value of 2.28 ∼ 6.28 (%). Six QTNs related to SSI-VI were detected on chromosomes 5, 6, 8, 10, and 11, five of which were identified by mrMLM and pKWmEB, with LOD values of 3.22 ∼ 7.16 and 3.11 ∼ 7.11, respectively. Only one QTN was detected by ISIS EM-BLASSO, with an LOD value of 8.59. Seven QTNs located on chromosomes 1, 2, 4, 6, 9, and 11 were correlated with SSI-MGT. All seven of these QTNs were detected by ISIS EM-BLASSO and pKWmEB, with LOD values of 3.18 ∼ 7.97 and 3.54 ∼ 6.62, respectively. The mrMLM, FASTmrMLM, FASTmrEMMA, and pLARmEB methods detected 3, 5, 1, and 2 QTNs related to SSI-MGT, respectively. Among the seven QTNs, rs1_15357371 was identified by all methods, except for mrMLM, with a PVE value of 2.95 ∼ 5.64 (%). For SSI-IR-24h, four significant QTNs were detected on chromosomes 4, 6, and 9 by mrMLM, pKWmEB, and ISIS EM-BLASSO, with LOD values of 6.97 ∼ 18.97, 3.42 ∼ 7.16, and 3.90 ∼ 10.18, respectively. Two of these QTNs were identified by FASTmrMLM, while none of the QTNs were detected by FASTmrEMMA and pLARmEB. Thirteen QTNs located on chromosomes 1, 2, 3, 4, 6, 7, 10, 11, and 12 were associated with SSI-IR-48h, including 10 that were detected by ISIS EM-BLASSO, with LOD values of 3.54 ∼ 10.0, and nine QTNs that were identified by FASTmrMLM, pLARmEB, and pKWmEB, with LOD values of 3.29 ∼ 6.51, 3.58 ∼ 6.1, and 5.04 ∼ 9.04, respectively. The mrMLM and FASTmrEMMA methods separately detected eight and six QTNs, with LOD values of 3.14 ∼ 6.68 and 3.39 ∼ 6.97, respectively. Of the 13 QTNs, rs1_5453364, rs11_28865880, and rs12_19111880 were identified by all six methods, with PVE values of 0.86 ∼ 2.16, 1.38 ∼ 4.83, and 0.62 ∼ 2.97 (%), respectively. Moreover, 12 QTNs associated with SSI-GR-5d were detected on chromosomes 1, 3, 5, 7, 8, 9, 10, and 11. Of these QTNs, nine, eight, seven, six, six, and four QTNs were separately detected by pLARmEB, FASTmrMLM, mrMLM, pKWmEB, FASTmrEMMA, and ISIS EM-BLASSO, respectively, with LOD values of 3.26 ∼ 7.57, 3.61 ∼ 5.96, 3.03 ∼ 6.43, 3.34 ∼ 6.13,

TABLE 1 | Phenotypic variance, estimated residual error, and heritability of 21 rice traits.

Trait	PV	Heritability (%)	Residual error					
			FASTmrEMMA	FASTmrMLM	ISIS EM-BLASSO	mrMLM	pKWmEB	pLARmEB
S_GI	0.17	42	0.13	0.13	0.11	0.10	0.12	0.13
S_VI	2.64	41	2.11	1.63	1.39	1.55	1.48	1.70
S_MGT	1.03	17	0.85	0.74	0.72	0.86	0.66	0.85
S_IR_24h	53351.6322	51	36552.52	23546.99	22673.23	26406.72	21410.21	24658.42
S_IR_48h	52655.13	56	32734.46	24011.68	21900.10	23275.70	24914.75	19703.57
S_GR_5d	1158.11	34	900.85	935.09	690.32	765.26	746.93	815.23
S_GR_10d	1233.41	35	844.53	961.70	893.90	796.73	719.22	907.11
N_GI	0.07	43	0.05	0.04	0.04	0.04	0.04	0.04
N_VI	5.84	34	3.72	3.10	3.00	3.86	2.72	3.18
N_MGT	0.82	50	0.57	0.48	0.36	0.41	0.43	0.49
N_IR_24h	63786.10	55	41714.56	30057.04	26812.22	28640.95	30489.47	31072.55
N_IR_48h	66260.99	48	44506.99	30736.87	23802.86	34238.02	32400.73	26220.37
N_GR_5d	452.67	27	326.57	307.38	233.09	330.44	267.86	285.44
N_GR_10d	86.61	31	66.58	58.11	48.46	59.51	51.38	52.48
SSI_GI	0.41	33	0.31	0.25	0.22	0.27	0.25	0.28
SSI_VI	0.10	22	0.08	0.06	0.05	0.08	0.06	0.07
SSI_MGT	4.07	32	3.61	3.02	2.57	2.75	2.66	3.22
SSI_IR_24h	199.2613	15	176.41	164.03	143.96	23.17	121.67	173.06
SSI_IR_48h	13.76	42	10.59	8.46	6.62	7.95	7.18	8.67
SSI_GR_5d	0.54	27	0.44	0.40	0.32	0.39	0.30	0.38
SSI_GR_10d	0.4509	34	0.34	0.31	0.23	0.30	0.24	0.28

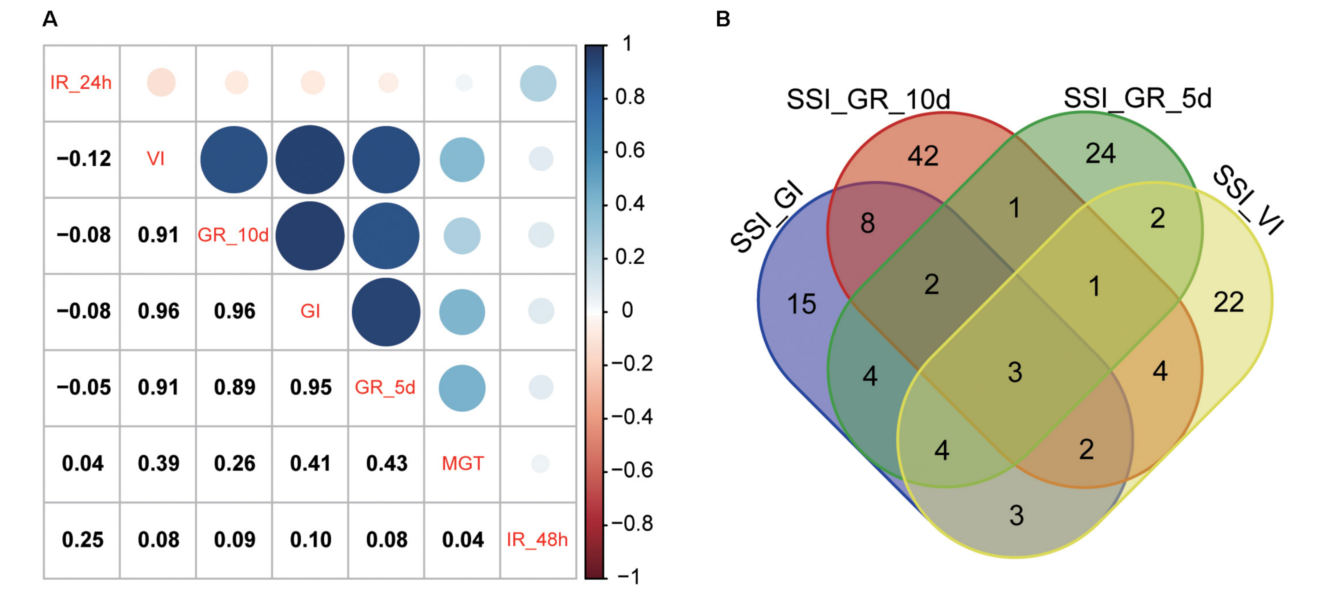


FIGURE 1 | Correlation among SSI-related traits **(A)** and a Venn diagram of the QTNs for four SSI-related traits **(B)** estimated by a multi-locus GWAS.

3.26 ∼ 6.57, and 3.09 ∼ 5.76, respectively. Three of the 12 QTNs, rs3_4264086, rs5_29609065, and rs11_27392033, were detected by five methods, with PVE values of 1.42 ∼ 4.47, 1.07 ∼ 4.65, and 0.96 ∼ 3.86 (%), respectively. For SSI-GR-10d, 12 QTNs were detected on chromosomes 1, 2, 4, 6, 7, 8, 9, 10, and 11. Of these 12 QTNs, rs10_22754603 and rs11_27380577 were identified by five methods, with PVE values of 0.93 ∼ 3.08 and 1.11 ∼ 4.4 (%), respectively (**Table 2**).

Validation of the Common QTNs

Among the 56 QTNs, 14 were identified by at least five methods, of which four, three, two, four, and one were associated with SSI_GI, SSI_GR_5d, SSI_GR_10d, SSI_IR_48h,

TABLE 2 | Significant QTNs for SSI-related traits in rice co-detected by at least three multi-locus GWAS methods.

Trait	SNPs[1]	Chromosome	Position	QTN effect	LOD score	PVE (%)[2]	Method[3]
SSI_GI	**rs1_11882948**	1	11882948	0.1 ∼ 0.11	3.67 ∼ 4.68	0.98 ∼ 1.55	2,3,5
	rs2_22250136	2	22250136	−0.16 ∼ −0.09	3.27 ∼ 4.43	0.61 ∼ 2.19	1,2,3,4
	rs2_24480757	2	24480757	0.08 ∼ 0.08	3.66 ∼ 4.01	0.98 ∼ 1.5	2,3,6
	rs3_29294598	3	29294598	0.08 ∼ 0.17	3.97 ∼ 6.21	2.45 ∼ 5.01	1,2,3,5,6
	rs6_30827714	6	30827714	−0.13 ∼ −0.09	3.15 ∼ 6.15	1.19 ∼ 2.82	1,2,3,5,6
	rs8_7832802	8	7832802	0.1 ∼ 0.21	3.48 ∼ 4.94	2.17 ∼ 4.69	1,3,4,6
	rs8_24915626	8	24915626	0.09 ∼ 0.16	3.12 ∼ 7.04	1.44 ∼ 4.48	1,2,3,5,6
	rs8_25014297	8	25014297	−0.35 ∼ −0.21	5.31 ∼ 10.4	4.56 ∼ 8.91	1,2,3,6
	rs8_27233581	8	27233581	0.1∼0.29	3.71 ∼ 7.67	2.28 ∼ 6.28	1,2,3,4,6
	rs9_5893568	9	5893568	0.05 ∼ 0.08	3.17 ∼ 3.43	0.57 ∼ 1.09	2,3,5
	rs11_17680260	11	17680260	0.19 ∼ 0.26	6.74 ∼ 10.53	4.9 ∼ 9.8	1,3,5,6
	rs12_21121298	12	21121298	−0.13 ∼ −0.09	3.04 ∼ 4.35	1.03 ∼ 2.06	1,2,5
SSI_VI	rs5_29590002	5	29590002	−0.1 ∼ 0.09	3.5 ∼ 4.36	1.26 ∼ 5.77	1,2,4,5
	rs6_26952785	6	26952785	0.1 ∼ 0.11	3.89 ∼ 8.59	3.22 ∼ 3.57	1,3,6
	rs8_27233581	8	27233581	0.06 ∼ 0.13	5.12 ∼ 5.45	2.38 ∼ 5.17	1,4,5,6
	rs10_2806159	10	2806159	0.08 ∼ 0.12	3.16 ∼ 5.42	1.85 ∼ 3.55	1,2,6
	rs10_11718859	10	11718859	−0.13 ∼ 3.56	4.28 ∼ 6.36	1.74 ∼ 3.2	2,4,5,6
	rs11_17680260	11	17680260	0.07 ∼ 0.1	3.12 ∼ 5.68	3.03 ∼7.16	1,5,6
SSI_MGT	rs1_15357371	1	15357371	0.43 ∼ 1.37	4.07 ∼ 6.54	2.95 ∼ 5.64	2,3,4,5,6
	rs2_23991498	2	23991498	0.26 ∼ 0.33	3.06 ∼ 6.62	1.61 ∼ 3.48	2,3,5,6
	rs4_13696726	4	13696726	−0.98 ∼ −0.44	3.54 ∼ 6.03	1.79 ∼ 4.88	1,3,6
	rs6_27962052	6	27962052	−0.43 ∼ −0.26	3.75 ∼ 4.55	1.06 ∼ 4.81	1,2,3,6
	rs9_4258702	9	4258702	−0.53 ∼ −0.41	5.71 ∼ 7.96	2.76 ∼ 3.35	2,3,6
	rs9_11450011	9	11450011	−0.36 ∼ −0.31	3.18 ∼ 5.53	1.94 ∼ 3.55	2,3,6
	rs11_24660808	11	24660808	−0.79 ∼ −0.44	3.3 ∼ 4.19	2.47 ∼ 2.88	1,3,6
SSI_IR_24h	rs4_31794832	4	31794832	2.59 ∼ 3.07	3.9 ∼ 6.97	0.11 ∼ 2.65	1,3,6
	rs6_5699431	6	5699431	2.33 ∼ 4.08	3.42 ∼ 10.18	0.09 ∼ 6.9	1,2,3,6
	rs9_12353804	9	12353804	−4.57 ∼ −3.5	4.95 ∼ 18.97	0.3 ∼ 5.3	1,3,6
	rs9_6746183	9	6746183	−10.84 ∼ −4.22	3.29 ∼ 16.12	0.97 ∼ 5.91	1,2,3,6
SSI_IR_48h	rs1_2103242	1	2103242	0.64 ∼ 0.96	3.39 ∼ 6.24	1.46 ∼ 3.92	1,2,4
	rs1_5453364	1	5453364	−1.26 ∼ 0.48	3.23 ∼ 6.79	0.86 ∼ 3.51	1,2,3,4,5,6
	rs1_31748567	1	31748567	−0.83 ∼ −0.56	4.04 ∼ 6.43	0.98 ∼ 3.55	1,3,5,6
	rs2_24073194	2	24073194	−0.62 ∼ 0.72	3.58 ∼ 5.04	0.62 ∼ 2.2	3,5,6
	rs3_20204466	3	20204466	0.89 ∼ 1.02	3.81 ∼ 6.73	0.82 ∼ 2.76	3,5,6
	rs4_4695323	4	4695323	−1.5 ∼ −0.78	3.29 ∼ 5.61	1.48 ∼ 1.9	2,3,6
	rs4_31202952	4	31202952	−0.67 ∼ −0.52	3.43 ∼ 5.63	1.61 ∼ 2.37	1,2,3
	rs6_1459330	6	1459330	0.62 ∼ 1.1	3.34 ∼ 5.38	1.51 ∼ 1.71	1,2,4
	rs7_21649301	7	21649301	1.08 ∼ 1.6	4.4 ∼ 6.08	1 ∼ 3.78	1,2,5
	rs10_10209541	10	10209541	0.67 ∼ 1.03	3.54 ∼ 6.61	1.51 ∼ 3.98	3,5,6
	rs11_28865880	11	28865880	0.7 ∼ 1.65	5.83 ∼ 10	1.38 ∼ 4.83	1,2,3,4,5,6
	rs12_7176832	12	7176832	−1.43 ∼ −0.84	4.43 ∼ 8.64	1.2 ∼ 4.91	2,3,4,5,6
	rs12_19111880	12	19111880	−1.66 ∼ −0.66	3.14 ∼ 7.23	0.62 ∼ 2.97	1,2,3,4,5,6
SSI_GR_5d	**rs1_11882948**	1	11882948	0.12 ∼ 0.15	3.03 ∼ 3.94	1.69 ∼ 1.99	1,2,6
	rs1_22648607	1	22648607	0.17 ∼ 0.23	4.49 ∼ 6.57	1.36 ∼ 1.85	3,5
	rs3_4264086	3	4264086	0.11 ∼ 0.21	3.97 ∼ 6.43	1.42 ∼ 4.47	1,2,4,5,6
	rs3_29294598	3	29294598	0.1 ∼ 0.11	3.59 ∼ 3.61	1.34 ∼ 1.72	2,5
	rs5_29609065	5	29609065	0.08 ∼ 0.22	3.34 ∼ 5.74	0.96 ∼ 3.86	1,2,4,5,6
	rs7_1171356	7	1171356	0.15 ∼ 0.28	4.48 ∼ 4.77	3.26 ∼ 3.58	1,4
	rs8_24915626	8	24915626	0.11 ∼ 0.15	3.31 ∼ 4.1	1.5 ∼ 3.01	2,3,5
	rs8_27233581	8	27233581	0.09 ∼ 0.28	3.09 ∼ 4.66	1.36 ∼ 3.33	3,4,6
	rs9_8174432	9	8174432	0.1 ∼ 0.15	3.26 ∼ 4.23	1.29 ∼ 2.84	1,2,5
	rs9_21139613	9	21139613	0.11 ∼ 0.19	3.58 ∼ 5.02	1.61 ∼ 4.79	1,5,6
	rs10_11718859	10	11718859	−0.23 ∼ −0.11	4.32 ∼ 5.76	1.17 ∼ 2.5	2,3,4,5

(Continued)

TABLE 2 | Continued

Trait	SNPs[1]	Chromosome	Position	QTN effect	LOD score	PVE (%)[2]	Method[3]
	rs11_27392033	11	27392033	−0.31 ∼ 0.12	3.93 ∼ 6.37	1.07 ∼ 4.65	1,2,4,5,6
SSI_GR_10d	rs1_3401561	1	3401561	−0.24 ∼ −0.14	3.21 ∼ 4.65	2.29 ∼ 6.08	1,3,5
	rs1_11882948	1	11882948	0.1 ∼ 0.12	3.35 ∼ 4.56	1.31 ∼ 1.72	2,3,6
	rs2_8009453	2	8009453	0.08 ∼ 0.15	3.34 ∼ 4.96	1.18 ∼ 3.31	1,2,3
	rs2_22247315	2	22247315	−0.17 ∼ −0.12	3.54 ∼ 6.51	0.95 ∼ 3.19	3,5,6
	rs4_19568498	4	19568498	0.16 ∼ 0.21	5.39 ∼ 6.37	1.9 ∼ 2.96	2,3,6
	rs6_26597879	6	26597879	0.11 ∼ 0.13	3.83 ∼ 5.41	2.44 ∼ 2.82	2,3,5
	rs7_3788168	7	3788168	0.22 ∼ 0.23	5.46 ∼ 7.54	2.54 ∼ 3.59	2,3,6
	rs7_22276671	7	22276671	0.09 ∼ 0.1	3.65 ∼ 4.66	1.57 ∼ 2.33	2,3,6
	rs8_24915626	8	24915626	0.13 ∼ 0.14	3.88 ∼ 5.8	2.07 ∼ 3.16	1,2,3
	rs9_21139613	9	21139613	0.1 ∼ 0.18	3.55 ∼ 8.69	1.92 ∼ 4.59	1,2,5,6
	rs10_22754603	10	22754603	0.08 ∼ 0.17	3.46 ∼ 5.94	0.93 ∼ 3.08	1,2,3,5,6
	rs11_27380577	11	27380577	−0.21 ∼ −0.13	3.24 ∼ 7.31	1.11 ∼ 4.4	1,2,3,5,6

[1] *SNPs in bold font are pleiotropic QTNs which were detected associate with multiple traits.*
[2] *PVE: Phenotypic variation explained.*
[3] *1:mrMLM; 2:FASTmrMLM; 3:ISIS EM-BLASSO; 4:FASTmrEMMA; 5:pLARmEB; 6:pKWmEB.*

and SSI_MGT, respectively. We divided the population into two groups according to allelic genotypes to test whether the mean phenotypes of the two groups were significantly different. The mean value of the group carrying the favorable allele was less than that of the other group (**Figure 2**).

GO and KEGG Pathway Enrichment Analyses

According to the Nipponbare reference genome, the 371 identified QTNs for traits related to salt tolerance were part of or were adjacent to 581 genes (**Supplementary Table S1**). These genes were significantly enriched for GO biological processes related to the plant lipid metabolic process and transmembrane transport process (**Supplementary Table S3**). They were also significantly enriched for the plant tryptophan metabolism pathway ($P < 0.03$). Moreover, two genes (LOC_Os01g45760 and LOC_Os10g04860) were associated with auxin biosynthesis. A total of 66 genes were identified around the 56 QTNs based on the enriched GO terms and KEGG pathways as well as the functional annotations (**Supplementary Table S4**). This information may be very useful for identifying the genes responsible for salt tolerance in rice.

DISCUSSION

Multi-locus GWAS models, which are relatively close to the true genetic models of plants and animals, are superior to single-locus GWAS models because of their higher statistical power and lower FPR (Segura et al., 2012; Wang et al., 2016a). These models were developed by geneticists, who added the polygenic effect and population structure to the single-locus GWAS model to decrease the bias in effect estimations by controlling the genetic background (Zhang et al., 2005; Yu et al., 2006; Zhang et al., 2010). Although advancements in the single-locus GWAS models have improved the detection accuracy to some extent, the multiple test correction for the threshold value of the significance test in single-locus models (e.g., Bonferroni correction) is too stringent to capture all true QTNs. Another unavoidable problem is that single-locus GWAS methods are inappropriate when the target traits are controlled by a series of polygenes. In this study, 478 rice accessions with 162,529 SNPs were used to identify QTNs for traits related to salt tolerance based on six multi-locus GWAS methods. We compared the QTNs identified by the multi-locus GWAS methods in our study with the previously reported QTNs detected by the efficient mixed-model EMMA eXpedited (EMMAX) program comprising a single-locus GWAS method. The comparison revealed that four of the previously reported six QTNs related to SSI-VI were detected by a multi-locus GWAS, and two QTNs associated with SSI-MGT overlapped with the previously reported QTNs. Additionally, 12, 4, 13, 12, and 12 QTNs separately associated with SSI-GI, SSI-IR-24h, SSI-IR-48h, SSI-GR-5d, and SSI-GR-10d, respectively, were simultaneously detected by at least three multi-locus GWAS methods. In contrast, none of the QTNs associated with the five traits were identified by a single-locus GWAS method. These observations were as expected, and can be explained by the following two points: (i) salt tolerance is a quantitative genetic characteristic that is controlled by multiple genes with small effects, which are difficult to detect in a single-locus GWAS model (Wang et al., 2011; Kumar et al., 2015); (ii) some true QTNs for traits related to salt tolerance are missed by a single-locus GWAS model because of an overly conservative critical value. Furthermore, our results suggest that a multi-locus GWAS model may be useful for detecting loci with small effects.

In this study, we used six multi-locus GWAS methods included in the mrMLM package to detect QTNs. The six methods involve two-step algorithms, and marker effects are treated as random effects in each method. However, each method has its own characteristics. We observed that mrMLM detected the most QTNs (**Supplementary Table S1**), but this method has one shortcoming. When the number of putative QTNs is much larger than the sample size, the multi-locus model in this method will be over-fitted. The residual error estimated by mrMLM was

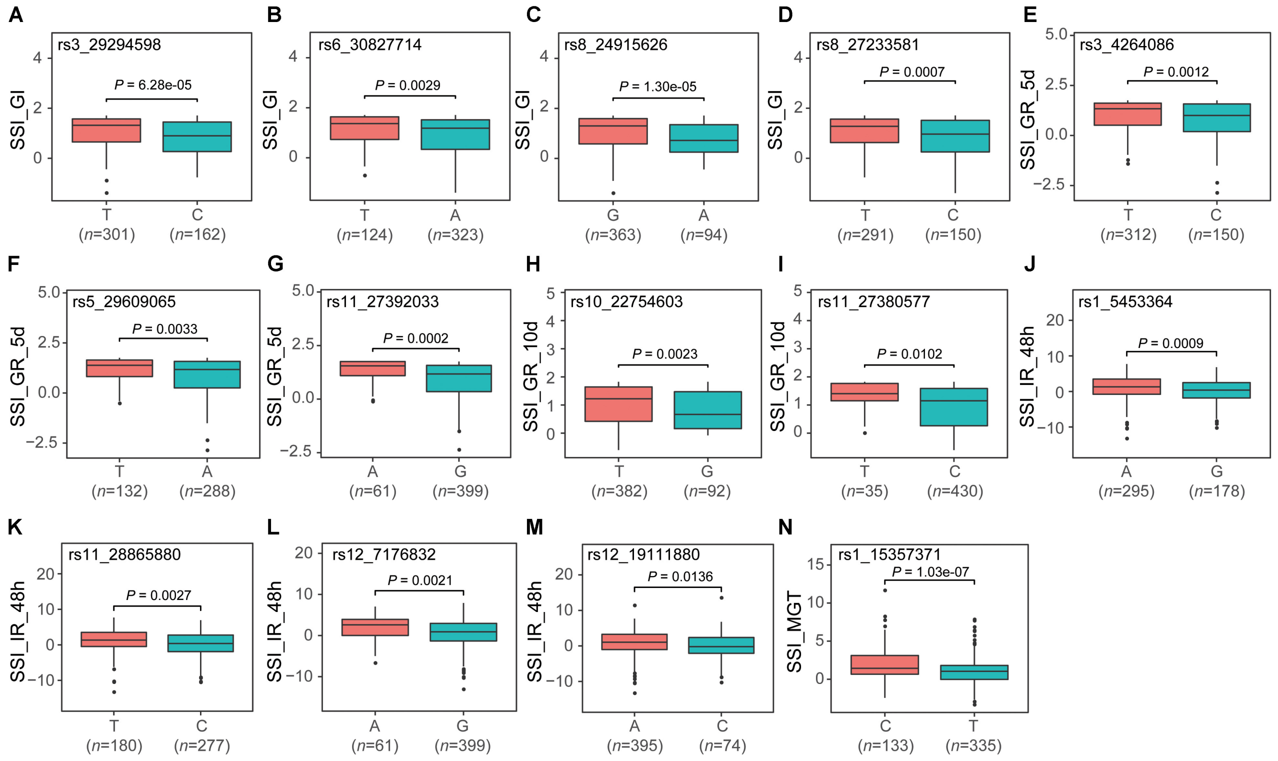


FIGURE 2 | Boxplot for validating 14 co-detected QTNs **(A–N)**. For each QTN, the population was divided into two groups according to allele types. The X-axis represents the two alleles for each QTN, while the Y-axis corresponds to the phenotype.

much smaller than that estimated by the five other methods (**Table 1**). During the first step, 7,588 QTNs with a threshold value $P < 0.01$ were selected, which is 16 times larger than the sample size. Over-fitting may occur when too many variables are added to a multi-locus model. This issue was solved by using FASTmrMLM, in which the least angle regression (LARS) algorithm is implemented between the first single-locus scanning step and the EM-Empirical Bayes estimation in the second step. The LARS algorithm (Efron et al., 2004) is a flexible method for selecting variables, and can be applied in the lars package[1]. In this method, $n-1$ variables (n is the number of samples), which are most likely associated with the target traits, are added to the multi-locus model.

The FASTmrEMMA method detected the fewest QTNs. This method involves an approximation algorithm in which the covariance matrix of the polygenic matrix K and environmental noise are whitened by a matrix transformation to increase the computing speed. In the pLARmEB method, the same transformed model as that used in FASTmrEMMA is implemented to control the polygenic background, and the LARS algorithm is applied to select potential SNPs related to the target trait for the subsequent multi-locus GWAS detection. Among the six multi-locus GWAS methods, ISIS-EM-BLASSO had the shortest running time and the smallest estimated residual errors (**Supplementary Figure S2** and **Table 1**). In the first step of this method, an iterative-modified sure independence screening (ISIS) approach is used to decrease the number of SNPs to a moderate level, after which the Expectation-Maximization (EM)-Bayesian least absolute shrinkage and selection operator (BLASSO) is used to estimate all of the selected SNP effects to detect true QTNs. The last method, pKWmEB, is a non-parametric method, in which a Kruskal–Wallis test and the LARS algorithm are used to identify potential SNPs. All identified markers are added to the multi-locus model to detect true QTNs.

The two-step multi-locus GWAS methods included in this study significantly improved the statistical power and decreased the FPR. Moreover, ISIS EM-BLASSO identified the most co-detected QTNs, followed by pKWmEB, while FASTmrEMMA identified the fewest QTNs (**Table 2**). Additionally, ISIS EM-BLASSO performed best, with the smallest estimated residual errors and highest computing speed. However, selecting an appropriate critical value is still problematic for the two-step multi-locus GWAS model. A threshold value that is too stringent will lead to the omission of loci information, whereas a relaxed threshold value will result in numerous loci being selected, which may lead to the over-fitting of multi-locus models. A simple solution to this problem involves developing a hybrid method that combines the results from different methods. Directly decreasing the number of SNPs instead of applying a single-locus GWAS scanning step represents another potential solution. We are currently developing a new bin analysis method that can be applied to any type of population. In the bin analysis method,

[1] http://cran.r-project.org/web/packages/lars/

the number of markers is decreased, but the information for all markers is fully retained. Adding a bin analysis to the multi-locus GWAS model represents a new option.

CONCLUSION

In this study, six multi-locus GWAS methods were used to detect loci related to rice salt tolerance at the seed germination stage. A total of 371 QTNs were identified, with 56 QTNs co-detected by at least three methods. Moreover, 66 genes were identified in the vicinity of the 56 QTNs based on functional annotations. Two of these genes (LOC_Os01g45760 and LOC_Os10g04860) are involved in auxin biosynthesis according to the enriched GO terms and KEGG pathways. These observations may be useful for identifying the genes responsible for rice salt tolerance.

AUTHOR CONTRIBUTIONS

YC drafted the manuscript. FZ and YC analyzed the data. YZ and FZ conceived the study and were in charge of the direction and planning. All authors read and approved the final version of this manuscript.

FUNDING

This research was supported by grants from the National Key Research and Development Program of China (Project No. 2016YFD0100101), the National Natural Science Foundation of China (Grant No. 31771762), the National High-tech Program of China (No. 2014AA10A603), the Shenzhen Peacock Plan (No. 20130415095710361), the Bill & Melinda Gates Foundation (OPP1130530), and the CAAS Agricultural Science and Technology Innovative Program. YC was supported by the Beachell-Borlaug International Student Fellowship from Monsanto.

ACKNOWLEDGMENTS

We thank Liwen Bianji, Edanz Editing China (www.liwenbianji.cn/ac) for editing the English text of a draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01464/full#supplementary-material

FIGURE S1 | Bar plot of the number of QTNs associated with seven salt tolerance traits detected by different methods.

FIGURE S2 | Computing time of the six multi-locus GWAS methods.

TABLE S1 | Significant salt tolerance QTNs detected by six multi-locus GWAS methods.

TABLE S2 | Significant QTNs detected by at least two multi-locus GWAS methods.

TABLE S3 | Results of a GO enrichment analysis.

TABLE S4 | Gene annotations for the 56 significant QTNs associated with salt tolerance traits.

REFERENCES

Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., Fuentes, R. R., et al. (2015). SNP-seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res.* 43(Database issue), D1023–D1027. doi: 10.1093/nar/gku1039

Ayers, K. L., and Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* 34, 879–891. doi: 10.1002/gepi.20543

Chen, W., Gao, Y. Q., Xie, W. B, Gong, L., Lu, K., Wang, W. S., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46, 714–721. doi: 10.1038/ng.3007

Chien, C., Chow, C., Wu, N., Chiang-Hsieh, Y., Hou, P., and Chang, W. (2015). EXPath: a database of comparative expression analysis inferring metabolic pathways for plants. *BMC Genomics* 16(Suppl. 2):S6. doi: 10.1186/1471-2164-16-S2-S6

Cho, S., Kim, H., Oh, S., Kim, K., and Park, T. (2009). Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proc.* 3(Suppl. 7):S25. doi: 10.1186/1753-6561-3-s7-s25

Cho, S., Kim, K., Kim, Y. J., Lee, J. K., Cho, Y. S., Lee, J. Y., et al. (2010). Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Annu. Hum. Genet.* 74, 416–428. doi: 10.1111/j.1469-1809.2010.00597

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Cui, Y. R., Zhang, F., Xu, J. L., Li, Z. K., and Xu, S. Z. (2015). Mapping quantitative trait loci in selected breeding populations: a segregation distortion approach. *Heredity* 115, 538–546. doi: 10.1038/hdy.2015.56

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.* 32, 407–451. doi: 10.1214/009053604000000067

Giglio, C., and Brown, S. D. (2018). Using elastic net regression to perform spectrally relevant variable selection. *J. Chemom.* 32:*e*3034. doi: 10.1002/cem.3034

Han, B., and Huang, X. H. (2013). Sequencing-based genome-wide association study in rice. *Curr. Opin. Plant Biol.* 16, 133–138. doi: 10.1016/j.pbi.2013.03.006

Hu, S. K., Tao, H. J., Qian, Q., and Guo, L. B. (2012). Genetics and molecular breeding for salttolerance in rice. *Rice Genom. Genet.* 3, 39–49.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., et al. (2013). Improvement of the *Oryza sativa* nipponbare reference genome using next generation sequence and optical map data. *Rice* 6:4. doi: 10.1186/1939-8433-6-4

Kumar, V., Singh, A., Mithra, S., Krishnamurthy, S., Parida, S., and Jain, S. (2015). Genome-wide association mapping of salinity tolerance in rice (*Oryza sativa*). *DNA Res.* 22, 133–145. doi: 10.1093/dnares/dsu046

Lee, K., Choi, W., Ko, J., Kim, T., and Gregorio, G. (2003). Salinity tolerance of japonica and indica rice (*Oryza sativa* L.) at the seedling stage. *Planta* 216, 1043–1046. doi: 10.1007/s00425-002-0958-3

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2018). pKWmEB: integration of Kruskal–Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Segura, V., Vilhjalmsson, B. J., Platt, A., Korte, A., Seren, U., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

Shi, Y. R., Gao, L. L., Wu, Z. C., Zhang, X. J., Wang, M. M., Zhang, C. S., et al. (2017). Genome-wide association study of salt tolerance at the seed germination stage in rice. *BMC Plant Biol.* 17:92. doi: 10.1186/s12870-017-1044-0

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017). agriGO v2.0*: a GO analysis toolkit for the agricultural community. Nucleic Acids Res.* 45, W122–W129. doi: 10.1093/nar/gkx382

Wang, D., Eskridge, K. M., and Crossa, J. (2010). Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. *J. Agric. Biol. Environ. Stat.* 16, 170–184. doi: 10.1007/s13253-010-0046-2

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., and Zhang, Z. (2014). A super powerful method for genome wide association study. *PLoS One* 9:e107684. doi: 10.1371/journal.pone.0107684

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L.,Wen, Y. J., et al. (2016a). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi.org/10.1038/srep19444 doi: 10.1038/srep19444

Wang, S. B., Wen, Y. J., Ren, W. L., Ni, Y. L., Zhang, J., Feng, J. Y., et al. (2016b). Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via a multi-locus GWAS methodology. *Sci. Rep.* 6:29951. doi: 10.1038/srep29951

Wang, Z., Wang, J., Bao, Y., Wu, Y., and Zhang, H. (2011). Quantitative trait loci controlling rice seed germination under salt stress. *Euphytica* 178, 297–307. doi: 10.1007/s10681-010-0287-8

Wei, J. L., Wang, A. G., Li, R. D., Qu, H., and Jia, Z. Y. (2017). Metabolome-wide association studies for agronomic traits of rice. *Heredity* 120, 342–355. doi: 10.1038/s41437-017-0032-3

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018a). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Wen, Y. J., Zhang, Y. W., Zhang, J., Feng, J. Y., Dunwell, J. M., and Zhang, Y. M. (2018b). An efficient multi-locus mixed model framework for the detection of small and linked QTLs in F2. *Brief. Bioinform.* bby058. doi: 10.1093/bib/bby058

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25, 714–721. doi: 10.1093/bioinformatics/btp041

Xu, S. (2013). Mapping quantitative trait loci by controlling polygenic background effects. *Genetics* 195, 1209–1222. doi: 10.1534/genetics.113.157032

Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., et al. (2014). Combining highthroughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.* 5:5087. doi: 10.1038/ncomms6087

Yi, N., and Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* 179, 1045–1055. doi: 10.1534/genetics.107.085589

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., and Doebley, J. F. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y. M., Mao, Y., Xie, C., Smith, H., Luo, L., and Xu, S. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize. *Genetics* 169, 2267–2275. doi: 10.1534/genetics.104.033217

Zhang, Y. M., and Tamba, C. L. (2018). A fast mrMLM algorithm for multi-locus genome-wide association studies. *biorxiv* [Preprint]. doi: 10.1101/341784

Zhang, Z., Ersoz, E., Lai, C. -Q., Todhunter, R. J., Tiwari, H. K., and Gore, M. A. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Plant Science

ORIGINAL RESEARCH
published: 20 November 2018
doi: 10.3389/fpls.2018.01683

Multi-Locus Genome-Wide Association Studies for 14 Main Agronomic Traits in Barley

Xin Hu[1,2†], *Jianfang Zuo*[1†], *Jibin Wang*[1], *Lipan Liu*[1], *Genlou Sun*[3], *Chengdao Li*[4,5], *Xifeng Ren*[1*] and *Dongfa Sun*[1,5*]

[1] *College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China,* [2] *Guiyang College of Traditional Chinese Medicine, Guiyang, China,* [3] *Biology Department, Saint Mary's University, Halifax, NS, Canada,* [4] *School of Veterinary and Life Sciences, Murdoch University, Murdoch, WA, Australia,* [5] *Hubei Collaborative Innovation Center for Grain Industry, Jingzhou, China*

The agronomic traits, including morphological and yield component traits, are important in barley breeding programs. In order to reveal the genetic foundation of agronomic traits of interest, in this study 122 doubled haploid lines from a cross between cultivars "Huaai 11" (six-rowed and dwarf) and "Huadamai 6" (two-rowed) were genotyped by 9680 SNPs and phenotyped 14 agronomic traits in 3 years, and the two datasets were used to conduct multi-locus genome-wide association studies. As a result, 913 quantitative trait nucleotides (QTNs) were identified by five multi-locus GWAS methods to be associated with the above 14 traits and their best linear unbiased predictions. Among these QTNs and their adjacent genes, 39 QTNs (or QTN clusters) were repeatedly detected in various environments and methods, and 10 candidate genes were identified from gene annotation. Nineteen QTNs and two genes (*sdw1/denso* and *Vrs1*) were previously reported, and eight candidate genes need to be further validated. The *Vrs1* gene, controlling the number of rows in the spike, was found to be associated with spikelet number of main spike, spikelet number per plant, grain number per plant, grain number per spike, and 1,000 grain weight in multiple environments and by multi-locus GWAS methods. Therefore, the above results evidenced the feasibility and reliability of genome-wide association studies in doubled haploid population, and the QTNs and their candidate genes detected in this study are useful for marker-assisted selection breeding, gene cloning, and functional identification in barley.

OPEN ACCESS

Edited by:
Zhenyu Jia,
University of California, Riverside,
United States

Reviewed by:
Yin-Gang Hu,
Northwest A&F University, China
Yang Xu,
Yangzhou University, China

***Correspondence:**
Dongfa Sun
sundongfa1@mail.hzau.edu.cn
Xifeng Ren
renxifeng@mail.hzau.edu.cn

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: 30 July 2018
Accepted: 29 October 2018
Published: 20 November 2018

Citation:
Hu X, Zuo J, Wang J, Liu L, Sun G, Li C, Ren X and Sun D (2018) Multi-Locus Genome-Wide Association Studies for 14 Main Agronomic Traits in Barley. Front. Plant Sci. 9:1683. doi: 10.3389/fpls.2018.01683

Keywords: genome-wide association study, barley, multi-locus model, doubled haploid population, quantitative trait nucleotide, candidate gene

INTRODUCTION

Barley (*Hordeum vulgare* L, $2n = 2x = 14$), one of the first domesticated grains in the Fertile Crescent (Zohary et al., 2012), has been used widely as animal feed, human health foods, and a source of beer. Its yield and quality are the most important breeding objectives in crop breeding programs.

Most agronomic traits, such as plant height- and yield-related traits, are controlled by quantitative trait loci (QTLs) in barley, so it is difficult to obtain their genetic foundation and molecular mechanism. Plant height and its component traits serve as major plant morphological traits affecting barley seed yield. An appropriate plant height is a prerequisite for obtaining the desired yield in barley-breeding programs. To date, more than 30 types of dwarfing or semi-dwarfing genes have been detected, while only a few have been successfully used in barley breeding program, such as *uzu* and *sdw1/denso* (Jia et al., 2009; Ren et al., 2010, 2013). Moreover, a large number of QTLs for plant height related traits were reported to be located on all the seven chromosomes (Sameri et al., 2006; Baghizadeh et al., 2007; Wang et al., 2010; Ren et al., 2014). Grain yield is the key trait for the breeder in barley breeding program, therefore, the yield related traits including spike number per plant (SP), grain number per plant (GP), grain weight per plant (GWP), and 1,000 grain weight (TGW), have gained more attentions in the genetic dissection of yield related traits. A huge number of QTLs for yield related traits were detected to be located across all the chromosomes (Li et al., 2005; Sameri et al., 2006; Baghizadeh et al., 2007; Wang et al., 2010, 2016a; Ren et al., 2013).

Traditionally, QTL mapping has been widely applied in the genetic dissection of quantitative traits in barley (Zhuang et al., 1997; Li et al., 2005, 2007; Peng et al., 2011). As the development of DNA sequencing technologies, it is relatively easy to obtain high-density SNP genotypes for association mapping (AM) population, which offers a huge convenience for genomic and genetic research in different species. Therefore, genome-wide association studies (GWAS) present a powerful tool to reconnect the complex quantitative traits with their genes. Due to the development of cheaper, faster and higher-throughput molecular markers, AM has been widely used for mapping QTLs and genes in many crops, such as maize, soybean, rice, barley and wheat (Huang et al., 2010; Yang et al., 2010; Pasam et al., 2012; Hu et al., 2015). In comparison with traditional QTL mapping, AM has three obvious advantages, including shorter construction time, much higher mapping resolution and a greater number of alleles (Zhang et al., 2005; Yu and Buckler, 2006). In barley, AM has been widely applied for complex traits including disease resistance (Massman et al., 2011), drought tolerance (Varshney et al., 2012; Wójcik-Jagła et al., 2018), salinity tolerance (Fan et al., 2016) and especially agronomic traits (Gawenda et al., 2015; Xu et al., 2018).

Beside for natural population, nowadays GWAS have been widely applied to the genetic analysis for complex traits in family-based populations, such as nested association mapping (NAM) and multi-parent advanced generation intercross (MAGIC) populations, and proved to be powerful tool for uncovering the basis of key agronomic traits in maize and barley (Tian et al., 2011; Cook et al., 2012; Maurer et al., 2015, 2016). Moreover, there are also successful cases that combine linkage analysis with GWAS in several bi-parental segregation populations, such as recombinant inbred line (RIL) population (Lu et al., 2010; Reif et al., 2010). For single segregating population, successful but fewer cases were performed using GWAS (Gao et al., 2015; Henning et al., 2016; Liu et al., 2018). Henning et al. (2016) conducted GWAS for downy mildew resistance in a segregating population of "Teamaker" × USDA 21422M in hop (*Humulus lupulus* L.), Gao et al. (2015) determined the location of TTKSK resistance in the 108 doubled haploid (DH) lines using a GWAS method implemented by R package rrBLUP (Endelman, 2011), and Liu et al. (2018) used two strategies (QTL mapping and GWAS) to reveal the genetic bases of fiber quality traits and yield components in 231 RILs. Therefore, it is feasible to use GWAS to dissect the genetic foundations of complex traits in single bi-parental segregating population. A combination of linkage and association methodologies should provide the more accurate and powerful approach for revealing the genetic bases of complex traits (Ott et al., 2011). Association mapping of drought tolerance-related traits was performed in barley to complement a traditional bi-parental QTL mapping study in Wójcik-Jagła et al. (2018).

The objectives of this study were to: (a) use GWAS to further dissect the genetic foundations for main agronomic traits in our previous studies of Ren et al. (2013, 2014) and Wang et al. (2016a), and compare the quantitative trait nucleotide (QTN) results with those in previous studies, (b) evaluate if GWAS is feasible and reliable in the genetic dissection of complex traits in DH population, and (c) mine the candidate genes in the regions of the QTNs. The outcome of this study will provide more precise and complete information for further gene cloning, and marker-assisted selection in barley breeding.

MATERIALS AND METHODS

Plant Materials and Field Experiments

One hundred and twenty-two DH lines, derived from a cross between barley cultivar "Huaai 11" (six-rowed and dwarfing) and barley cultivar "Huadamai 6" (two-rowed), was used in this study. The details of the materials and field experiments were described in the previous studies of Ren et al. (2010, 2013, 2014) and Wang et al. (2016a).

Phenotyping Data

Fourteen agronomic traits for the above DH population were measured in 2008–2009, 2009–2010, and 2011–2012, and all the three datasets had been reported by Ren et al. (2013, 2014) and Wang et al. (2016a). These traits included plant height (PH), first internode length (IL1), second internode length (IL2), third internode length (IL3), fourth internode length (IL4), main spike length (MSL), spike number per plant (SP), spikelet number of main spike (SMS), spikelet number per plant (SLP), grain number per plant (GP), grain number per spike (GS), grain weight per plant (GWP), grain weight per spike (GWS), and 1,000 grain weight (TGW). All the details have been described in Ren et al. (2013, 2014).

The best linear unbiased predictions (BLUPs) for each trait of 3 years were calculated using the R package Lme4 (Bates et al., 2014) with the following model: $y = \text{lmer}\ (\text{Trait} \sim (1|\text{Genetype}) + (1|\text{Year}))$. The three single-year phenotypic values (Ren et al., 2013, 2014; Wang et al., 2016a) and their BLUP values were used for GWAS. The results of phenotype statistics of BLUP for each trait were summarized in **Table S1**.

Genotyping Data

All the above DH lines were genotyped by 10,367 polymorphic SNPs. After excluding low quality SNP markers, 9680 SNPs were used in this study. Base on the recent genome sequence release in barley (Ibsc, 2016; Beier et al., 2017; Mascher et al., 2017), all the SNP markers were aligned to the most reliable genome position (http://webblast.ipk-gatersleben.de/barley_ibsc/). The *Vrs1* locus controlling row number of barley was integrated with SNP markers for GWAS. All the above information has been described in Ren et al. (2016).

GWAS

Q matrix was calculated by STRUCTURE software (Falush et al., 2003), and the optimal K was inferred in **Figure S1**. The kinship (K) matrix between the lines was calculated as previously described in Wang et al. (2016b). All the 9680 SNPs for the above 122 DH lines were used to conduct GWAS for the above 14 traits in 3 years and their BLUP values using five multi-locus GWAS methods, including mrMLM (Wang et al., 2016b), FASTmrMLM (Zhang and Tamba, 2018), FASTmrEMMA (Wen et al., 2018), pLARmEB (Zhang et al., 2017) and ISIS EM-BLASSO (Tamba et al., 2017), which were included in the R package mrMLM v3.1 (https://cran.r-project.org/web/packages/mrMLM/index.html). All parameters in GWAS were set at default values. The critical thresholds of significant association for the five methods were set as LOD = 3 (or P-value $= 2 \times 10^{-4}$; Wang et al., 2016b).

The significant QTNs, repeatedly detected in at least two environments or methods, were viewed as reliable. The associated regions on chromosomes, repeatedly located on same or similar traits in at least 2 years or methods, were viewed as reliable QTN clusters. These QTNs (or clusters) were named as "qtn (qtnc)" + trait name abbreviation + chromosome + detected QTL order on chromosome.

Phenotypes Difference Corresponding to QTNs

For each QTN, all the DH lines were firstly divided into two groups based on their QTN genotypes, then t-test was used to test the phenotypic difference between the two genotypes.

Identification of Candidate Genes

According to the recent genome sequence release of barley (Ibsc, 2016; Beier et al., 2017; Mascher et al., 2017) and the gene annotation information (http://plants.ensembl.org/Hordeum_vulgare/Info/Index and https://www.uniprot.org/uniprot), some genes around reliable QTNs (or clusters) were selected for each trait. By combining gene annotation information, protein domain function in database and previous reports, and expressional information (http://barlex.barleysequence.org), then, candidate genes for each trait were mined.

RESULTS

GWAS for 14 Agronomic Traits

Using five multi-locus GWAS methods in the R package mrMLM v3.1, GWAS for 14 agronomic traits were performed. A total of 913 significant QTNs were found to be associated with the 14 agronomic traits in 3 years and their BLUP values (**Table S2**). The number of significant QTNs varied across various traits, ranging from 4 for 2012_IL4 to 37 for BLUP_PH (**Figure 1**; **Table S2**), the chromosomal distribution of all identified QTNs revealed that 2H had the maximum number of significant QTNs, which weren't evenly distributed on the genome, and five QTN hotspots on chromosomes 2H, 3H, 6H, and 7H were observed (**Figures 1**, **2**).

The significant QTNs repeatedly detected by multiple methods were list in **Table S3**. The reliable QTNs (or clusters) for 14 agronomic traits were summarized in **Table 1**. Totally, there were 39 reliable QTNs (or clusters) for 14 agronomic traits [8 for PH, 4 for IL1, 1 for IL2, 2 for IL3, 3 for IL4, 4 for MSL, 2 for SP, 2 for SMS, 3 for SLP, 3 for TGW, 2 for GP, 3 for GS, 1 for GWS, and 1 QTN cluster for (GP-GWP-GWS)].

The QTNs for Plant Height and Its Components Traits

Six reliable QTNs and two reliable QTN clusters, distributed on four chromosomes, were significantly associated with PH (**Table 1**; **Figure 2**). Six reliable QTNs for PH, *qtnPH-1H-1* (1H: 554,371,992 bp), *qtnPH-2H-1* (2H: 540,094,243 bp), *qtnPH-3H-2* (3H: 651,696,476 bp), *qtnPH-7H-1* (7H: 81,959,684 bp), *qtnPH-7H-2* (7H:108,670,637 bp), and *qtnPH-7H-3* (7H: 622,802,079 bp), located on chromosome 1H, 2H, 3H, 7H, 7H, and 7H, explained 0.63–4.07, 2.82–4.64, 3.13–9.03, 1.07–2.39, 2.31–6.07, and 0.69–2.68% of total phenotypic variation, respectively. Two reliable QTN clusters for PH, *qtncPH-2H-2* (2H: 560,195,592–564,116,957 bp) and *qtncPH-3H-1* (3H: 631,870,705–633,068,955 bp), were identified in at least 2 years and methods, explained 2.12–3.79 and 2.15–7.23% of total phenotypic variation, respectively (**Tables 1**, **S3**).

4, 1, 2, and 3 reliable QTNs (or clusters) were found to be associated with plant height component traits IL1, IL2, IL3, and IL4, respectively. For IL1, *qtncIL1-7H-4*, close to *qtnPH-7H-1* (7H: 81,959,684 bp) for PH, was located on 7H (81,889,341–84,350,472 bp) accounting for a highest phenotypic variation (3.53–22.62%) among the four QTNs and QTN clusters (**Table 1**). For IL2, *qtnIL2-7H-6* on chromosome 7H: 258,071,311 bp, explained phenotypic variation of 25.82–55.77%. For IL3, *qtncIL3-3H-3* was located at 3H: 631,342,028–636,535,362 bp, close to the region of *qtncPH-3H-1* (3H: 631,870,705–633,068,955 bp) for PH, accounting for 4.88–21.70% of the phenotypic variation. *qtncIL3-6H-1* located on the chromosome 6H (16,165,407–17,542,081 bp), explained phenotypic variation of 1.89–3.68%. For IL4, *qtncIL4-3H-4* (3H: 631,870,705–636,535,362 bp), co-located in the same region of *qtncIL3-3H-3* for IL3 and *qtncPH-3H-1* for PH mentioned above, explained 5.73–15.97% of phenotypic variation. Therefore, the region on chromosome 3H: 631,342,028–636,535,362 bp is a more credible QTN cluster for PH and the components traits, regulating the PH through controlling IL3 and IL4 (**Table 1**). Another QTN cluster for IL4, *qtncIL4-2H-4* on 2H (4,629,895–4,950,022 bp) explained less (1.79–6.02%) phenotypic variation, while QTN *qtnIL4-7H-7* (7H: 360793216 bp) showed a high explanation (6.10–19.57%) for IL4 (**Tables 1**, **S3**; **Figure 2**).

The QTNs for Spike and Yield Related Traits

Main spike length (MSL): Three reliable QTNs and one QTN cluster were detected for MSL. Among which, *qtnMSL-2H-7*

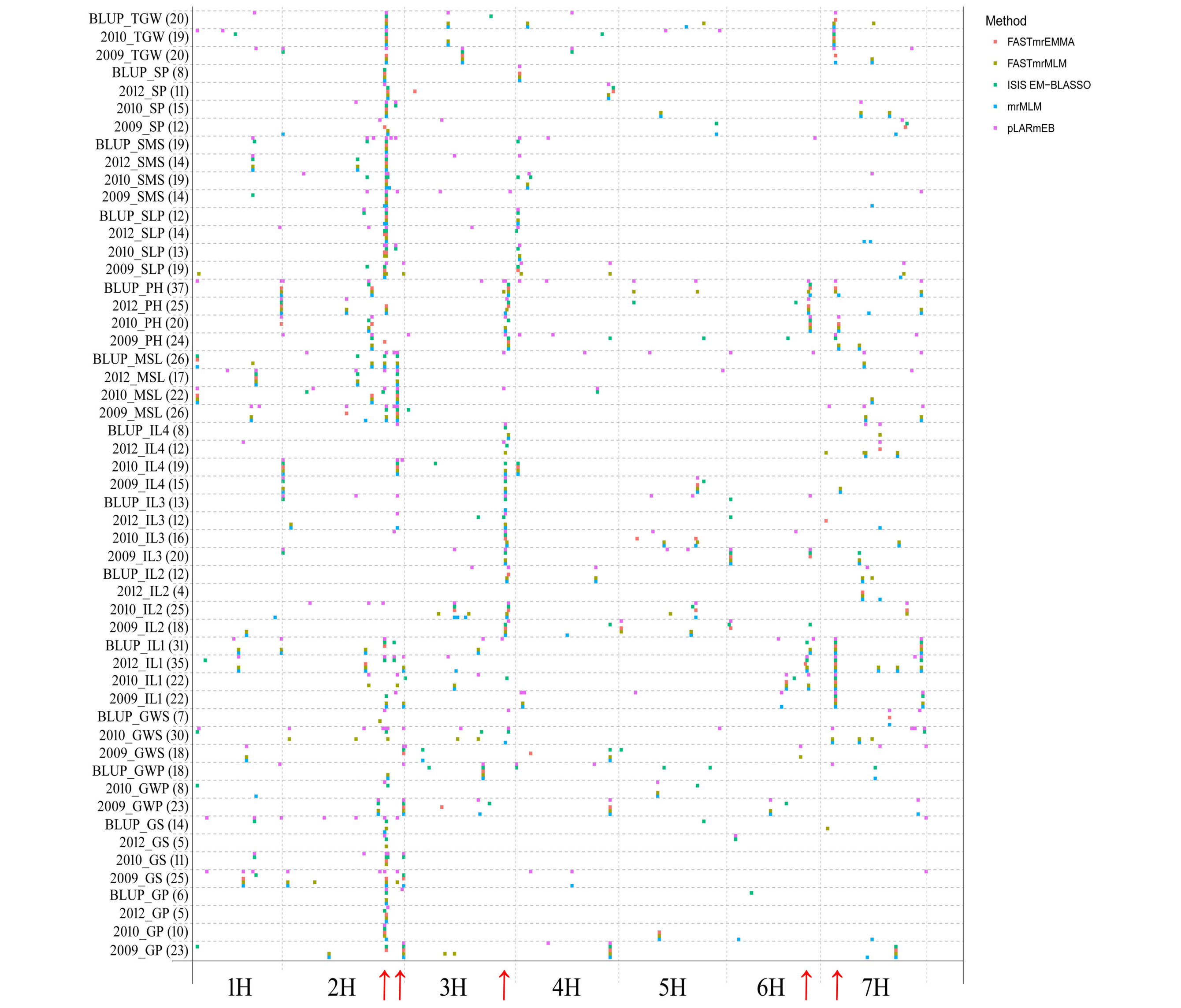


FIGURE 1 | Chromosomal distribution of QTNs identified in this study. The *x*-axis indicates genomic locations by chromosomal order, and the significant QTNs are plotted against genome location. Each row represents one QTN identified by a different method. The red arrows show the QTN hotspots.

located on chromosome 2H: 727,985,438 bp, was repeatedly detected not only in three environments and BLUP values but also by multiple methods to be significantly associated with MSL, and explained phenotypic variation about 2.09–18.16% (LOD score: 3.29–22.92; **Tables 1**, **S3**; **Figure 2**).

Spikelet number of main spike (SMS): *qtnSMS-2H-9* referred to *Vrs1* (the morphological markers for row number of barley), located at 2H (position: 652,030,802 bp), was identified in all the situations and methods to be significantly associated with SMS (**Tables 1**, **S3**), accounting for the largest phenotypic variation (65.07–90.41%). As already known, *Vrs1*, a gene controlling row number of barley, was validated controlling row number in the DH population derived from a cross between the six-rowed barley cultivar "Huaai 11" and the two-rowed barley cultivar "Huadamai 6." Thus, *Vrs1* should control the spike related traits, such as SMS and SLP. Moreover, *qtnSMS-2H-8* (2H: 535,680,815 bp) with minor effect was associated with SMS (**Table 1**).

Spikelet number per plant (SLP): One reliable QTN *qtnSLP-2H-10* (2H: 649,558,019 bp) and one reliable QTN clusters *qtncSLP-2H-11* (2H: 652,030,802–653,982,961 bp) close to *Vrs1* (2H: 652,030,802 bp) were identified in multiple environments and by multiple methods to be significantly associated with SLP, explaining high proportions of total phenotypic variation, 14.87–55.05 and 8.97–80.30%, respectively. The reliable QTN cluster *qtncSLP-4H-1* with a minor effect (2.82–4.14%), mapped on the region 15,498,372–16,168,735 bp of chromosome 4H, was detected in 2 years and the BLUP (**Tables 1**, **S3**; **Figure 2**).

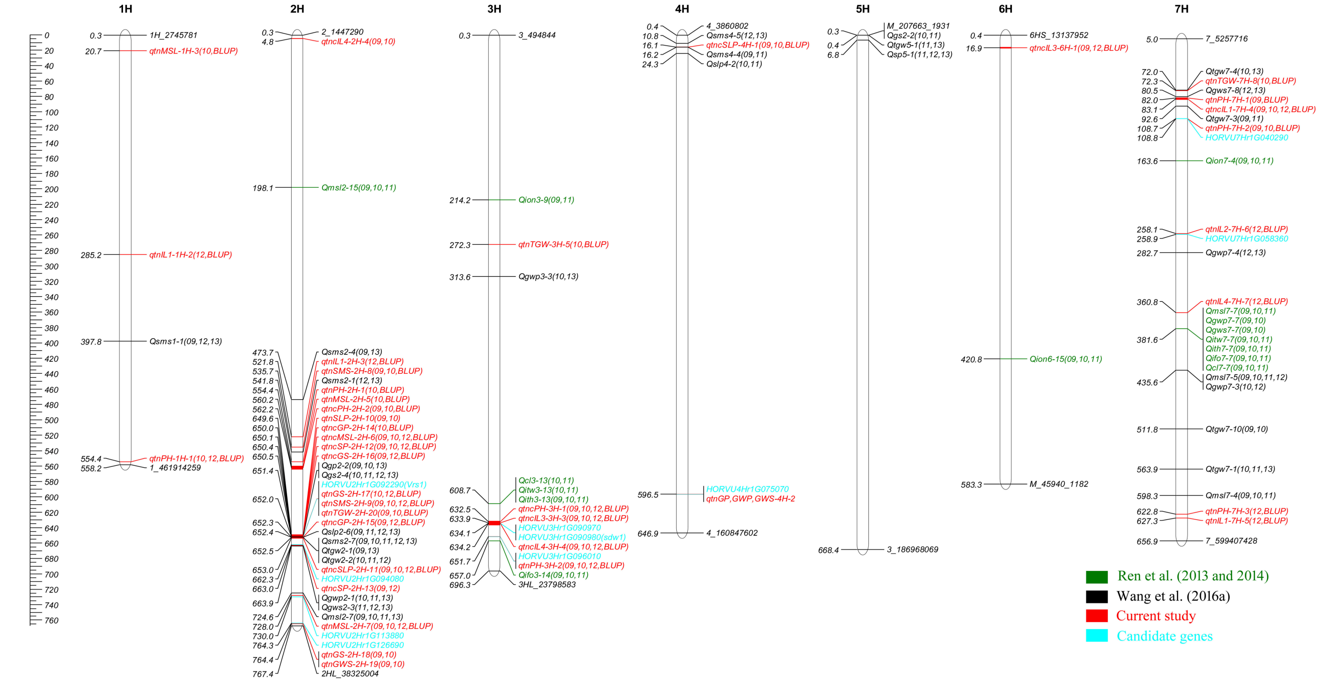


FIGURE 2 | Chromosomes location of reliable QTLs for 14 agronomic traits in both previous studies (Ren et al., 2013, 2014; Wang et al., 2016a) and the current studies. The peak positions of previous QTLs were used for mapping, Genetic distance scale in physic position (Mb) is placed at left margin. Green is for the QTLs detected in Ren et al. (2013, 2014), black is for the QTLs Wang et al. (2016a), red is for the QTNs and QTNs clusters of current study, the region of QTNs clusters was marked with red on the bar, the cyan is for the candidate genes.

Spike number per plant (SP): Two reliable QTN clusters for SLP were detected. *qtncSP-2H-12*, located on 2H: 648,821,931–652,030,802 bp overlapping with *Vrs1*, was detected in three environments and BLUP value and by multiple methods to be associated with SP, accounting for 7.61–52.32% of the phenotypic variation. *qtncSP-2H-13*, located on chromosome 2H (Position: 662,335,248–663,628,734 bp), was detected in two environments and by multiple methods to be associated with SP, explaining 6.49–11.95% of the phenotypic variation (**Tables 1**, **S3**; **Figure 2**).

Grain number per plant (GP) and grain number per spike (GS): The QTN cluster *qtncGP-2H-14* with a high proportion of total phenotypic variation (19.56–37.44%), located at 2H: 649,558,019–650,438,830 bp, was detected in 2010 and BLUP value to be associated with GP. The *qtncGP-2H-15* with a high explanation (3.41–34.59%), close to *Vrs1* (2H: 652,030,802–652,604,015 bp), was detected in three environments and by multiple methods to be associated with GP (**Table 1**). One small-effect QTN on 4H (Position: 596,447,744 bp), was detected in 2009 to be significantly associated with GP. Meanwhile, this QTN was also detected for GWP and GWS (**Table 1**). One QTN cluster and two QTNs were detected for GS in at least two environments and by multiple methods. The *qtncGS-2H-16* (2H: 649,558,019–651,399,477 bp), close to *Vrs1*, showed a high proportion of phenotypic variation (9.77–71.5%) for GS. The *qtnGS-2H-17* referred to *Vrs1* (2H: 652,030,802 bp) was significantly associated with GS, explaining a high percentage (49.99–72.25%) of the phenotypic variation. The reliable QTN *qtnGS-2H-18*, located at chromosome 2H (Position: 764,361,924 bp), was significantly associated with GS in 2009 and 2010 and multiple methods, accounting for 1.06–10.05% of the phenotypic variation (**Tables 1**, **S3**; **Figure 2**).

Grain weight per plant (GWP) and grain weight per spike (GWS): One QTN on 4H (Position: 596,447,744 bp) was found to be associated with GP, GWP and GWS with 3.97–9.17% proportions for the phenotypic variation. The QTN *qtnGWS-2H-19*, derived from the same associated SNP *2_625783669* (2H: 764,361,924 bp) as GS, was detected for GWS in 2009 and 2010 and by multiple GWAS methods with 0.74–6.25% proportion of phenotypic variation (**Tables 1**, **S3**; **Figure 2**).

1,000 grain weight (TGW): three reliable QTNs were detected for TGW. The *qtnTGW-2H-20*, located at *Vrs1* (2H: 652,030,802 bp), was significantly associated with TGW in multiple environments and GWAS methods, explaining high percentage (32.77–56.66%) of the phenotypic variation. The *qtnTGW-3H-5* (3H: 272,283,784 bp) and *qtnTGW-7H-8* (7H: 72,344,563 bp) were significantly associated with TGW, explaining 2.19–4.39 and 4.39–11.22% of the phenotypic variation, respectively (**Tables 1**, **S3**; **Figure 2**).

Phenotypic Difference Corresponding to QTNs

According to the QTN genotypes, all the DH lines were divided into two different groups to test whether the significant difference

TABLE 1 | Reliable QTNs and QTN clusters for 14 agronomic traits using multi-locus GWAS methods.

QTN (QTN cluster)[a]	Trait	Marker associated	Physic position (bp)	LOD score	r^2 (%)	Year	Previous QTL
qtnPH-1H-1	PH	1_463006138	1H: 554371992	3.39–7.64	0.63–4.07	2010, 2012, BLUP	
qtnPH-2H-1	PH	M_1999039_472	2H: 540094243	5.54–16.21	2.82–4.64	2010, BLUP	
qtncPH-2H-2	PH	2_447773331–M_1663886_573	2H: 560195592–564116957	3.51–12.47	2.12–3.79	2009, 2010, BLUP	
qtncPH-3H-1	PH	3HL_37004393–3_511749149	3H: 631870705–633068955	3.38–15.96	2.15–7.23	2009, 2010, 2012, BLUP	Qcl3-13, Qitw3-13, Qith3-13 (Ren et al., 2014)
qtnPH-3H-2	PH	2HS_32409186	3H: 651696476	4.18–16.72	3.13–9.03	2009, 2010, 2012, BLUP	Qifo3-14 (Ren et al., 2014)
qtnPH-7H-1	PH	7HS_12212266	7H: 81959684	3.33–9.13	1.07–2.39	2009, BLUP	qIN3-7HS, qIN4-7HS (Sameri et al., 2009)
qtnPH-7H-2	PH	7HS_33062962	7H: 108670637	3.72–6.60	2.31–6.07	2009, 2010, BLUP	qIN3-7HS, qIN4-7HS (Sameri et al., 2009)
qtnPH-7H-3	PH	M_249593_1037	7H: 622802079	3.38–10.17	0.69–2.68	2012, BLUP	
qtnIL1-1H-2	IL1	M_173442_1610	1H: 285199675	4.31–5.02	2.20–4.69	2012, BLUP	
qtnIL1-2H-3	IL1	2_399823529	2H: 521774247	3.23–5.44	2.26–5.96	2012, BLUP	
qtncIL1-7H-4	IL1	7HS_21829337–7_95992736	7H: 81889341–84350472	3.21–14.46	3.53–22.62	2009, 2010, 2012, BLUP	qIN3-7HS, qIN4-7HS (Sameri et al., 2009)
qtnIL1-7H-5	IL1	7_575487388	7H: 627311039	3.11–7.81	1.60–6.27	2012, BLUP	
qtnIL2-7H-6	IL2	M_114215_455	7H: 258071311	3.13–7.27	25.82–55.77	2012, BLUP	
qtncIL3-3H-3	IL3	3HL_37004393–3_511668322	3H: 631870705–636535362	3.01–21.75	4.88–21.70	2009, 2010, 2012, BLUP	Qith3-13 (Ren et al., 2014)
qtncIL3-6H-1	IL3	6_14536026–6_18118681	6H: 16165407–17542081	3.15–6.71	1.89–3.68	2009, 2012, BLUP	
qtncIL4-2H-4	IL4	M_1778358_754–2_4900503	2H: 4629895–4950022	3.58–5.09	1.79–6.02	2009, 2010	
qtncIL4-3H-4	IL4	3HL_37004393–3_511668322	3H: 631870705–636535362	5.69–14.07	5.73–15.97	2009, 2010, 2012, BLUP	Qifo3-14 (Ren et al., 2014)
qtnIL4-7H-7	IL4	7HL_11281033	7H: 360793216	3.00–3.46	6.10–19.57	2012, BLUP	Qifo7-7 (Ren et al., 2014)
qtnMSL-1H-3	MSL	1_35132055	1H: 20685614	3.21–6.62	1.49–3.46	2010, BLUP	
qtnMSL-2H-5	MSL	2_447773331	2H: 560195592	3.99–8.37	2.92–6.35	2010, BLUP	
qtncMSL-2H-6	MSL	2_522610509–2HL_34260490	2H: 648821931–651436685	3.61–10.16	1.57–4.65	2009, 2010, 2012, BLUP	
qtnMSL-2H-7	MSL	2_600749073	2H: 727985438	3.29–22.92	2.09–18.16	2009, 2010, 2012, BLUP	Qmsl2-7 (Wang et al., 2016a)
qtnSMS-2H-8	SMS	2_406934594	2H: 535680815	3.62–12.96	3.14–5.39	2009, 2010, BLUP	Qsms2-1 (Wang et al., 2016a)
qtnSMS-2H-9	SMS	Vrs1	2H: 652030802	6.24–82.16	65.07–90.41	2009, 2010, 2012, BLUP	Qsms2-7 (Wang et al., 2016a)
qtnSLP-2H-10	SLP	2_524762464	2H: 649558019	4.51–21.01	14.87–55.05	2009, 2010	Qslp2-6 (Wang et al., 2016a)
qtncSLP-2H-11	SLP	Vrs1–2HL_17075593	2H: 652030802–653982961	3.32–51.41	8.97–80.30	2009, 2010, 2012, BLUP	Qslp2-6 (Wang et al., 2016a)
qtncSLP-4H-1	SLP	4_16553551–M_1605646_794	4H: 15498372–16761959	3.26–9.76	2.82–4.14	2009, 2010, BLUP	Qslp4-2 (Wang et al., 2016a)
qtncSP-2H-12	SP	2_522610509–Vrs1	2H: 648821931–652030802	4.02–30.41	7.61–52.32	2009, 2010, 2012, BLUP	
qtncSP-2H-13	SP	2_531255437–M_124056_833	2H: 662335248–663628734	3.73–7.32	6.49–11.95	2009, 2012	
qtncGP-2H-14	GP	2_524762464–M_1589358_1352	2H: 649558019–650438830	3.46–17.05	19.56–37.44	2010, BLUP	Qgp2-2 (Wang et al., 2016a)
qtncGP-2H-15	GP	Vrs1–2_527241334	2H: 652030802–652604015	3.42–16.71	3.41–34.59	2009, 2012, BLUP	Qgp2-2 (Wang et al., 2016a)
qtncGS-2H-16	GS	2_524762464–2_527636020	2H: 649558019–651399477	3.04–17.50	9.77–71.50	2009, 2012, BLUP	
qtnGS-2H-17	GS	Vrs1	2H: 652030802	11.28–48.96	49.99–72.25	2010, 2012, BLUP	Qgs2-4 (Wang et al., 2016a)

(Continued)

TABLE 1 | Continued

QTN (QTN cluster)[a]	Trait	Marker associated	Physic position (bp)	LOD score	r^2 (%)	Year	Previous QTL
qtnGS-2H-18	GS	*2_625783669*	2H: 764361924	3.48–5.87	1.06–10.05	2009, 2010	
qtnGWS-2H-19	GWS	*2_625783669*	2H: 764361924	3.29–7.37	0.74–6.25	2009, 2010	
qtn(GP-GWP-GWS)-4H-2	GP, GWP, GWS	*4_497278091*	4H: 596447744	3.04–6.81	3.97–9.17	2009	
qtnTGW-2H-20	TGW	*Vrs1*	2H: 652030802	4.82–49.42	32.77–56.66	2009, 2010, BLUP	*Qtgw2-1, Qtgw2-2* (Wang et al., 2016a)
qtnTGW-3H-5	TGW	*3HS_23539468*	3H: 272283784	3.88–5.98	2.19–4.39	2010, BLUP	
qtnTGW-7H-8	TGW	*7HS_10887541*	7H: 72344563	6.27–13.16	4.39–11.22	2010, BLUP	*Qtgw7-4* (Wang et al., 2016a)

[a] Reliable QTNs and QTN clusters which was detected at least in 2 years environments and multiple GWAS methods; b, physic position of chromosome based on the blast result for the sequence of marker in barley genome database (http://webblast.ipk-gatersleben.de/barley_ibsc/).

of corresponding phenotypes of the QTN genotypes exist using *t*-test. Here, six reliable QTNs were used to underlying the phenotypes difference as an example. The details were showed in **Figure 3**.

Among six QTNs, three yield related QTNs (*qtnSMS-2H-9* for SMS, *qtnGS-2H-17* for GS, and *qtnTGW-2H-20* for TGW) had the significant differences of phenotypic averages between their two genotypes in all four environments (**Figures 3A–C**), and three PH related QTNs (*qtnPH-7H-2*, *qtnIL2-7H-6*, and *qtnMSL-2H-7*) had the significant differences in all four environments (**Figures 3D–F**), indicating their reliability. It was worth noting that the *qtnMSL-2H-7* and *qtnSMS-2H-9* was detected in all the four environments, the *qtnPH-7H-2*, *qtnTGW-2H-20*, and *qtnGS-2H-17* was only detected in three environments, and the *qtnIL2-7H-6* was detected only in 2012 and the BLUP (**Table 1**).

Identification of Candidate Genes Around Reliable QTNs (or Clusters)

According to the recently released genome sequence of barley (Ibsc, 2016; Beier et al., 2017; Mascher et al., 2017) and the gene annotation information, ten candidate genes for the traits of interest were detected around the reliable QTNs and QTN clusters (**Table 2**). The chromosomal distribution of candidate genes was showed in **Figure 2**. Among which, two genes correspond to the previously reported genes, such as *sdw1/denso* and *Vrs1*, while eight candidate genes were new and their functions were derived from the annotated information, which need to be further validated (**Table 2**). Among eight new genes, *HORVU3Hr1G090970*, *HORVU7Hr1G040290*, *HORVU7Hr1G058360*, and *HORVU3Hr1G096010* are related to plant height and its component traits, encoding SAM-dependent_Mtases, Alpha-mannosidase, DHHC-cysteine-rich domain S-acyltransferase and Homeobox-like, respectively. *HORVU2Hr1G113880*, *HORVU2Hr1G094080*, *HORVU2Hr1G126690*, and *HORVU4Hr1G075070* are associated with spike and yield related traits, encoding AP2-like ethylene-responsive transcription factor, BTB/POZ domain protein, Acyl-CoA N-acetyltransferase and Patatin, respectively (**Table 2**).

DISCUSSION

Previously Reported and Novel QTNs Detected With Multi-Locus GWAS Analysis

The comparison between the reliable QTNs (or clusters) for the main agronomic traits and the reliable QTLs in previous studies (Ren et al., 2010, 2013, 2014; Wang et al., 2016a) were conducted (**Table 1**; **Figure 2**). According to the physic positions of associated markers, the reliable QTNs (or clusters) in this study were integrated to the physic map with the reliable QTLs using MapChart 2.32 (Voorrips, 2002; **Figure 2**). Among 39 reliable QTNs (QTN clusters) detected by GWAS, 19 were located on the same regions of QTLs in previous studies (Sameri et al., 2009; Ren et al., 2014; Wang et al., 2016a), while 20 reliable QTNs (QTN clusters) including some minor effect QTNs were novel (**Table 1**). Totally, 8 of 18 QTNs (QTN clusters) associated with plant height related traits, were same as those in previous studies (Sameri et al., 2009; Ren et al., 2014), and 10 were new. For spike and yield related traits, 11 of 21 QTNs (QTN clusters) were the same as the QTLs in Wang et al. (2016a), the others were new in the current study (**Table 1**).

Among 8 previously reported QTNs (QTN clusters) for PH related traits, QTN clusters *qtncPH-3H-1* (3H: 631,870,705–633,068,955 bp), *qtncIL3-3H-3* (3H: 631,870,705–636,535,362 bp), and *qtncIL4-3H-4* (3H: 631,870,705–636,535,362 bp), located on the hotspot of 3H (**Figure 2**), were significantly associated with PH, IL3, and IL4, respectively (**Table 1**). Close to the region of the QTNs clusters, *Qcl3-13* for CL (the length from the ground to the collar equal PH minus MSL), *Qitw3-13* for IL2, *Qith3-13* for IL3, and *Qifo3-14* for IL4 were detected close to the region of SSR markers *Bmag13* (Position: 608,671,381 bp) and *Bmag877* (Position: 657,045,459 bp) on 3H, respectively (Ren et al., 2014). Moreover, *qIN6-3HL* was detected to be significantly associated with IN6 (sixth internode) between *Bmag13* (3H: 608,671,381 bp) and *e06m30.8.3* in Sameri et al. (2009). Compared to the physical positions of the QTLs with those of QTNs, the QTN clusters for PH, IL3, and IL4 should correspond to the QTL *Qcl3-13* for CL, *Qith3-13* for IL3, and *Qifo3-14* for IL4, respectively (Ren et al., 2014; **Table 1**).

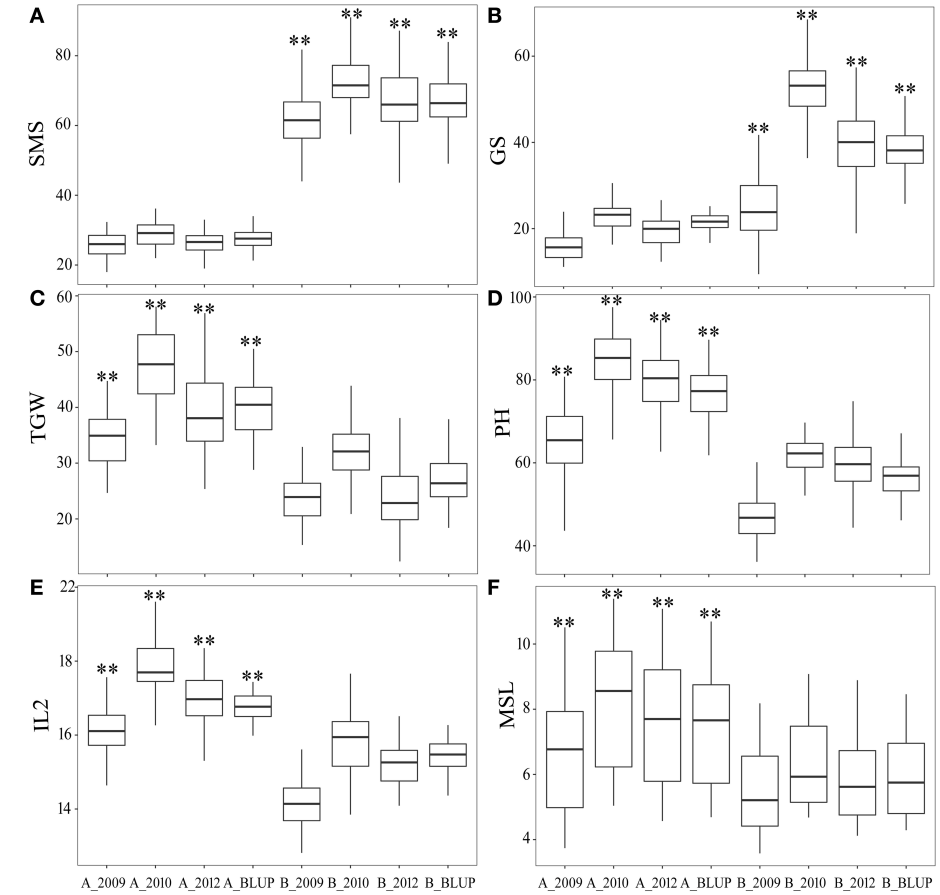


FIGURE 3 | The difference of phenotypes between two kinds of genotypes for each of the six QTNs. **(A)**: SMS at *qtnSMS-2H-9*, **(B)**: GS at *qtnGS-2H-17*, **(C)**: TGW at *qtnTGW-2H-20*, **(D)**: PH at *qtnPH-7H-2*, **(E)**: IL2 at *qtnIL2-7H-6*, **(F)**: MSL at *qtnMSL-2H-7*. **Significant difference at $P < 0.01$.

The *qtnPH-7H-1* (7H: 81,959,684 bp), *qtnPH-7H-2* (7H: 108,670,637 bp), and *qtncIL1-7H-4* (7H: 81,889,341–84,350,472 bp) were detected in same region of chromosome 7HS to be significantly associated with PH and IL1, respectively. These three QTNs are likely the same to the two QTLs *qIN3-7Hs* and *qIN4-7Hs* identified between markers *HVCMA* (7H: 75,227,158 bp) and *ABG701* (7H: 90,406,550 bp) in Sameri et al. (2009) (**Table 1**). No consistent QTLs were detected in the region of these QTNs by Ren et al. (2014). It seems that GWAS can detect more minor QTNs for interesting traits than traditional QTL analysis.

Among 11 previously reported QTNs (or clusters) for spike and yield related traits, it is worth noting that most were located close to the region of *Vrs1* gene. Four QTNs and three QTN clusters, such as *qtnSMS-2H-9* (2H: 652,030,802 bp) for SMS, *qtnGS-2H-17* (2H: 652,030,802 bp) for GS, *qtnTGW-2H-20* (2H: 652,030,802 bp) for TGW, *qtnSLP-2H-10* (2H: 649,558,019 bp) and *qtncSLP-2H-11* (2H: 652,030,802–653,982,961 bp) for SLP, and *qtncGP-2H-14* (2H: 649,558,019–650,438,830 bp) and *qtncGP-2H-15* (2H: 652,030,802–652,604,015 bp) for GP were detected at *Vrs1* (2H: 652,030,802 bp) and the nearby region of 2H with highly phenotypic variation (**Table 1**; **Figure 2**), which was consistent with the QTLs detected in Wang et al. (2016a), including *Qsms2-7* (2H: 652,507,869 bp) for SMS, *Qgs2-4* (2H: 651,436,685 bp) for GS, *Qtgw2-1* (2H: 652,507,869 bp) and *Qtgw2-2* (2H: 652,508,158 bp) for TGW, *Qslp2-6* (2H: 651,436,685 bp) for SLP, and *Qgp2-2* (2H: 651,436,685 bp) for GP (**Table 1**; **Figure 2**). Moreover, QTNs (or clusters) *qtnMSL-2H-7* (2H: 727,985,438 bp), *qtnSMS-2H-8* (2H: 535,680,815 bp), *qtncSLP-4H-1* (4H: 15,498,372–16,761,959 bp), and *qtnTGW-7H-8* (7H: 72,344,563 bp) were consistent with the QTLs *Qmsl2-7* (2H: 724,577,184 bp), *Qsms2-1* (2H: 541,758,123 bp), *Qslp4-2* (4H: 24,332,575 bp), and *Qtgw7-4* (7H: 71,957,427 bp) of Wang et al. (2016a) (**Table 1**; **Figure 2**).

For 20 novel reliable QTNs (QTN clusters) detected by multi-locus GWAS, most had minor effects (**Table 1**). It is worth noting that *qtnIL2-7H-6* was a novel QTN associated with IL2 accounting for a higher proportion of phenotypic variation (22.85–55.77%). And the candidate gene *HORVU7Hr1G058360* (7H: 258,860,422–258,866,854 bp), close to *qtnIL2-7H-6* (7H: 258,071,311 bp), may involve in regulating IL2. For other minor effect QTNs (QTN clusters), three reliable candidate genes (*HORVU2Hr1G094080*, *HORVU2Hr1G126690*, and *HORVU4Hr1G075070*) were identified to be close to the

TABLE 2 | Candidate genes around the reliable QTNs and QTN clusters.

QTN (QTN cluster)[a]	Trait	Marker	Physic position (bp)[b]	candidate gene[c]	Physic position (bp)	Annotation[d]
qtncPH-3H-1	PH	*3HL_37004393–3_511749149*	3H: 631870705–633068955	*HORVU3Hr1G090980*, *HORVU3Hr1G090970*	3H: 634,077,598–634,081,600, 3H: 634,071,757–634,080,040	*sdw1/denso*, GA20-oxidases; SAM-dependent_Mtases, like *PvSAMS*
qtncIL3-3H-3	IL3	*3HL_37004393–3_511668322*	3H: 631870705–636535362	*HORVU3Hr1G090980*, *HORVU3Hr1G090970*	3H: 634,077,598–634,081,600, 3H: 634,071,757–634,080,040	*sdw1/denso*, GA20-oxidases; SAM-dependent_Mtases, like *PvSAMS*
qtncIL4-3H-4	IL4	*3HL_37004393–3_511668322*	3H: 631870705–636535362	*HORVU3Hr1G090980*, *HORVU3Hr1G090970*	3H: 634,077,598–634,081,600, 3H: 634,071,757–634,080,040	*sdw1/denso*, GA20-oxidases; SAM-dependent_Mtases, like *PvSAMS*
qtnPH-7H-2	PH	*7HS_33062962*	7H: 108670637	*HORVU7Hr1G040290*	7H:108834015–108839990	*AMS1p*, Alpha-mannosidase
qtnIL2-7H-6	IL2	*M_114215_455*	7H: 258071311	*HORVU7Hr1G058360*	7H: 258,860,422–258,866,854	DHHC-cysteine-rich domain S-acyltransferase
qtnPH-3H-2	PH	*2HS_32409186*	3H: 651696476	*HORVU3Hr1G096010*	3H: 651,659,644–651,663,119	Homeobox-like, SANT/Myb
qtnSMS-2H-9	SMS	*Vrs1*	2H: 652030802	*HORVU2Hr1G092290*	2H: 652,031,058–652,032,990	*Vrs1*, Homeobox
qtncSLP-2H-11	SLP	*Vrs1–2HL_17075593*	2H: 652030802–653982961	*HORVU2Hr1G092290*	2H: 652,031,058–652,032,990	*Vrs1*, Homeobox
qtncGP-2H-15	GP	*Vrs1–2_527241334*	2H: 652030802–652604015	*HORVU2Hr1G092290*	2H: 652,031,058–652,032,990	*Vrs1*, Homeobox
qtnGS-2H-17	GS	*Vrs1*	2H: 652030802	*HORVU2Hr1G092290*	2H: 652,031,058–652,032,990	*Vrs1*, Homeobox
qtnTGW-2H-20	TGW	*Vrs1*	2H: 652030802	*HORVU2Hr1G092290*	2H: 652,031,058–652,032,990	*Vrs1*, Homeobox
qtnMSL-2H-7	MSL	*2_600749073*	2H: 727985438	*HORVU2Hr1G113880*	2H:730027508–730030208	AP2-like ethylene-responsive transcription factor
qtncSP-2H-13	SP	*2_531255437–M_124056_833*	2H: 662335248–663628734	*HORVU2Hr1G094080*	2H: 662,298,334–662,300,891	BTB/POZ
qtnGS-2H-18	GS	*2_625783669*	2H: 764361924	*HORVU2Hr1G126690*	2H: 764,279,329–764,290,102	Acyl-CoA N-acetyltransferase
qtnGWS-2H-19	GWS	*2_625783669*	2H: 764361924	*HORVU2Hr1G126690*	2H: 764,279,329–764,290,102	Acyl-CoA N-acetyltransferase
qtn(GP-GWP-GWS)-4H-2	GP, GWP, GWS	*4_497278091*	4H: 596447744	*HORVU4Hr1G075070*	4H: 596,446,043–596,448,382	Patatin

[a] Reliable QTNs and QTN clusters which were detected at least in 2 years environments and multiple GWAS methods; [b] Physic position of chromosome based on the blast result for the sequence of marker in barley genome database (http://webblast.ipk-gatersleben.de/barley_ibsc/); [c] Candidate gene was acquired from http://plants.ensembl.org/Hordeum_vulgare/Info/Index; [d] Annotation information was from the database http://plants.ensembl.org/Hordeum_vulgare/Info/Index and https://www.uniprot.org/uniprot.

region of *qtncSP-2H-13*, *qtnGS-2H-18* (or *qtnGWS-2H-19*), and *qtn(GP-GWP-GWS)-4H-2*, respectively (**Table 2**). Therefore, the novel QTNs (QTN clusters) by multi-locus GWAS were reliable, even with minor effect. In other words, multi-locus GWAS can detect more minor effect QTNs than traditional QTL analysis.

The comparison between QTLs in previous studies (Sameri et al., 2009; Ren et al., 2013, 2014; Wang et al., 2016a) and in current study indicated that the consistent results should be much more reliable, which is valuable for further gene cloning and molecular marker assistant selection for breeding. Meanwhile, it illustrated that GWAS is feasible and reliable to detect significant associations for complex quantitative traits in DH population. Moreover, some new QTNs with minor effects were detected, suggesting that GWAS should be a good complementary to traditional QTL mapping. The combination of linkage and association analysis should provide the more accurate and powerful approach for reveling the genetic base of complex quantitative traits (Ott et al., 2011).

Two Previous Reported Genes Reveal the Reliability of Multi-Locus GWAS

HORVU3Hr1G090980 (3H: 634,077,598–634,081,600 bp) was identified on the region (3H: 631,870,705–636,535,362 bp) of three QTN clusters for plant height related traits (*qtncPH-3H-1* for PH, *qtncIL3-3H-3* for IL3, and *qtncIL4-3H-4* for IL4; **Table 2**;

Figure 2). This gene *HORVU3Hr1G090980* correspondeds to the previously reported gene *sdw1/denso,* which was a semi-dwarf gene encoding a gibberellin 20-oxidase enzyme in barley (Jia et al., 2009; Xu et al., 2017). It was clarified that GA 20-oxidases encoded by *sdw1/denso* affected the plant height involving in the later steps of GA biosynthesis (Spielmeyer et al., 2004; Jia et al., 2009; Xu et al., 2017; **Table 2**). Therefore, the association between three QTN clusters and three height related traits (PH, IL3, and IL4) might be the effect of the gene *sdw1/denso*. *Vrs1* (2H: 652,030,802 bp) was detected to be associated with SMS (*qtnSMS-2H-9*), GS (*qtnGS-2H-17*), and TGW (*qtnTGW-2H-20*) with high proportion of phenotypic variation in multiple environments and by multi-locus GWAS methods in current study (**Tables 1**, **2**). Around the gene *Vrs1*, moreover, five QTNs (or clusters), including *qtnSLP-2H-10* (2H: 649,558,019 bp) and *qtncSLP-2H-11* (2H: 652,030,802–653,982,961 bp) for SLP, *qtncGP-2H-14* (2H: 649,558,019–650,438,830 bp), and *qtncGP-2H-15* (2H: 652,030,802–652,604,015 bp) for GP, and *qtncSP-2H-12* (2H: 648,821,931–652,030,802 bp) for SP, were detected in multiple situations and by multi-locus GWAS methods with high proportion of phenotypic variation (**Table 1**; **Figure 2**). *Vrs1*, controlling row number of barley, encodes homeobox and profoundly affects barley spike morphology (Komatsuda et al., 2007). The identification of the reliable QTNs (or clusters) around two previously reported genes using multi-locus GWAS further revealed the reliability of multi-locus GWAS in bi-parental segregation population.

Novel Candidate Genes Reveal the Possible Molecular Basis of Plant Height- and Yield-Related Traits

Gene *HORVU3Hr1G090970* (3H: 634,071,757–634,080,040 bp), encoded a S-adenosylmethionine (SAM)-dependent-MTases, which was identified on the region (3H: 631,870,705–636,535,362 bp) of three QTN clusters for plant height related traits (*qtncPH-3H-1* for PH, *qtncIL3-3H-3* for IL3, and *qtncIL4-3H-4* for IL4; **Table 2**; **Figure 2**). It was reported that SAM biosynthetic pathways affects lignin biosynthesis in switchgrass (*Panicum virgatum* L.; Bai et al., 2018; **Table 2**). Cystathionine c-synthase (CGS) is the first committed enzyme for the biosynthesis of Met that can be metabolized to SAM, CGS-RNAi transgenic switchgrass lines showed much shorter plant height and internode length (Bai et al., 2018). Therefore, *HORVU3Hr1G090970* might be a new semi-dwarf gene in barley.

HORVU7Hr1G008720, was identified at 7H:108,834,015–108,839,990 bp, which was 0.15Mb from the QTN *qtnPH-7H-2* for PH, encoding Alpha-mannosidase like *AMS1p* (**Table 2**; **Figure 2**). Alpha-mannosidase is the component of cell wall, involving in cell wall biosynthesis or modification, which participated in the cell growth of internodes with pectinesterase and alpha-xylosidase in plant (Wu and Cao, 2008). Moreover, Alpha-mannosidase is the member of cytoplasm-to-vacuole targeting (Cvt) pathway with *AuTophaGy8* (*ATG8*) gene, and soybean transgenic lines over-expressed *GmATG8c* showed higher plant height than the wild type (Xia et al., 2012). Therefore, *HORVU7Hr1G008720* should be a reliable candidate gene, which affected PH by regulating the cell growth as *AMS1p* does.

HORVU7Hr1G058360 at 7H: 258,860,422-258,866,854 bp was close to *qtnIL2-7H-6* (7H: 258,071,311 bp) for IL2, encoding a S-acyltransferase with DHHC-cysteine-rich domain (**Table 2**; **Figure 2**). DHHC-cysteine-rich domain S-acyltransferase proteins are involved in plant development and stress responses in *Arabidopsis* (Li et al., 2016). *AtPAT10* is an S-acyl transferase, which affects the vascular development through controlling the cell division and expansion in *Arabidopsis*. *AtPAT10* mutants are semi-dwarfed, and the reduction of plant height is due to the reduced length of the internodes, which appears to be the result of reduction in both cell number and cell size in these tissues (Qi et al., 2013). Therefore, *HORVU7Hr1G058360* is a reliable candidate gene regulating the IL2 as *AtPAT10* dose.

HORVU3Hr1G096010 (3H: 651,659,644–651,663,119 bp) encoding homeobox-like protein with SANT/MYB domain, was close to the QTN *qtnPH-3H-2* (3H: 651,696,476 bp) for PH (**Table 2**; **Figure 2**). Homeobox gene was reported to be involved in the regulation of morphological development in plants, homeobox gene *OSH15* affects the architecture of internodes resulting in d6 dwarf plants (Sato et al., 1999). Moreover, the *RAD* gene in *Arabidopsis*, encoding small plant-specific single SANT/MYB domain protein, affects the growth and development of *Arabidopsis*. Overexpression of the *RAD* gene can repress *Arabidopsis* growth, resulting in dwarfing and delaying flowering (Baxter et al., 2007; Zhang et al., 2011). Thus, *HORVU3Hr1G096010* is a reliable candidate gene regulating plant height as the function of homeobox gene or *RAD* gene.

For main spike length (MSL), *HORVU2Hr1G113880* corresponding to *Cly1* gene, was identified at 2H:730,027,508–730,030,208 bp, which was 3 Mb from the QTN *qtnMSL-2H-7* (2H: 727,985,438 bp) for MSL (**Table 2**; **Figure 2**). *Cly1* encodes for an AP2-protein that inhibits development of flower (Nair et al., 2010; Terzi et al., 2017). *HvAP2* regulates the length of a critical developmental window required for the elongation of the inflorescence internodes in barley (Houston et al., 2013). Therefore, the association between *qtnMSL-2H-3* and MSL might be the effect of *Cly1* gene.

For spike number per plant (SP), *HORVU2Hr1G094080* (2H: 662,298,334–662,300,891 bp), encoding a protein with BTB/POZ domain (Broad complex, Tramtrack, Bric à brac, Pox virus and Zinc finger), was detected about 40 Kb from the QTN cluster *qtncSP-2H-13* (2H: 662,335,248–663,628,734 bp; **Table 2**; **Figure 2**). *HvCul4* gene encodes a BLADE-ON-PETIOLE-like (BOP-Like) protein containing BTB/POZ domain, which shares high similarity with *Arabidopsis BOP1* and *BOP2* (Tavakol et al., 2015; Jost et al., 2016). It was reported that *HvCul4* controlled the tiller and leaf pattern in barley (Tavakol et al., 2015; Jost et al., 2016), and *Arabidopsis BOP1* and *BOP2* acted at boundary regions to regulate axillary development and leaf morphogenesis (Ha et al., 2004). In addition, according to the barley expression database from Barlex (http://barlex.barleysequence.org), *HORVU2Hr1G094080.1* showed highest level of expression in developing inflorescences. Therefore, candidate gene *HORVU2Hr1G094080* performed the similar function as *HvCul4*, *BOP1*, and *BOP2* to control the spike number per plant (SP).

For grain number per spike (GS) and grain weight per spike (GWS), marker *2_625783669* (2H: 764,361,924 bp) was detected significantly associated with these 2 traits. *HORVU2Hr1G126690* (2H: 764,279,329–764,290,102 bp), encoding a protein with N-acetyltransferase domain, was detected about 83 Kb from the marker *2_625783669* (**Table 2**; **Figure 2**). *OsSNAT1* encodes N-acetyltransferase1, it was reported that overexpression of T2 homozygous *OsSNAT1* in rice increased panicle number and seed weight per plant, while decreased spikelet numbers per panicle under paddy field conditions (Lee and Back, 2017). Moreover, the expression of *HORVU2Hr1G126690.4* is much higher in developing inflorescences than in other tissues according to the barley expression database from Barlex (http://barlex.barleysequence.org). Thus, the candidate gene *HORVU2Hr1G126690* may affect the GS and GWS through the similar function of *OsSNAT1*.

Marker *4_497278091* (4H: 596,447,744 bp) was significant associated with GP, GWP, and GWS. *HORVU4Hr1G075070* (4H: 596,446,043–596,448,382 bp), encoding Patatin, was detected at *4_497278091* (**Table 2**). Overexpression of a patatin-like protein in *Camelina sativa* (Li et al., 2015) or in *Arabidopsis* (Li et al., 2013) reduced growth and overall seed production, but increased seed oil content. Therefore, *HORVU4Hr1G075070* is a reliable candidate gene which might affect GP, GWP and GWS as the function of patatin-like gene.

Among the above ten candidate genes, two were previously reported, such as *sdw1/denso* and *Vrs1*, eight were new, which were derived from the annotated information. Based on the annotations of these candidate genes, homologous genes or proteins with same function or function domain were reported to be regulated the corresponding traits in barley, *Arabidopsis* and rice. The reliable QTNs and QTN clusters for these traits may be the effect of the candidate genes with similar function as the homologous genes or proteins does. The functions of eight reliable candidate genes need to be further validated. In summary, it is feasible and reliable to use multi-locus GWAS in bi-parental segregation populations.

The New Multi-Locus GWAS for Bi-Parental Segregation Population

Traditionally, segregation populations were used for QTL analysis, and GWAS are commonly used in natural populations. Nowadays, as the development of high-throughput SNP markers and high-throughput phenotypes, GWAS have been widely applied to the genetic analysis for complex traits in family-based populations (such as NAM and MAGIC populations) and proved to be powerful tool for uncovering the basis of key agronomic traits in maize and barley (Tian et al., 2011; Cook et al., 2012; Maurer et al., 2015, 2016). However, for single segregating population, successful but fewer cases were performed using GWAS (Gao et al., 2015; Henning et al., 2016; Liu et al., 2018). It indicated that GWAS for segregating population are feasible. However, high false positive rate is an obvious problem in the traditional single-locus GWAS using general linear models (GLMs) and mixed linear models (MLMs) (Zhang et al., 2010; Pace et al., 2015). And the P threshold ($P = 0.05/n$, n is the number of SNPs) leads to missing many significant QTNs, particularly small-effect QTNs (Wang et al., 2016b). Some multi-locus GWAS methodologies, such as mrMLM (Wang et al., 2016b), FASTmrMLM (Zhang and Tamba, 2018), FASTmrEMMA (Wen et al., 2018), ISIS EM-BLASSO (Tamba et al., 2017), pLARmEB (Zhang et al., 2017), and pKWmEB (Ren et al., 2018) have been developed to remedy the shortcomings mentioned above. These multi-locus GWAS methods have been used to analyze the published data, indicated that these methods constituted effective approaches with high detection power and less stringent criteria (Wang et al., 2016b; Tamba et al., 2017; Zhang et al., 2017; Wen et al., 2018). Totally, five multi-locus GWAS methods were used in our study, which improved the detection power and accuracy of QTNs for interesting traits. Moreover, the QTNs (QTN clusters), which were repeatedly detected in multiple environments and GWAS methods, were selected as reliable QTNs (QTN clusters). This greatly improved the accuracy of the association results and reduced its false positive, and more small-effect QTNs were detected within a certain rate of false positive. In addition, the t-test results of the phenotypic difference corresponding to QTNs, demonstrated that GWAS have the more stringent threshold of significance than t-test, and the advantages of accuracy and false positive controlling. In our study, 39 reliable QTNs and QTN clusters were detected, among which 10 reliable candidate gene were identified. Meanwhile, several new reliable QTNs with small-effect were also detected, which were different from the previous reports (**Table 1**). There was a limitation to identify candidate genes for all the reliable QTNs and QTN clusters, especially the small-effect ones, based on the imperfect annotation database of barley. However, these results indicated that multi-locus GWAS methods are feasible and reliable for DH population, and good complementary to traditional QTL mapping for the detection of new reliable QTNs even with small-effect. which will provide more useful information for future works.

CONCLUSIONS

Available online at: In this study, five multi-locus GWAS methods were performed for 14 main agronomic traits in 122 doubled haploid (DH) lines. Thirty-nine reliable QTNs and/or QTN clusters were repeatedly detected in multiple environments and methods 10 candidate genes for the interest traits were detected, 19 QTNs and two genes (*sdw1/denso* and *Vrs1*) were previously reported, and eight candidate genes need to be further validated.

The results validated the feasibility and reliability of GWAS in DH population and the good complementary to traditional QTL analysis. All the results will facilitate elucidating genetic basis of agronomic traits and improving marker-assisted selection breeding in barley.

AUTHOR CONTRIBUTIONS

DS and XR conceived and designed the experiments. XH, XR, JW, and LL conducted the experiments and phenotyping measurements. XH and JZ performed the analysis. XH and JZ wrote the paper. GS, XR, and DS modified the manuscript. CL produced the Huaai 11 and Huadamai 6 DH population. All the authors read and approved the final version of this manuscript.

FUNDING

This project was supported by the earmarked fund for China Agriculture Research System (CARS-5).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01683/full#supplementary-material

REFERENCES

Baghizadeh, A., Taleei, A. R., and Naghavi, M. R. (2007). QTL analysis for some agronomic traits in barley (*Hordeum vulgare* L.). *Int. J. Agric. Biol.* 9, 372–374.

Bai, Z., Qi, T., Liu, Y., Wu, Z., Ma, L., Liu, W., et al. (2018). Alteration of S-adenosylhomocysteine levels affects lignin biosynthesis in switchgrass. *Plant Biotechnol. J.* 16, 2016–2026. doi: 10.1111/pbi.12935

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version. *J. Stat. Softw.* 1, 1–23.

Baxter, C. E., Costa, M. M. R., and Coen, E. S. (2007). Diversification and co-option of RAD-like genes in the evolution of floral asymmetry. *Plant J.* 52, 105–113. doi: 10.1111/j.1365-313X.2007.03222.x

Beier, S., Himmelbach, A., Colmsee, C., Zhang, X. Q., Barrero, R. A., Zhang, Q., et al. (2017). Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. *Sci. Data* 4, 170044. doi: 10.1038/sdata.2017.44

Cook, J. P., McMullen, M. D., Holland, J. B., Tian, F., Bradbury, P., Ross-Ibarra, J., et al. (2012). Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiol.* 158, 824–834. doi: 10.1104/pp.111.185033

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587. doi: 10.3410/f.1015548.197423

Fan, Y., Zhou, G., Shabala, S., Chen, Z.-H., Cai, S., Li, C., et al. (2016). Genome-wide association study reveals a new QTL for salinity tolerance in barley (*Hordeum vulgare L.*). *Front. Plant Sci.* 7:946. doi: 10.3389/fpls.2016.00946

Gao, L. L., Kielsmeier-Cook, J., Bajgain, P., Zhang, X. F., Chao, S. M., Rouse, M. N., et al. (2015). Development of genotyping by sequencing (GBS)- and array-derived SNP markers for stem rust resistance gene *Sr42*. *Mol. Breeding* 35, 207. doi: 10.1007/s11032-015-0404-4

Gawenda, I., Thorwarth, P., Günther, T., Ordon, F., and Schmid, K. J. (2015). Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. *Plant Breeding* 134, 28–39. doi: 10.1111/pbr.12237

Ha, C. M., Jun, J. H., Nam, H. G., and Fletcher, J. C. (2004). *BLADE-ON-PETIOLE1* encodes a BTB/POZ domain protein required for leaf morphogenesis in *Arabidopsis thaliana*. *Plant Cell Physiol.* 45, 1361–1370. doi: 10.1093/pcp/pch201

Henning, J. A., Gent, D. H., Twomey, M. C., Townsend, M. S., Pitra, N. J., and Matthews, P. D. (2016). Genotyping-by-sequencing of a bi-parental mapping population segregating for downy mildew resistance in hop (*Humulus lupulus* L.). *Euphytica* 208, 545–559. doi: 10.1007/s10681-015-1600-3

Houston, K., McKim, S. M., Comadran, J., Bonar, N., Druka, I., Uzrek, N., et al. (2013). Variation in the interaction between alleles of *HvAPETALA2* and *microRNA172* determines the density of grains on the barley inflorescence. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16675–16680. doi: 10.1073/pnas.1311681110

Hu, X., Ren, J., Ren, X., Huang, S., Sabiel, S. A., Luo, M., et al. (2015). Association of agronomic traits with SNP markers in durum wheat (*Triticum turgidum* L. *durum (Desf.)*). *PLoS ONE* 10:e0130854. doi: 10.1371/journal.pone.0130854

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967. doi: 10.1038/ng.695

Ibsc (2016). Pseudomolecules of the map-based reference genome assembly of barley cv. *Morex*. doi: 10.5447/ipk/2016/34

Jia, Q. J., Zhang, J. J., Westcott, S., Zhang, X. Q., Bellgard, M., Lance, R., et al. (2009). GA-20 oxidase as a candidate for the semidwarf gene *sdw1/denso* in barley. *Funct. Integr. Genomic.* 9, 255–262. doi: 10.1007/s10142-009-0120-4

Jost, M., Taketa, S., Mascher, M., Himmelbach, A., Yuo, T., Shahinnia, F., et al. (2016). A homolog of *Blade-On-Petiole 1* and *2* (*BOP1/2*) controls internode length and homeotic changes of the barley inflorescence. *Plant Physiol.* 171, 1113–1127. doi: 10.1104/pp.16.00124

Komatsuda, T., Pourkheirandish, M., He, C., Azhaguvel, P., Kanamori, H., Perovic, D., et al. (2007). Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1424–1429. doi: 10.1073/pnas.0608580104

Lee, K., and Back, K. (2017). Overexpression of rice serotonin N-acetyltransferase 1 in transgenic rice plants confers resistance to cadmium and senescence and increases grain yield. *J. Pineal. Res.* 62:e12392. doi: 10.1111/jpi.12392

Li, J., Huang, X., Heinrichs, F., Ganal, M., and Röder, M. (2005). Analysis of QTLs for yield, yield components, and malting quality in a BC_3-DH population of spring barley. *Theor. Appl. Genet.* 110, 356–363. doi: 10.1007/s00122-004-1847-x

Li, M., Bahn, S. C., Fan, C., Li, J., Phan, T., Ortiz, M., et al. (2013). Patatin-related phospholipase pPLAIIIä increases seed oil content with long-chain fatty acids in *Arabidopsis*. *Plant Physiol*. 162, 39–51. doi: 10.1104/pp.113.216994

Li, M., Wei, F., Tawfall, A., Tang, M., Saettele, A., and Wang, X. (2015). Overexpression of patatin-related phospholipase AIIIδ altered plant growth and increased seed oil content in camelina. *Plant Biotechnol. J.* 13, 766–778. doi: 10.1111/pbi.12304

Li, S. S., Jia, J. Z., Wei, X. Y., Zhang, X. C., Li, L. Z., Chen, H. M., et al. (2007). A intervarietal genetic map and QTL analysis for yield traits in wheat. *Mol. Breeding* 20, 167–178. doi: 10.1007/s11032-007-9080-3

Li, Y. X., Lin, J. Z., Li, L., Peng, Y. C., Wang, W. W., Zhou, Y. B., et al. (2016). DHHC-cysteine-rich domain S-acyltransferase protein family in rice: organization, phylogenetic relationship and expression pattern during development and stress. *Plant. Syst. Evol.* 302, 1405–1417. doi: 10.1007/s00606-016-1339-x

Liu, R., Gong, J., Xiao, X., Zhang, Z., Li, J., Liu, A., et al. (2018). GWAS analysis and QTL identification of fiber quality traits and yield components in upland cotton using enriched high-density SNP markers. *Front. Plant Sci.* 9:1067. doi: 10.3389/fpls.2018.01067

Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., et al. (2010). Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proc. Natl. Acad. Sci. U.S.A.* 107, 19585–19590. doi: 10.1073/pnas.1006105107

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433. doi: 10.1038/nature22043

Massman, J., Cooper, B., Horsley, R., Neate, S., Dill-Macky, R., Chao, S., et al. (2011). Genome-wide association mapping of Fusarium head blight resistance in contemporary barley breeding germplasm. *Mol. Breeding* 27, 439–454. doi: 10.1007/s11032-010-9442-0

Maurer, A., Draba, V., Jiang, Y., Schnaithmann, F., Sharma, R., Schumann, E., et al. (2015). Modelling the genetic architecture of flowering time control in barley through nested association mapping. *BMC Genomics* 16:290. doi: 10.1186/s12864-015-1459-7

Maurer, A., Draba, V., and Pillen, K. (2016). Genomic dissection of plant development and its impact on thousand grain weight in barley through nested association mapping. *J. Exp. Bot.* 67, 2507–2518. doi: 10.1093/jxb/erw070

Nair, S. K., Wang, N., Turuspekov, Y., Pourkheirandish, M., Sinsuwongwat, S., Chen, G., et al. (2010). Cleistogamous flowering in barley arises from the suppression of microRNA-guided *HvAP2* mRNA cleavage. *Proc. Natl. Acad. Sci. U.S.A.* 107, 490–495. doi: 10.1073/pnas.0909097107

Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nat. Rev. Genet.* 12, 465–474. doi: 10.1038/nrg2989

Pace, J., Yu, X., and Lubberstedt, T. (2015). Genomic prediction of seedling root length in maize (*Zea mays* L.). *Plant J.* 83, 903–912. doi: 10.1111/tpj.12937

Pasam, R. K., Sharma, R., Malosetti, M., van Eeuwijk, F. A., Haseneyer, G., Kilian, B., et al. (2012). Genome-wide association studies for agronomical traits in a world wide spring barley collection. *BMC Plant Biol.* 12:16. doi: 10.1186/1471-2229-12-16

Peng, B., Li, Y., Wang, Y., Liu, C., Liu, Z., Tan, W., et al. (2011). QTL analysis for yield components and kernel-related traits in maize across multi-environments. *Theor. Appl. Genet.* 122, 1305–1320. doi: 10.1007/s00122-011-1532-9

Qi, B., Doughty, J., and Hooley, R. (2013). A Golgi and tonoplast localized S-acyl transferase is involved in cell expansion, cell division, vascular patterning and fertility in Arabidopsis. *New Phytol.* 200, 444–456. doi: 10.1111/nph.12385

Reif, J. C., Liu, W., Gowda, M., Maurer, H. P., Mohring, J., Fischer, S., et al. (2010). Genetic basis of agronomically important traits in sugar beet (*Beta vulgaris* L.) investigated with joint linkage association mapping. *Theor. Appl. Genet.* 121, 1489–1499. doi: 10.1007/s00122-010-1405-7

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2018). pKWmEB: integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Ren, X., Sun, D., Guan, W., Sun, G., and Li, C. (2010). Inheritance and identification of molecular markers associated with a novel dwarfing gene in barley. *BMC Genet.* 11:89. doi: 10.1186/1471-2156-11-89

Ren, X., Wang, J., Liu, L., Sun, G., Li, C., Luo, H., et al. (2016). SNP-based high density genetic map and mapping of *btwd1* dwarfing gene in barley. *Sci. Rep.* 6:31741. doi: 10.1038/srep31741

Ren, X. F., Sun, D. F., Dong, W. B., Sun, G. L., and Li, C. D. (2014). Molecular detection of QTL controlling plant height components in a doubled haploid barley population. *Genet. Mol. Res.* 13, 3089–3099. doi: 10.4238/2014.April.17.5

Ren, X. F., Sun, D. F., Sun, G. L., Li, C. D., and Dong, W. B. (2013). Molecular detection of QTL for agronomic and quality traits in a doubled haploid barley population. *Aust. J. Crop Sci.* 7, 878–886.

Sameri, M., Nakamura, S., Nair, S. K., Takeda, K., and Komatsuda, T. (2009). A quantitative trait locus for reduced culm internode length in barley segregates as a Mendelian gene. *Theor. Appl. Genet.* 118, 643–652. doi: 10.1007/s00122-008-0926-9

Sameri, M., Takeda, K., and Komatsuda, T. (2006). Quantitative trait loci controlling agronomic traits in recombinant inbred lines from a cross of oriental- and occidental-type barley cultivars. *Breeding Sci.* 56, 243–252. doi: 10.1270/jsbbs.56.243

Sato, Y., Sentoku, N., Miura, Y., Hirochika, H., Kitano, H., and Matsuoka, M. (1999). Loss-of-function mutations in the rice homeobox gene *OSH15* affect the architecture of internodes resulting in dwarf plants. *EMBO J.* 18, 992–1002. doi: 10.1093/emboj/18.4.992

Spielmeyer, W., Ellis, M., Robertson, M., Ali, S., Lenton, J. R., and Chandler, P. M. (2004). Isolation of gibberellin metabolic pathway genes from barley and comparative mapping in barley, wheat and rice. *Theor. Appl. Genet.* 109, 847–855. doi: 10.1007/s00122-004-1689-6

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tavakol, E., Okagaki, R., Verderio, G., Shariati, J. V., Hussien, A., Bilgic, H., et al. (2015). The barley *Uniculme4* gene encodes a BLADE-ON-PETIOLE-like protein that controls tillering and leaf patterning. *Plant Physiol.* 168, 164–174. doi: 10.1104/pp.114.252882

Terzi, V., Tumino, G., Pagani, D., Rizza, F., Ghizzoni, R., Morcia, C., et al. (2017). Barley developmental mutants: The high road to understand the cereal spike morphology. *Diversity* 9:21. doi: 10.3390/d9020021

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., et al. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* 43, 159–162. doi: 10.1038/ng.746

Varshney, R. K., Paulo, M. J., Grando, S., van Eeuwijk, F. A., Keizer, L. C. P., Guo, P., et al. (2012). Genome wide association analyses for drought tolerance related traits in barley (*Hordeum vulgare* L.). *Field Crop. Res.* 126, 171–180. doi: 10.1016/j.fcr.2011.10.008

Voorrips, R. E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. *J. Hered.* 93, 77–78. doi: 10.1093/jhered/93.1.77

Wang, J., Sun, G., Ren, X., Li, C., Liu, L., Wang, Q., et al. (2016a). QTL underlying some agronomic traits in barley detected by SNP markers. *BMC Genet.* 17:103. doi: 10.1186/s12863-016-0409-y

Wang, J., Yang, J., McNeil, D. L., and Zhou, M. (2010). Identification and molecular mapping of a dwarfing gene in barley (*Hordeum vulgare* L.) and its correlation with other agronomic traits. *Euphytica* 175, 331–342. doi: 10.1007/s10681-010-0175-2

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016b). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Wójcik-Jagła, M., Fiust, A., Koscielniak, J., and Rapacz, M. (2018). Association mapping of drought tolerance-related traits in barley to complement a traditional biparental QTL mapping study. *Theor. Appl. Genet.* 131, 167–181. doi: 10.1007/s00122-017-2994-1

Wu, T., and Cao, J. S. (2008). Differential gene expression of tropical pumpkin (*Cucurbita moschata* Duchesne) bush mutant during internode development. *Sci. Hortic.* 117, 219–224. doi: 10.1016/j.scienta.2008.04.002

Xia, T., Xiao, D., Liu, D., Chai, W., Gong, Q., and Wang, N. N. (2012). Heterologous expression of *ATG8c* from soybean confers tolerance to nitrogen deficiency and increases yield in Arabidopsis. *PLoS ONE* 7:e37217. doi: 10.1371/journal.pone.0037217

Xu, X., Sharma, R., Tondelli, A., Russell, J., Comadran, J., Schnaithmann, F., et al. (2018). genome-wide association analysis of grain yield-associated traits in a Pan-European barley cultivar collection. *Plant Genome* 11:170073. doi: 10.3835/plantgenome2017.08.0073

Xu, Y., Jia, Q., Zhou, G., Zhang, X. Q., Angessa, T., Broughton, S., et al. (2017). Characterization of the *sdw1* semi-dwarf gene in barley. *BMC Plant Biol.* 17:11. doi: 10.1186/s12870-016-0964-4

Yang, X., Yan, J., Shah, T., Warburton, M. L., Li, Q., Li, L., et al. (2010). Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. *Theor. Appl. Genet.* 121, 417–431. doi: 10.1007/s00122-010-1320-y

Yu, J., and Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. *Curr. Opin. Biotechnol.* 17, 155–160. doi: 10.1016/j.copbio.2006.02.003

Zhang, F., Liu, X., Zuo, K., Sun, X., and Tang, K. (2011). Molecular cloning and expression analysis of a novel *SANT/MYB* gene from *Gossypium barbadense*. *Mol. Biol. Rep.* 38, 2329–2336. doi: 10.1007/s11033-010-0366-x

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y.-M., and Tamba, C. L. (2018). A fast mrMLM algorithm for multi-locus genome-wide association studies. *bioRxiv [preprint]* 341784. doi: 10.1101/341784

Zhang, Y. M., Mao, Y., Xie, C., Smith, H., Luo, L., and Xu, S. (2005). Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). *Genetics* 169, 2267–2275. doi: 10.1534/genetics.104.033217

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhuang, J. Y., Lin, H. X., Lu, J., Qian, H.-R., Hittalmani, S., Huang, N., et al. (1997). Analysis of QTL × environment interaction for yield components and plant height in rice. *Theor. Appl. Genet.* 95, 799–808. doi: 10.1007/s001220050628

Zohary, D., Hopf, M., and Weiss, E. (2012). *Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in South-West Asia, Europe, and the Mediterranean Basin. 4th Edn.* New York, NY: Oxford University Press.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 21 November 2018
doi: 10.3389/fpls.2018.01690

Identification of QTNs Controlling Seed Protein Content in Soybean Using Multi-Locus Genome-Wide Association Studies

Kaixin Zhang[1,2†], Shulin Liu[3†], Wenbin Li[1,2], Shiping Liu[1,2], Xiyu Li[1,2], Yanlong Fang[1,2], Jun Zhang[4], Yue Wang[1,2], Shichao Xu[1,2], Jianan Zhang[1,2], Jie Song[1,2], Zhongying Qi[1,2], Xiaocui Tian[1,2], Zhixi Tian[3], Wen-Xia Li[1,2]* and Hailong Ning[1,2]*

[1] Key Laboratory of Soybean Biology in the Chinese Ministry of Education, Northeast Agricultural University, Harbin, China, [2] Northeastern Key Laboratory of Soybean Biology and Breeding/Genetics in the Chinese Ministry of Agriculture, Northeast Agricultural University, Harbin, China, [3] State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China, [4] College of Agronomy, Jilin Agricultural University, Changchun, China

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Dongying Gao,
University of Georgia, United States
Yongqing Jiao,
Oil Crops Research Institute (CAAS),
China

***Correspondence:**
Wen-Xia Li
liwenxianeau@126.com
Hailong Ning
ninghailongneau@126.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: 15 June 2018
Accepted: 31 October 2018
Published: 21 November 2018

Citation:
Zhang K, Liu S, Li W, Liu S, Li X, Fang Y, Zhang J, Wang Y, Xu S, Zhang J, Song J, Qi Z, Tian X, Tian Z, Li W-X and Ning H (2018) Identification of QTNs Controlling Seed Protein Content in Soybean Using Multi-Locus Genome-Wide Association Studies. Front. Plant Sci. 9:1690. doi: 10.3389/fpls.2018.01690

Protein content (PC), an important trait in soybean (*Glycine max*) breeding, is controlled by multiple genes with relatively small effects. To identify the quantitative trait nucleotides (QTNs) controlling PC, we conducted a multi-locus genome-wide association study (GWAS) for PC in 144 four-way recombinant inbred lines (FW-RILs). All the FW-RILs were phenotyped for PC in 20 environments, including four locations over 4 years with different experimental treatments. Meanwhile, all the FW-RILs were genotyped using SoySNP660k BeadChip, producing genotype data for 109,676 non-redundant single-nucleotide polymorphisms. A total of 129 significant QTNs were identified by five multi-locus GWAS methods. Based on the 22 common QTNs detected by multiple GWAS methods or in multiple environments, pathway analysis identified 8 potential candidate genes that are likely to be involved in protein synthesis and metabolism in soybean seeds. Using superior allele information for 22 common QTNs in 22 elite and 7 inferior lines, we found higher superior allele percentages in the elite lines and lower percentages in the inferior lines. These findings will contribute to the discovery of the polygenic networks controlling PC in soybean, increase our understanding of the genetic foundation and regulation of PC, and be useful for molecular breeding of high-protein soybean varieties.

Keywords: protein content, soybean, multi-locus GWAS, QTNs, four-way recombinant inbred lines

INTRODUCTION

Soybean [*Glycine max* (L.) Merr.] is a globally important high-protein crop, with protein accounting for about 40% of the seed's dry weight. Soybean is one of humans' main sources of dietary protein; therefore, breeding high-protein varieties of soybean is an ongoing, important objective of plant breeders. The efficiency of plant breeding has been greatly accelerated by the emergence of molecular markers and molecular technology, such as random amplified polymorphic DNA, restriction fragment length polymorphism, amplified fragment length polymorphism, simple sequence repeats, specific-locus amplified fragment, and single-nucleotide polymorphism (SNP). For genome-wide association studies (GWAS) in soybean, the acquisition of a large number of molecular markers is extremely important (Song et al., 2013), and SNPs are well suited for such analyses because of their high densities throughout the genome (Gaur et al., 2012).

Breeders and molecular geneticists have routinely used populations derived from biparental crosses for development of new varieties and mapping quantitative trait loci (QTLs) for traits of interest. However, the richness of allelic and phenotypic variation in biparental inter-mated populations is somewhat limited. To overcome this limitation, animal breeders have developed the multi-parental inter-mated population design, for example by using a population descended from eight mouse strains (Yalcin et al., 2005). Recently, multi-parent advanced generation inter-cross (MAGIC) lines have also been developed in plants, including wheat (*Triticum aestivum*; Huang et al., 2012; Mackay et al., 2014), maize (*Zea mays*; Dell'Acqua et al., 2015), *Arabidopsis thaliana* (Kover et al., 2009), barley (*Hordeum vulgare*; Sannemann et al., 2015), tomato (*Solanum lycopersicum*; Pascual et al., 2015), and rice (*Oryza sativa*; Bandillo et al., 2013; Meng et al., 2016a,b). MAGIC populations have more diverse alleles than bi-parental populations, which increases genetic variation. However, obtaining a MAGIC population is labor intensive because of the repeated crossing and requires large population sizes to include recombinants for all the desirable traits. An intermediate design, the four-way cross population, is easier to obtain while providing some of the same benefits as an eight-way cross population.

In recent years, GWAS with high-density SNPs have emerged as very powerful tools for dissecting the genetic basis of complex traits. This approach has been applied to MAGIC populations of many crop plants (Bandillo et al., 2013; Pascual et al., 2015; Meng et al., 2016a,b) but not, as of yet, to soybean. As in soybean, either biparental populations or natural populations have been used in all previous QTL mapping studies (Li et al., 2018; Wang et al., 2018). This previous research indicates that PC in soybean is a typical quantitative trait controlled by multiple genes with relatively small genetic effects, whose identification will require a more efficient method to detected QTLs. Multi-locus GWAS is a suitable method for identifying significant QTNs, especially for relatively small effects; it also has a low false positive rate, and has been used in many studies (Liu et al., 2016; Wang et al., 2016; Tamba et al., 2017; Zhang Y. et al., 2017; Wen et al., 2018).

In this study, we used 144 recombinant inbred lines (FW-RILs) from a four-way cross, which were genotyped by SNPs and phenotyped seed protein content (PC) in different environments. We then combined these data to identify significant QTNs for PC in soybean using multi-locus GWAS methods. The objective was to find common QTNs that were identified by multiple methods or in multiple environments and then deduce potential candidate genes and identify elite lines in the FW-RIL population, as a means to accelerate molecular breeding to increase PC in soybean.

MATERIALS AND METHODS

Plant Materials

Four soybean varieties, Kenfeng14 (PC 41.08%), Kenfeng15 (PC 41.42%), Kenfeng19 (PC 43.06%), and Heinong48 (PC 43.55%), were used to construct a four-way recombinant inbred line (FW-RIL) population. Among these, Kenfeng14, Kenfeng15, and Kenfeng19 were bred by the Heilongjiang Academy of Agricultural Reclamation and derived from the crosses Suinong 10 × Changnong 5, Suinong 14 × Kenjiao 9307, and Hefeng 25 × (Kenfeng 4 × Gong 8861-0), respectively; and Heinong48 was bred by Heilongjiang Academy of Agricultural Science and derived from the cross Ha 90-6719 × Sui 90-5888.

First, two single crosses, Kengfeng14 × Kenfeng15 and Kenfeng19 × Heinong48, were carried out in Harbin (45.75°N, 126.63°E), Heilongjiang Province, China, and the F_1 seeds were harvested in 2008. Second, a cross was conducted between two sets of single-cross F_1 seeds, and F_1 seeds of the resulting four-way cross (Kengfeng14 × Kenfeng15) × (Kenfeng19 × Heinong48) were harvested in 2009. Third, the four-way cross F_1 seeds were self-crossed for six generations continuously by alternate sowing in Yacheng (17.5°N, 109.00°E), Hainan Province, China, in the winter and in Harbin in the summer from 2010 to 2014, using the single-seed descent method to select single seeds from individual plants in each generation. Finally, a total of 144 FW-RILs were obtained for this study.

Field Experiment and Phenotype Data Collection

The four parental lines and 144 FW-RILs were planted in 20 environments with different locations, years, seedling densities, fertilizers, and sowing dates. The detailed planting schedule is summarized in **Supplementary Table S1**. All plant materials in each environment were grown in three-row 5 m × 0.7 m plots in a completely randomized block design with three replications. The experimental plots were managed identically to local soybean crops. Ten plants from the middle of the plots for each line (four parents and 144 FW-RILs) were harvested and the seeds threshed separately for each of the 20 environments. The total PC of seeds (dry seeds, with water content of about 10%) was determined in three random samples from mixed seed of each line by the near-infrared analyzer (Infratec 1241, Foss, Denmark) at the Key Laboratory of Soybean Biology of the Chinese Education Ministry at the Northeast Agricultural University in China. The calibration regression technique was Partial Least Square (PLS), which involved combining spectral data with laboratory data (Kjeldahl method) to calculate seed PC, described by the percentage of seed weight. The phenotypic values given for each parental and FW-RIL used in this study were all the mean values of three repetitions.

Statistical Analyses of Phenotypic Data

Mean, standard deviation, minimum, maximum, range, skewness, kurtosis, and coefficient of variation (CV) for FW-RILs in each environment were calculated. A correlation analysis between each pair of environments was performed. Analysis of variance (ANOVA) for single environment and jointly multiple environments was conducted with "varietal effect model of genotype" and "environment + genotype + environment × genotype interaction," respectively, and the broad-sense heritability (h^2) was estimated by the following equation:

$$h^2 = \sigma_G^2/(\sigma_G^2 + \sigma_{GE}^2/e + \sigma^2/er)$$

where σ_G^2 is the genotypic variance, σ_{GE}^2 is the variance due to the genotype × environment interaction, σ^2 is the error variance, e is the number of environments, and r is the number of replications within an environment. The statistical analysis was implemented by SAS 9.2 (SAS Institute, Cary, NC, United States).

Genotyping

Juvenile leaves from parents and FW-RIL plants were collected, frozen in liquid nitrogen, and immediately ground into powder. Total genomic DNA was extracted using the CTAB method (Doyle et al., 1990) and eluted in 50 μl deionized water. The DNA concentration was determined using a UV752N spectrophotometer (Shanghai Jingke Science Instrument Co. Ltd.) and was diluted to 100 ng ± 1 ng in deionized water. SNP genotyping was performed at Beijing Boao Biotechnology Co. Ltd, using the SoySNP660K BeadChip. A total of 109,676 SNPs across 20 chromosomes remained after quality filtering; the SNP markers identified were filtered for minor allele frequency (MAF $>$ 0.05), and the maximum missing sites per SNP was $<$ 10% (Belamkar et al., 2016). Heterozygous loci were then marked as missing to obtain better estimates of marker effects, and the SNP markers were re-filtered using the same filtering criteria and used for the next analysis of population structure, kinship, and GWAS.

Analysis of Population Structure and Linkage Disequilibrium

The analysis of population structure was performed with the software STRUCTURE 2.3.4 (Pritchard et al., 2000). For each run, the number of burn-in iterations was 10,000, followed by 2000 Markov chain Monte Carlo (MCMC) replications after burn-in. The admixture and allele frequencies correlated models were considered in the analysis. Ten impended iterations were used in the STRUCTURE analysis. The hypothetical number of subpopulations (K) ranged from 1 to 10. The best K was identified according to Evanno et al. (2005) using STRUCTURE HARVESTER (Earl and Vonholdt, 2012).

TASSEL 5.0 was utilized to analyze linkage disequilibrium (LD) (Bradbury et al., 2007) by analyzing r^2 values of all pairs of SNPs located within 10 Mb physical distance, the LD decay trend was found following regression of the negative natural logarithm, and the physical distance of LD decay was estimated as the position where r^2 dropped to half of its maximum value.

Genome-Wide Association Studies

The software mrMLM.GUI (version 3.0) was used to perform the GWAS. Five multi-locus GWAS methods within mrMLM.GUI were used to identify significant QTNs, including mrMLM (Wang et al., 2016), FASTmrMLM (Tamba et al., 2017), FASTmrEMMA (Wen et al., 2018), pLARmEB (Zhang J. et al., 2017), and ISIS EM-BLASSO (Tamba et al., 2017). The critical P-value parameters for these methods at the first stage were set to 0.01 except for FASTmrEMMA, where the critical P-value was set to 0.005, and the critical LOD score was set to 3 for significant QTN at the last stage. All these five methods involved the population structure and kinship matrices in this study, and the kinship matrix was calculated with the software mrMLM.GUI 3.0.

Superior Allele Analysis

We considered the QTNs we detected from multiple environments or by multiple methods as common QTNs. Based on the effect values of each common QTN and the genotype for code 1, we could determine the superior alleles of each QTN. If the effect value of the QTN is positive, the genotype for code 1 is the superior allele; if the effect value is negative, another genotype is the superior allele. For each QTN, the superior allele percentage in the 144 FW-RILs was equal to the number of lines containing the superior allele divided by the total number of lines. For each line, the proportion of superior alleles in these QTNs was calculated as the number of superior alleles divided by the total number of QTNs. A heat map visualizing the percentage of superior alleles was obtained in the R (heatmap package) program (Mellbye and Schuster, 2014).

Identification of Potential Candidate Genes

The search for potential candidate genes based on the common QTNs detected by multiple methods or in multiple environments was performed using four steps. First, the intervals that include each common QTN were selected on the Phytozome website[1]. These intervals were determined by the rate of LD decay. Second, genes highly expressed in the form process of seed protein through the Bio-Analytic Resource for Plant Biology (BAR) website[2] were identified. Third, based on the experimental data in the Plant Expression Database (PLEXdb) website[3], the differentially expressed genes among high and low protein lines were identified from the above high expression genes. Finally, all the differentially expressed genes were put together for analysis of pathways on the Kyoto Encyclopedia of Genes and Genomes (KEGG) website[4], and potential candidate genes were identified by the result of pathway analysis.

RESULTS

Protein Content Phenotype

We measured the PC phenotypes of the parents and the 144 FW-RILs in 20 environments, which are presented in **Supplementary Table S2** and **Supplementary Figure S1**. Graphing the average value for PC of each line in 20 environments revealed that the 144 lines show extensive variation in PC (**Supplementary Figure S1**). Examination of the values (**Supplementary Table S2**) showed that the

[1] https://phytozome.jgi.doe.gov

[2] http://www.bar.utoronto.ca

[3] http://www.plexdb.org

[4] http://www.kegg.jp

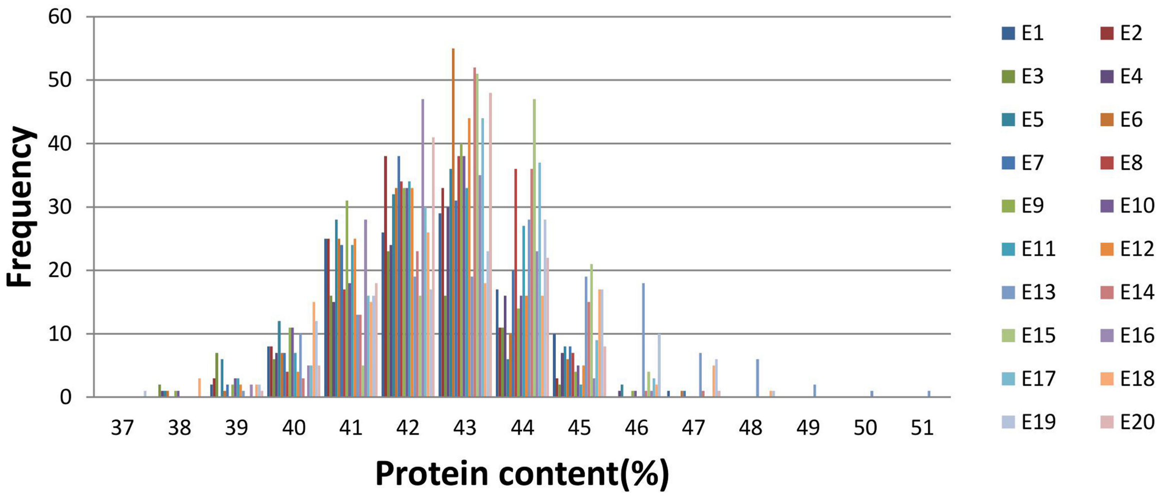


FIGURE 1 | Frequency distribution of seed protein content under 20 environments.

parental lines Kenfeng14 and Kenfeng15 had lower PC than Kenfeng19 and Heinong48 in all 20 environments. And the "Range" (Range = PC_{Max} − PC_{Min}) of the four parents in 20 environments was from 1.70 to 4.04%; the "Range" of FW-RILs was 5.20–11.56%, representing a large difference in PC, especially in FW-RILs. Kurtosis and skewness (absolute value) were less than 1, indicating the continuous normal distribution of the PC values (**Supplementary Table S2** and **Figure 1**). We found a high coefficient of variation in PC across all the environments. The ANOVA results of parents and FW-RILs both indicated that extremely significant variation exists in genotype, environment, and genotype-by-environment (**Tables 1**, **2**). The mean square values for the genotype-by-environment interaction were all less than the mean square values of genotype, and the estimated broad-based heritability was high, being 85.46%. The correlation coefficients between each pair of environments were almost all positive, and many were significant or extremely significant (**Supplementary Table S3**), indicating high consistency across various environments.

Population Structure and LD

To define the subpopulations within the panel of 144 lines, as described by Pritchard et al. (2009), we selected 5375 of the 109,676 SNPs that had better polymorphisms and were randomly distributed across the 20 soybean chromosomes. Delta *K* (Δ*K*) was calculated using STRUCTURE 2.3.4 (**Figure 2A**; *K* = 1–10), revealing the presence of two subpopulations (selected *K* = 2) based on ΔK values (**Figure 2B**). These two subgroups contained 53 (36.81%) and 91 (63.19%) lines.

TABLE 1 | Joint ANOVA of PC of parent lines in multiple environments.

Source	DF	SS	MS	*F*	Pr > F
Replication	2	5.55	2.77	6.59	0.0018
Environment	19	155.86	8.20	19.49	< 0.0001
Genotype	3	271.22	90.41	214.80	< 0.0001
Genotype * Environment	57	50.02	0.877	2.08	0.0002
Error	158	66.50	0.42		

We analyzed the r^2 values of all pairs of SNPs located within 10 Mb of each other and determined the LD decay trend based on regression to the negative natural logarithm. As shown in **Supplementary Figure S2**, r^2 decreased gradually with increased distance, and the LD decay distance was estimated at 1.2 Mb, where r^2 dropped to half of its maximum value (0.45). Because the population used in this study is derived from parents, the speed of LD decay is slower and the LD decay distance is much longer than that of a natural population.

QTNs Detected by Multi-Locus GWAS Methods

We identified 19, 18, 12, 37, and 43 significant QTNs for PC using mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, and ISIS EM-BLASSO, respectively, and 10, 11, 10, 2, 1, 2, 3, 6, 11, 10, 0, 8, 4, 9, 9, 7, 1, 5, 12, and 8 significant QTNs, respectively, in the 20 environments (**Figure 3** and **Supplementary Table S4**). No significant QTN was detected in the eleventh environment (E11).

We further checked the common QTNs across multiple environments. As discussed above, only one such common QTN was identified in three environments (**Table 3**). The single QTN (AX-157298785) was located on chromosome 18, with the LOD values ranging from 4.02 to 5.33 (**Table 3**). The proportion of phenotypic variance explained (PVE) by the QTN ranged from 8.40 to 11.02%, and the QTN direction of effect (positive or negative) was consistent across different environments and different methods (**Table 3**).

Comparing the results across the different approaches, we found that 22 common QTNs (including AX-157298785) were identified simultaneously by at least two approaches (**Table 4**); these were located on chromosomes 1, 2, 3, 4, 6, 7, 9, 10,

TABLE 2 | Joint ANOVA of PC of FW-RILs in multiple environments and heritability.

Source	DF	SS	MS	*F*	Pr > F	Variance component
Replication	2	1, 304.62	652.31	11, 639.60	< 0.0001	0.25
Environment	19	1, 893.98	99.681	1, 778.72	< 0.0001	0.23
Genotype	143	4, 565.61	31.93	569.70	< 0.0001	0.50
Genotype * Environment	2,436	12, 408.82	5.09	90.89	< 0.0001	1.68
Error	5,196	291.19	0.06			0.06
h^2						0.85

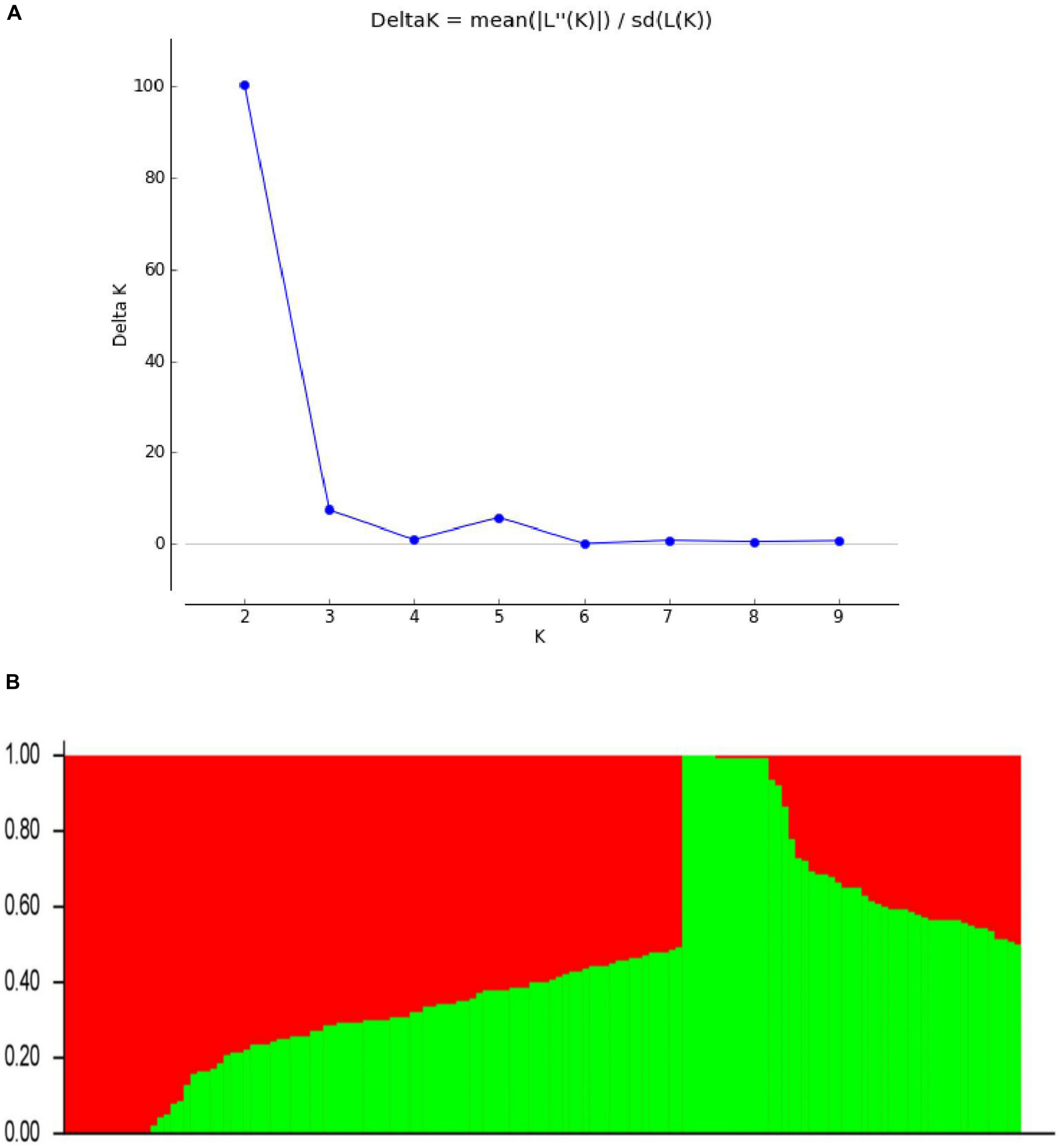


FIGURE 2 | Population structure based on 5375 SNPs distributed across 20 chromosomes. **(A)** Plot of ΔK calculated for *K* = 1–10. **(B)** Population structure (*K* = 2); the areas of the two colors (green and red) illustrate the proportion of each subgroup.

12, 14, 16, and 18. Their LOD values ranged from 3.06 to 6.90, the proportion of PVE by each QTN ranged from 3.84 to 19.21%, and the direction of effect (positive or negative) of each QTN was also consistent across the different methods (**Table 4**). Of the 22 common QTNs, 12, 5, and 5 were identified simultaneously by 2, 3, and 4 approaches, respectively.

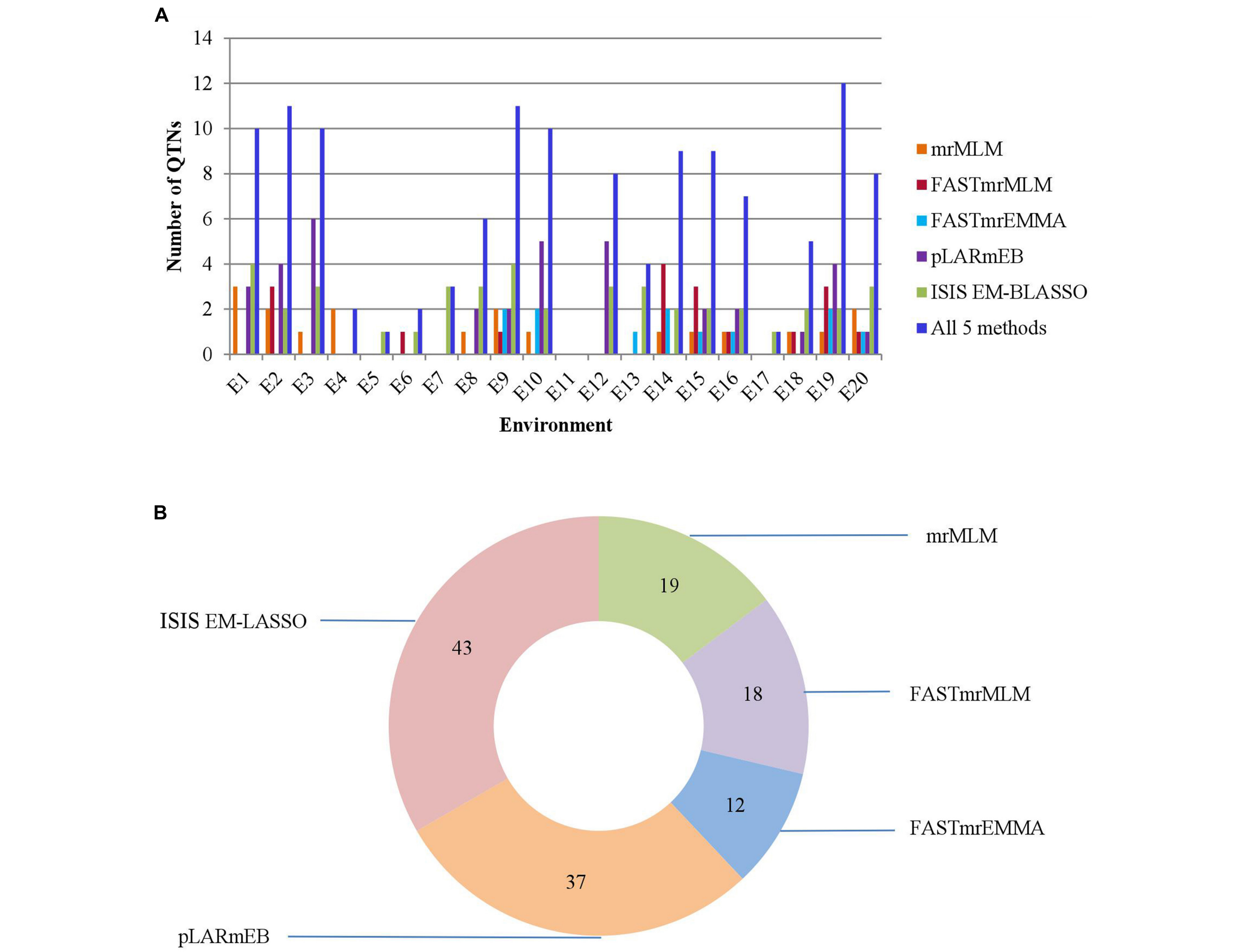


FIGURE 3 | (A) The total numbers of significant QTNs detected in 20 environments across 5 methods. **(B)** The total numbers of significant QTNs detected using each of 5 multi-locus GWAS methods in 20 environments.

TABLE 3 | Stable expressed QTNs identified in multiple environments and by multiple methods.

Method[a]	Env	Marker	Chr	Marker position	QTN effect	LOD score	r^2(%)[b]
1,1	E14,E20,	AX-157298785	18	6,620,851	−0.39, −0.40,	4.21, 4.02,	10.78, 11.02,
2,3	E14,E14,				−0.29, −0.67,	4.32, 4.43,	5.86, 7.78,
5,5	E14,E16				−0.34, −0.39	4.56, 5.33	8.40, 10.48

[a] *mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, and ISIS EM-BLASSO were indicated by 1 to 5, respectively.* [b] *r^2 (%), proportion of total phenotypic variation explained by each QTN.*

Among the five methods, ISIS EM-BLASSO detected the highest number of common QTNs (**Figure 4A**), and among the combinations of two methods, FASTmrMLM combined with pLARmEB detected the highest number of common QTNs (**Figure 4B**).

We found only one stable QTN that was identified not only in multiple environments but also by multiple methods (**Table 3**): AX-157298785, located on chromosome 18, which was detected by mrMLM, FASTmrMLM, FASTmrEMMA, and ISIS EM-BLASSO in environments E14, E16, and E20, with LOD values ranging from 4.02 to 5.33 and PVE values ranging from 5.86 to 11.02% (**Table 3**).

Distribution of Superior Alleles in the FW-RILs

Based on the PC averages in 20 environments for each FW-RIL, we found that 22 lines had higher phenotypic values (43.07–44.21%) and 7 lines had lower phenotypic values (40.60–40.98%) (**Table 5**). For each of the 22 elite lines, the percentages of superior alleles (PSA) across 22 common

TABLE 4 | Common QTNs for seed protein content in soybean across different multi-methods.

Method[a]	Marker	Chr	Position (bp)	QTN effect	LOD score	r^2 (%)[b]
2,3	AX-157088197	1	2,142,538	0.24,0.51	3.06,3.15	5.41,5.50
2,5	AX-157514742	1	5,244,469	0.41,0.44	4.76,4.76	8.58,9.99
2,4	AX-157393800	1	36,630,129	0.53,0.61	4.17,5.49	6.39,8.57
2,3,4	AX-157197609	2	19,981,350	−0.34,−0.65,−0.38	3.38,3.36,3.81	8.06,6.14,9.58
3,5	**AX-157074676**	**2**	**43,036,996**	**−0.64,−0.30**	**3.65,3.10**	**5.59,4.83**
2,4	AX-157487767	3	28,963,194	0.45,0.47	4.72,3.90	4.74,5.30
2,3,4,5	**AX-157594705**	**4**	**47,793,555**	**−0.35,−0.65,−0.42,−0.35**	**3.94,3.24,5.26,4.40**	**6.97,5.61,9.85,6.96**
1,2,4,5	**AX-157397239**	**4**	**47,801,472**	**−0.32,−0.27,−0.23,−0.25**	**3.13,3.07,3.43,3.73**	**9.42,6.54,4.75,5.88**
1,2,4,5	**AX-157489326**	**6**	**48,361,864**	**0.43,0.34,0.42,0.31**	**3.78,4.26,6.53,3.37**	**11.81,7.50,11.14,6.19**
1,4	AX-157083233	6	49,396,770	−0.78,−0.51	3.20,3.65	19.21,6.67
1,3,4,5	**AX-157462104**	**7**	**20,724,011**	**0.82,1.35,0.65,0.50**	**4.64,4.97,5.90,3.21**	**15.18,9.83,9.58,5.78**
1,2,5	**AX-157506141**	**9**	**34,120,396**	**−0.56,−0.44,−0.41**	**4.02,4.03,3.63**	**12.36,7.81,6.79**
3,5	**AX-157570733**	**10**	**43,785,659**	**0.75,0.34**	**5.09,3.88**	**9.27,7.63**
4,5	**AX-157566978**	**12**	**1,258,280**	**−0.24,−0.36**	**3.65,4.52**	**3.84,8.50**
3,5	**AX-157069070**	**12**	**10,655,900**	**0.83,0.42**	**4.24,5.50**	**8.30,8.67**
4,5	**AX-157357710**	**12**	**11,111,913**	**−0.61,−0.62**	**6.90,5.32**	**9.83,12.74**
2,4,5	**AX-157217990**	**14**	**7,160,557**	**−0.23,−0.24,−0.23**	**3.42,3.22,3.63**	**4.79,5.45,4.92**
1,4	AX-157512649	16	24,057,874	0.31,0.24	3.16,3.44	7.00,4.09
1,5	**AX-157168337**	**16**	**28,693,806**	**0.64,0.37**	**4.66,3.94**	**14.20,5.23**
1,3,5	AX-157333937	18	2,064,407	0.56,0.96,0.48	3.56,4.33,4.23	11.90,8.54,8.58
1,2,4	AX-157443296	18	6,597,875	−0.49,−0.36,−0.38	3.83,4.26,5.12	15.96,8.71,9.88
1,2,3,5	AX-157298785	18	6,620,851	−0.39(−0.40),−0.29, −0.67,−0.34 (−0.39)	4.21 (4.02),4.32, 4.43,4.56,(5.33)	10.78(11.02),5.86, 7.78,8.40 (10.48)

[a] mrMLM, FASTmrMLM, FASTmrEMMA, pLARmEB, and ISIS EM-BLASSO were indicated by 1–5, respectively. [b] r^2 (%), proportion of total phenotypic variation explained by each QTN. Bold text indicates the QTNs appeared to be near QTLs associated with protein content in soybean that had been mapped in earlier studies.

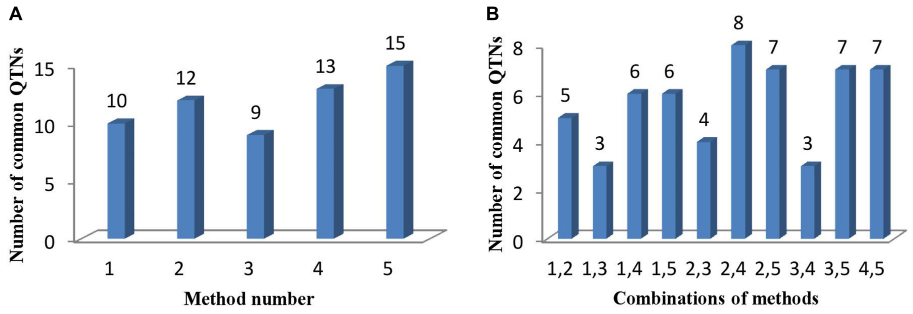


FIGURE 4 | (A) The number of common QTNs detected by different methods and **(B)** different combinations of methods. Method numbers correspond to (1) mrMLM, (2) FASTmrMLM, (3) FASTmrEMMA, (4) pLARmEB, and (5) ISIS EM-BLASSO.

QTNs ranged from 36 to 82% (**Table 5**), 91% (20 of the 22 lines) showed PSAs of ≥50%, and only 9% showed PSAs of <50%. For each of the 7 lines with lower phenotypic values, the PSAs ranged from 32 to 50% (**Table 5**), only 2 lines (28%) had PSAs of ≥50%, and the remaining 5 (71%) had PSAs of <50%. Thus, the elite lines with higher PC have more superior alleles than the lines with lower PC (**Figure 5**).

Based on the superior allele information for the 22 common QTNs in 29 lines, the PSAs for each QTN ranged from 36 to 95% in the 22 elite lines, with 16 QTNs showing ≥ 50% superior alleles while the remaining 6 QTNs showed < 50%. The range of PSAs for each QTN was 0–71% in the 7 lines with lower phenotypic values; 8 of the QTNs had PSAs ≥ 50% and the remaining 14 QTNs had PSAs < 50% (**Table 6** and **Figure 6**). The number of QTNs with ≥50% superior alleles was higher in the 22 elite lines than in the 7 inferior lines. Based on these results, we can easily find elite lines by identifying superior alleles for application in breeding higher PC soybean.

In addition, we found some common superior alleles in multiple elite lines: for example, the seven lines HN54, HN37, HN46, HN47, HN40, HN45, and HN103 all contained the superior alleles AX-157506141, AX-157168337, AX-157514742, AX-157397239, AX-157570733, and AX-15719760 (**Figure 6**), and the superior allele AX-157506141 occurred in 21 of the 22 elite lines. We suspect that common superior alleles may have a

TABLE 5 | Phenotypic averages of seed protein content and proportion of superior alleles in 29 lines across 22 common QTNs.

Line	PC (%)	PSA (%)	Line	PC (%)	PSA (%)
HN54	44.21	82	HN67	44.19	55
HN37	43.26	77	HN98	43.01	55
HN46	44.01	77	HN41	43.10	50
HN47	43.27	77	HN48	43.55	50
HN40	43.40	73	HN58	43.13	50
HN45	43.54	73	HN93	43.03	45
HN103	44.20	68	HN20	43.34	36
HN24	43.07	64	**HN2**	**40.69**	**50**
HN74	43.42	64	**HN112**	**40.60**	**50**
HN118	43.11	64	**HN65**	**40.98**	**45**
HN142	43.14	64	**HN17**	**40.66**	**36**
HN60	43.06	59	**HN32**	**40.81**	**36**
HN69	43.43	59	**HN75**	**40.60**	**36**
HN91	43.20	59	**HN106**	**40.71**	**32**
HN35	43.43	55			

Bold font indicates the 7 lines with lower protein contents, and non-bold font indicates the 22 lines with higher protein contents.

particularly strong influence on PC. In further research, we hope to make use of this information to breed better soybean using marker-assisted selection.

Potential Candidate Genes Determined Based on Common QTNs

We used the LD decay distance to select potential candidate genes within a specific distance of each common QTN. Because of the nature of the population (i.e., derived from parents), the LD decay distance is large, so we determined the range of potential candidate genes according to the position of the fastest decay rate. In **Supplementary Figure S2B**, we can see that LD decays fastest before 200 kb, and then tends to flatten, so we searched for potential candidate genes in the interval of 100 kb on either side of each QTN. Following the four steps described in the Materials and Methods, a total of 288 genes were found in these intervals and 96 genes were expressed highly in seed at the form process of seed protein. Among the 96 genes, 34 genes were differentially expressed among high and low protein lines, and these 34 genes were used to do pathway analysis.

From the annotation data, we found that 17 of 34 genes (51.4% of the genes we submitted) were previously annotated in 14 pathways and 3 protein families in the KEGG database (**Supplementary Tables S5**, **S6** and **Figure 7**). Of these, 8 were considered potential candidate genes based on the information of their annotation and functions in metabolic pathways (**Table 7** bold text).

DISCUSSION

In this study, we employed multi-locus GWAS with an FW-RIL population of the MAGIC population type to identify QTNs related to PC of soybean. Twenty-two common QTNs were detected by multiple methods or in multiple environments (**Tables 3**, **4**). Based on the SoyBase database and the results of recent studies, 12 of the 22 common QTNs appeared to be near QTLs associated with PC in soybean that had been mapped in earlier studies (Lee et al., 1996; Brummer et al.,

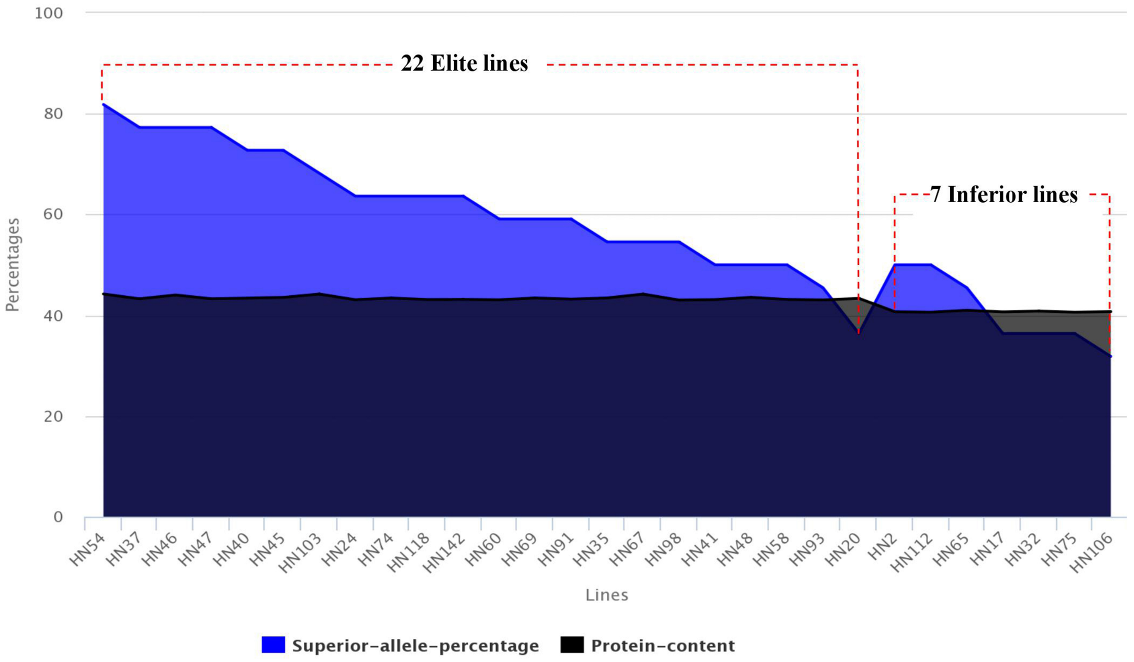


FIGURE 5 | Distribution of superior allele percentages and the PC in the 29 high- and low-PC lines.

TABLE 6 | Superior alleles and their proportions of 22 common QTNs in 22 elite and 7 inferior lines.

QTN	Superior allele	PSA (%)^a	PSA (%)^b	QTN	Superior allele	PSA (%)^a	PSA (%)^b
AX-157506141	CC	95	71	AX-157074676	GG	64	71
AX-157514742	CC	82	29	AX-157489326	TT	59	57
AX-157168337	TT	77	57	AX-157594705	GG	59	29
AX-157393800	AA	77	43	AX-157487767	AA	50	57
AX-157462104	GG	77	43	AX-157570733	TT	50	29
AX-157298785	GG	73	43	AX-157512649	GG	45	71
AX-157397239	AA	73	29	AX-157197609	AA	45	14
AX-157443296	CC	68	57	AX-157069070	TT	41	57
AX-157088197	GG	68	43	AX-157333937	CC	41	14
AX-157566978	AA	68	43	AX-157083233	TT	36	29
AX-157217990	TT	68	14	AX-157357710	AA	36	0

^{a}*Indicates the percentage of superior allele of each common QTN in 22 elite lines.* ^{b}*Indicates the percentage of superior allele of each common QTN in 7 inferior lines.*

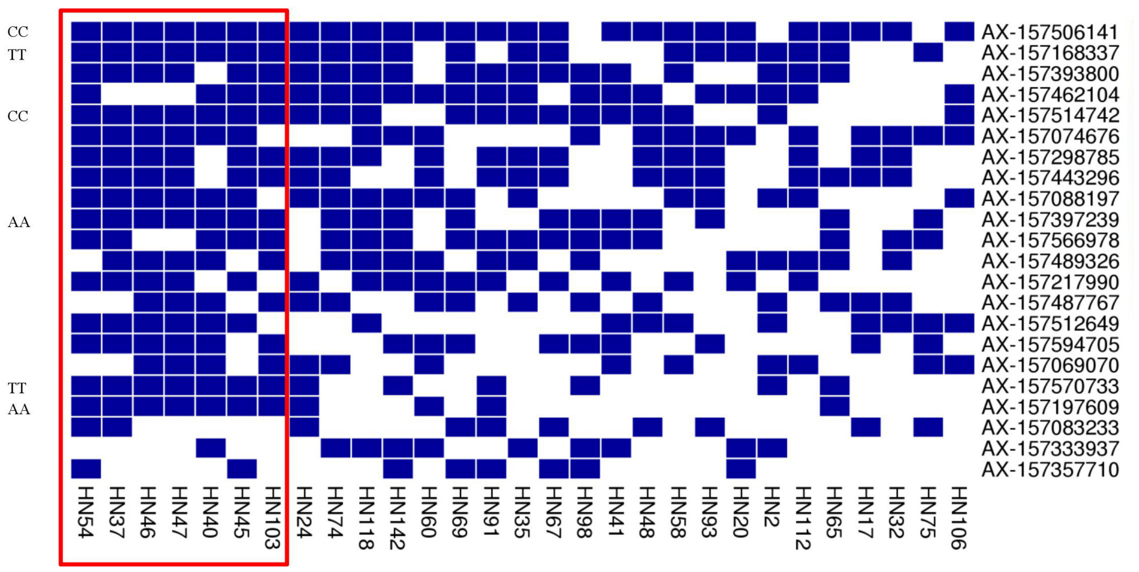


FIGURE 6 | Heat map of the superior allele distribution for the 22 common QTNs in the 29 high- and low-PC lines. Blue and white colors represent superior and inferior alleles, respectively.

1997; Csanádi et al., 2001; Jun et al., 2008; Lu et al., 2013; Mao et al., 2013; Pathan et al., 2013; Hwang et al., 2014; Qi et al., 2014; Vaughn et al., 2014; Warrington et al., 2015): AX-157074676, AX-157594705, AX-157397239, AX-157489326, AX-157462104, AX-157506141, AX-157069070, AX-157357710, AX-157217990, AX-157570733, AX-157566978, and AX-157168337 (**Table 4**, bold text). This indirectly confirmed the accuracy of our QTN detection. The remaining ten QTNs in this study were new (**Table 4**, non-bold text). For this population, the detection of significant QTNs is not only a way to identify the genes related to PC, but can also identify good lines based on the superior allele information to support breeding high PC soybean.

Based on the 22 common QTNs detected here and their pathway annotation, we have identified 8 genes that may be related to protein anabolism (**Table 7**, bold text). *Glyma.03G100800* is intimately involved in the biosynthesis of amino acids, and the pentose phosphate pathway which it is involved in also indirectly affects the biosynthesis of proteins (Xu et al., 2018). *Glyma.10G207300*, *Glyma.14G081600,* and *Glyma.12G019300* are mainly involved in the proteasome pathway and work as protease to degrade proteins to small peptides and amino acids, so we think that these three genes are closely related to protein degradation. *Glyma.18G071100* may play an important role in protein anabolism; it participates in the process of glycosylation in endoplasmic reticulum, and the function of glycosylation is to enable proteins to resist the digestive enzymes, so as to protect protein from degradation (Jayaprakash and Surolia, 2017). *Glyma.12G112900* is a kind of riboflavin synthase and it participates in the biosynthesis of riboflavin, which plays an important role in energy metabolism including carbohydrate, protein and fat metabolism (Tuan et al., 2014; Zhao et al., 2014). *Glyma.18G071300* participates in RNA transport, and this

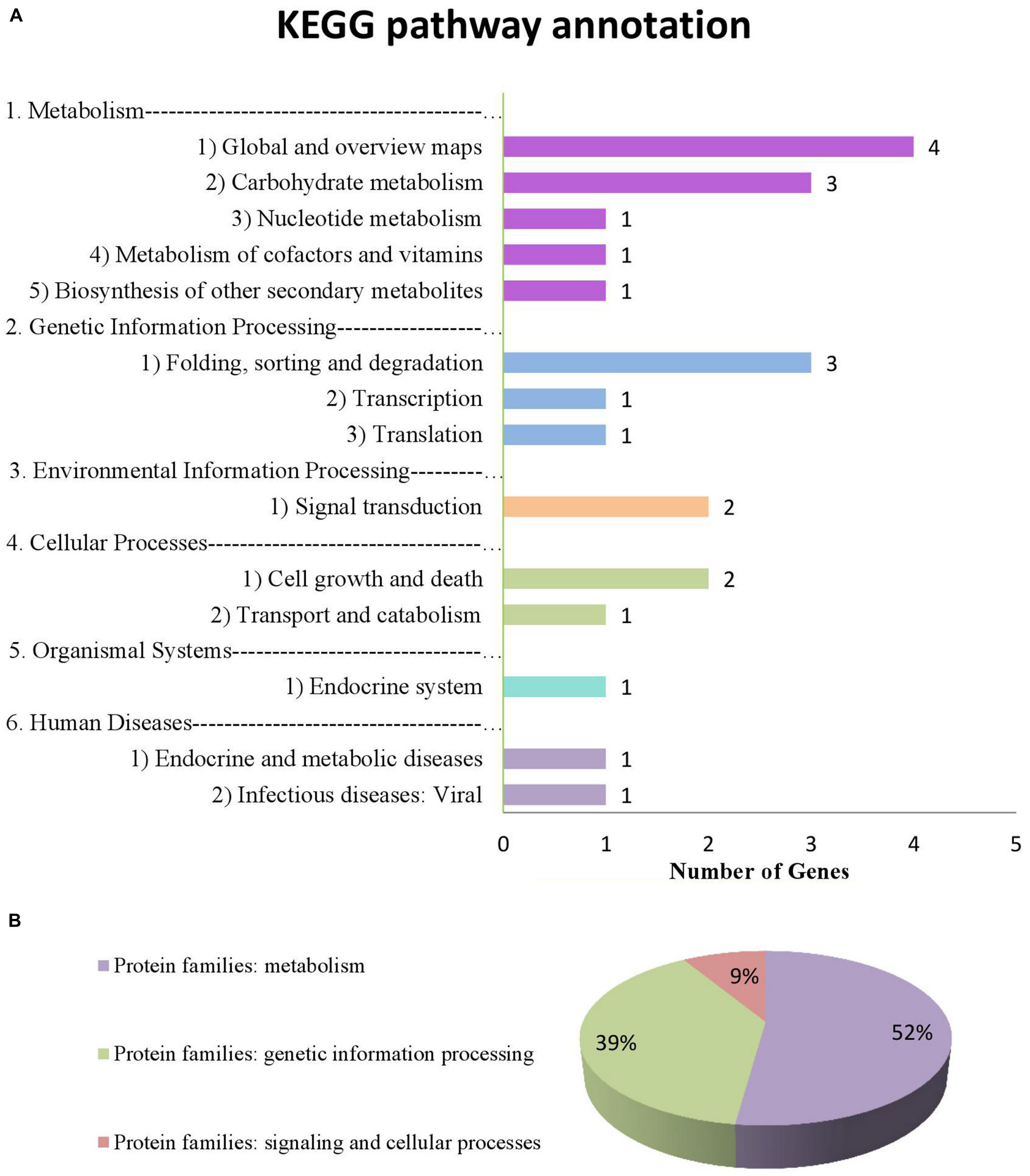


FIGURE 7 | Information on pathways and orthologous protein families of 17 genes. **(A)** shows the information on pathway. **(B)** shows the information on orthologous protein families.

process is essential in protein synthesis; as we know, it mainly carries amino acids into ribosomes and synthesizes proteins under the guidance of mRNA. *Glyma.18G028600* is a kind of translocation protein, and it mainly works in post-translational transport and post-translational modification in the process of protein synthesis, so it plays an important role in protein synthesis (Kwon et al., 1999). Based on these 8 genes, further work will be needed to determine which of these actually significantly affect PC in soybeans, and then to identify the target genes. This information can be used as the basis for further exploration of the gene network for the trait.

In recent years, GWAS has been widely applied to crop plants such as rice (Huang et al., 2010; Ma et al., 2016), maize (Tian et al., 2011), and soybean (Li et al., 2018), and the model mainly used for GWAS is mixed linear model (MLM). It belongs to single-locus GWAS, for which the screening criterion for significance is $P = 0.05/m$ (where m is the number

TABLE 7 | Details of 17 genes annotated in the KEGG database.

QTN name	Gene name^{a}	Chromosome	Position	KO number	Annotation
AX-157088197	Glyma.01G020900	chr01	2104937..2109784	K00083	E1.1.1.195; cinnamyl-alcohol dehydrogenase [EC:1.1.1.195]
AX-157514742	Glyma.01G046000	chr01	5317206..5321170	K21734	SLD; sphingolipid 8-(E/Z)-desaturase [EC:1.14.19.29]
AX-157487767	**Glyma.03G100800**	**chr03**	**28980378..28988564**	**K00948**	**PRPS; ribose-phosphate pyrophosphokinase [EC:2.7.6.1]**
AX-157397239	Glyma.04G206300	chr04	47880082..47890343	K21594	GUF1; translation factor GUF1, mitochondrial [EC:3.6.5.-]
AX-157570733	Glyma.10G207100	chr10	43860713..43863694	K01373	CTSF; cathepsin F [EC:3.4.22.41]
AX-157570733	**Glyma.10G207300**	**chr10**	**43879369..43883196**	**K16298**	**SCPL-IV; serine carboxypeptidase-like clade IV [EC:3.4.16.-]**
AX-157566978	Glyma.12G017200	chr12	1212436..1217081	K17790	TIM22; mitochondrial import inner membrane translocase subunit TIM22
AX-157566978	Glyma.12G018000	chr12	1252709..1256330	K04683	TFDP1; transcription factor Dp-1
AX-157566978	Glyma.12G018800	chr12	1321415..1326705	K1542	PPP4R2; serine/threonine-protein phosphatase 4 regulatory subunit 2
AX-157566978	**Glyma.12G019300**	**chr12**	**1354109..1356766**	**K11599**	**POMP; proteasome maturation protein**
AX-157357710	**Glyma.12G112900**	**chr12**	**11064160..11065437**	**K00793**	**ribE; riboflavin synthase [EC:2.5.1.9]**
AX-157357710	Glyma.12G113400	chr12	11135703..11140450	K19355	MAN; mannan endo-1,4-beta-mannosidase [EC:3.2.1.78]
AX-157217990	**Glyma.14G081600**	**chr14**	**7064342..7068643**	**K03030**	**PSMD14; 26S proteasome regulatory subunit N11**
AX-157333937	Glyma.18G027100	chr18	2033839..2038697	K05857	PLCD; phosphatidylinositol phospholipase C, delta [EC:3.1.4.11]
AX-157333937	**Glyma.18G028600**	**chr18**	**2154790..2160008**	**K00685**	**ATE1; arginyl-tRNA—protein transferase [EC:2.3.2.8]**
AX-157443296	**Glyma.18G071100**	**chr18**	**6667467..6671661**	**K12275**	**SEC62; translocation protein SEC62**
AX-157443296	**Glyma.18G071300**	**chr18**	**6687583..6692093**	**K12880**	**THOC3; THO complex subunit 3**

Bold font indicates the genes which correlate with the protein anabolism in soybean according to our deduction. ^{a}Indicates the gene which correlates with the QTN (before the gene in the same row).

of markers) (Perneger, 1998). For a large number of SNPs, some important loci may be undetectable under this screening criterion.

In this study, we also used MLM to carry out the GWAS of soybean PC, with the calculation of population structure based on STRUCTURE 2.3.4 and the calculation of kinship and GWAS based on TASSEL 5.0. However, we did not detect any significant SNPs using the model. We believe that this was related to the screening criteria of single-locus GWAS and to the type of target trait as well as the population type. In light of this negative result, we tried changing the screening criteria, and replaced the strict Bonferroni correction with a less stringent false discovery rate (FDR) correction. The q-values were equal to the P-values adjusted with the Benjamini and Hochberg (2000) procedure; we used the cut-off of $q < 0.05$ and $P < 0.0001$as the threshold value to identify significant QTNs. Based on these screening criteria, we still did not find significantly associated SNP markers. To further reduce the stringency of the screening criteria, we next tried using a P-value of < 0.0001 directly as a cut-off without any corrections. With this criterion, a total of 15 SNP markers were identified with 1 cluster (bold text in **Table 8**). The range for the one cluster was 83.40 kb.

Because of the nature of single-locus methods, even if we identified these significant SNP markers under less stringent screening methods, we lack confidence about our ability to control for the false positive rate of results obtained without the correction for multiple tests. A previous study yielded a similar outcome to ours (Fang et al., 2017), so it seems evident that the single-locus method is not always suitable for detecting the genetic basis of complex traits.

TABLE 8 | Details of significant SNPs detected by MLM with screening critical $P < 0.0001$.

Environment	Chr	Physical position (bp)	Physical distance (kb)	*P*-value	No. of QTNs (MLM)^{a}	No. of QTNs (multi-locis GWAS)^{b}
E6	1	5,703,406		2.82E-05	1	1
E19	8	43,712,243		1.17E-04	1	1
E4	13	1,147,825		9.80E-05	1	0
E12	14	15,013,495		1.35E-04	1	0
E12	16	5,609,879		1.12E-04	1	0
E19	16	28,008,354		1.07E-04	1	2
E6	18	5,396,919		3.54E-05	1	0
E16	**18**	**6,590,065 - 6,673,462**	**83.40**	**6.01E-05 - 1.35E-04**	**8**	**11**
Total					**15**	**15**

^{a}*Number of QTNs detected by MLM model.* ^{b}*Number of QTNs detected by multi-locus GWAS methods.*

To make up for the shortcomings of the above methods, multi-locus GWAS methods have recently been explored, including the five methods we used in this study: mrMLM (Wang et al., 2016), FASTmrMLM (Tamba et al., 2017), FASTmrEMMA (Wen et al., 2018), pLARmEB (Zhang J. et al., 2017), and ISIS EM-BLASSO (Tamba et al., 2017). Using the five methods, we detected a total of 19, 18, 12, 37, and 43 significant QTNs, respectively (**Figure 3B** and **Supplementary Table S4**). The differences in the numbers of QTNs detected by the five methods are presumed to be due to the different principles underlying the different methods: even though all five are two-stage combined approaches, they differ in the models and methods for screening and estimation. Because the main purpose of this study required the most accurate QTNs possible, we felt it desirable to take into consideration the results of all five methods, and took QTNs detected by multiple methods as the credible QTNs to use in further experiments. This practice adds an extra screening to the multi-locus GWAS approach and thus makes us more confident about the results.

Based on the data from the QTNs we detected, we found that the absolute values of QTN effects were all relatively low, in the range of 0.17–1.35, which indirectly explained why we could not detect significant QTNs by MLM with the standard Bonferroni correction. Thus, the greatest advantage of the multi-locus GWAS approach was its ability to find loci with relatively small effects. In addition, the multi-locus GWAS method was more suitable for the FW-RIL population used in this study than single-locus GWAS. This is because the single-locus GWAS method is generally based on SNPs for which there are only two alleles at one locus, meaning that the multi-allelic variation that exists at some genetic loci in the FW-RIL population cannot be detected. However, the multi-locus GWAS method overcomes this limitation: because it is based on a multi-locus and multi-allele model, it can identify genome-wide QTNs in a more comprehensive fashion along with the multiple alleles. This is the other reason that we were able to detect significant QTNs with the multi-locus GWAS method.

SUMMARY

Combining five multi-locus GWAS methods, we identified 22 common QTNs, including one stable expression QTN AX-157298785. Around these QTNs, 8 potential candidate genes were identified. Moreover, we selected elite lines for breeding higher seed protein content.

AUTHOR CONTRIBUTIONS

W-XL and HN conceived and designed the experiments. WL, ShiL, SX, JiZ, KZ, XL, YF, YW, JS, ZQ, and XT made major contributions to the field experiments and determination of quality traits. ShuL and ZT performed the genome sequencing. KZ, HN, and JuZ analyzed and interpreted the results. KZ and HN drafted the manuscript. All the authors contributed to the manuscript revision.

FUNDING

The authors gratefully acknowledge the financial support for this study provided by grants from the National Key Research and Development Program of China (2017YFD0101303-6) to W-XL and Project of Research and Development on Applied Technology of Harbin in Heilongjiang Province (2017RAXXJ019) to HN.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01690/full#supplementary-material

FIGURE S1 | The mean values of seed protein content for 144 lines and 4 parents across 20 environments.

FIGURE S2 | Linkage disequilibrium (LD) decays in the four-way recombinant line (FW-RIL) population. **(A)** The LD decay rate was estimated as the squared correlation coefficient (r^2) using all pairs of SNPs located within 10 Mb of physical distance. The dashed line in red indicates the position where r^2 dropped to half of its maximum value, and the dashed line in green indicates the position where r^2 dropped fast and then tended to flatten. **(B)** Enlarged display of the area in the purple frame in **(A)**.

TABLE S1 | Summary of details of planting conditions in field experiments.

TABLE S2 | Statistical characteristics of protein content in the parents and the FW-RIL populations grown in twenty environments.

TABLE S3 | The correlation analysis between each pair of environment.

TABLE S4 | Significant QTNs for seed protein content across 20 environments using the five multi-locus GWAS methods.

TABLE S5 | The information of pathways of 17 genes annotated in the KEGG database.

TABLE S6 | The information of orthologous protein families of 17 genes annotated in the KEGG database.

REFERENCES

Bandillo, N., Raghavan, C., Muyco, P. A., Sevilla, M. A. L., Lobina, I. T., Dilla-Ermita, C. J., et al. (2013). Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice* 6:11. doi: 10.1186/1939-8433-6-11

Belamkar, V., Farmer, A. D., Weeks, N. T., Kalberer, S. R., Blackmon, W. J., and Cannon, S. B. (2016). Genomics-assisted characterization of a breeding collection of Apios americana, an edible tuberous legume. *Sci. Rep.* 6:34908. doi: 10.1038/srep34908

Benjamini, Y., and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* 25, 60–83. doi: 10.2307/1165312

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. doi: 10.1093/bioinformatics/btm308

Brummer, E. C., Graef, G. L., Orf, J., Wilcox, J. R., and Shoemaker, R. C. (1997). Mapping QTL for seed protein and oil content in eight soybean populations. *Crop Sci.* 37, 370–378. doi: 10.1007/BF00224058

Csanádi, G., Vollmann, J., Stift, G., and Lelley, T. (2001). Seed quality QTLs identified in a molecular map of early maturing soybean. *Theor. Appl. Genet.* 103, 912–919. doi: 10.1007/s001220100621

Dell'Acqua, M., Gatti, D. M., Pea, G., Cattonaro, F., Coppens, F., Magris, G., et al. (2015). Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biol.* 16:167. doi: 10.1186/s13059-015-0716-z

Doyle, J. J., Doyle, J. L., and Brown, A. H. D. (1990). Analysis of a polyploid complex in Glycine with chloroplast and nuclear DNA. *Aust. Syst. Bot.* 3, 125–136. doi: 10.1071/SB9900125

Earl, D. A., and vonholdt, B. M. (2012) Structure harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 18:161. doi: 10.1186/s13059-017-1289-9

Gaur, R., Azam, S., Jeena, G., Khan, A. W., Choudhary, S., Jain, M., et al. (2012). High-throughput SNP discovery and genotyping for constructing a saturated linkage map of chickpea (*Cicer arietinum* L.). *DNA Res.* 19, 357–373. doi: 10.1093/dnares/dss018

Huang, B. E., George, A. W., Forrest, K. L., Kilian, A., Hayden, M. J., Morell, M. K., et al. (2012). A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnol. J.* 10, 826–839. doi: 10.1111/j.1467-7652.2012.00702.x

Huang, X., Wei, X., Sang, T., Zhao, Q. A., Feng, Q., Zhao, Y., et al. (2010). Genomewide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–976. doi: 10.1038/ng.695

Hwang, E., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., et al. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15:1. doi: 10.1186/1471-2164-15-1

Jayaprakash, N. G., and Surolia, A. (2017). Role of glycosylation in nucleating protein folding and stability. *Biochem. J.* 474, 2333–2347. doi: 10.1042/BCJ20170111

Jun, T., Van, K., Kim, M. Y., Lee, S., and Walker, D. R. (2008). Association analysis using SSR markers to find QTL for seed protein content in soybean. *Euphytica* 162, 179–191. doi: 10.1007/s10681-007-9491-6

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., et al. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet.* 5:e1000551. doi: 10.1371/journal.pgen.1000551

Kwon, Y. T., Kashina, A. S., and Varshavsky, A. (1999). Alternative splicing results in differential expression, activity, and localization of the two forms of arginyl-tRNA-protein transferase, a component of the N-end rule pathway. *Mol. Cell. Biol.* 19, 182–193. doi: 10.1128/MCB.19.1.182

Lee, S. H., Bailey, M. A., Mian, M. A., Carter, T. E. Jr., Shipe, E. R., Ashley, D. A., et al. (1996). RFLP loci associated with soybean seed protein and oil content across populations and locations. *Theor. Appl. Genet.* 93, 649–657. doi: 10.1007/BF00224058

Li, J., Li, H., Li, Y., Li, Y., Li, W., Reif, J. C., et al. (2018). Genome-wide association mapping of QTL underlying seed oil and protein contents of a diverse panel of soybean accessions. *Plant Sci.* 266, 95–101. doi: 10.1016/j.plantsci.2017.04.013

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12:e1005767. doi: 10.1371/journal.pgen.1005767

Lu, W., Wen, Z., Li, H., Yuan, D., Li, J., Zhang, H., et al. (2013). Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. *Theor. Appl. Genet.* 126, 425–433. doi: 10.1007/s00122-012-1990-8

Ma, X., Feng, F., Wei, H., Mei, H., Xu, K., Chen, S., et al. (2016). Genome-wide association study for plant height and grain yield in rice under contrasting moisture regimes. *Front. Plant Sci.* 7:1801. doi: 10.3389/fpls.2016.01801

Mackay, I. J., Bansept-Basler, P., Barber, T., Bentley, A. R., Cockram, J., Gosman, N., et al. (2014). An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: creation, properties, and validation. *G3* 4, 1603–1610. doi: 10.1534/g3.114.012963

Mao, T., Jiang, Z., Han, Y., Teng, W., Zhao, X., Li, W., et al. (2013). Identification of quantitative trait loci underlying seed protein and oil contents of soybean across multi-genetic backgrounds and environments. *Plant Breed.* 132, 630–641. doi: 10.1111/pbr.12091

Mellbye, B., and Schuster, M. (2014). Physiological framework for the regulation of quorum sensing-dependent public goods in *Pseudomonas aeruginosa*. *J. Bacteriol.* 196, 1155–1164. doi: 10.1128/JB.01223-13

Meng, L., Guo, L., Ponce, K., Zhao, X., and Ye, G. (2016a). Characterization of three rice multiparent advanced generation intercross (MAGIC) populations for quantitative trait loci identification. *Plant Genome* 9, 2–15. doi: 10.3835/plantgenome2015.10.0109

Meng, L., Zhao, X., Ponce, K., Ye, G., and Leung, H. (2016b). QTL mapping for agronomic traits using multi-parent advanced generation inter-cross (MAGIC) populations derived from diverse elite indica rice lines. *Field Crops Res.* 189, 19–42. doi: 10.1016/j.fcr.2016.02.004

Pascual, L., Desplat, N., Huang, B. E., Desgroux, A., Bruguier, L., Bouchet, J., et al. (2015). Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnol. J.* 13, 565–577. doi: 10.1111/pbi.12282

Pathan, S. M., Vuong, T., Clark, K., Lee, J., Shannon, J. G., Roberts, C. A., et al. (2013). Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. *Crop Sci.* 53, 765–774. doi: 10.2135/cropsci2012.03.0153

Perneger, T. V. (1998). What's wrong with bonferroni adjustments. *BMJ* 316, 1236–1238. doi: 10.1136/bmj.316.7139.1236

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181. doi: 10.1086/302959

Pritchard, J. K., Wen, X., and Falush, D. (2009). *Documentation for Structure Software: Version 2.3*. Chicago, IL: The University of Chicago Press.

Qi, Z., Hou, M., Han, X., Liu, C., Jiang, H., Xin, D., et al. (2014). Identification of quantitative trait loci (QTLs) for seed protein concentration in soybean and analysis for additive effects and epistatic effects of QTLs under multiple environments. *Plant Breed.* 133, 499–507. doi: 10.1111/pbr.12179

Sannemann, W., Huang, B. E., Mathew, B., and Léon, J. (2015). Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept. *Mol. Breed.* 35, 1–16. doi: 10.1007/s11032-015-0284-7

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One* 8:e54985. doi: 10.1371/journal.pone.0054985

Tamba, C. L., Ni, Y., and Zhang, Y. (2017). Iterative sure independence screening EM-bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., et al. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet* 159–162. doi: 10.1038/ng.746

Tuan, P. A., Zhao, S., Kim, J. K., Kim, Y. B., Yang, J., Li, C. H., et al. (2014). Riboflavin accumulation and molecular characterization of cDNAs encoding bifunctional GTP cyclohydrolase II/3,4-dihydroxy-2-butanone 4-phosphate synthase, lumazine synthase, and riboflavin synthase in different organs of Lycium chinense plant. *Molecules* 19, 17141–17153. doi: 10.3390/molecules191117141

Vaughn, J. N., Nelson, R. L., Song, Q., Cregan, P. B., and Li, Z. (2014). The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. *G3* 4, 2283–2294. doi: 10.1534/g3.114.013433

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wang, Y., Li, Y., Wu, H., Hu, B., Zheng, J., Zhai, H., et al. (2018). Genotyping of soybean cultivars with medium-density array reveals the population structure and QTNs underlying maturity and seed traits. *Front. Plant Sci.* 9:610. doi: 10.3389/fpls.2018.00610

Warrington, C. V., Abdel-Haleem, H., Hyten, D. L., Cregan, P. B., Orf, J. H., Killam, A. S., et al. (2015). QTL for seed protein and amino acids in the benning × danbaekkong soybean population. *Theor. Appl. Genet* 128, 839–850. doi: 10.1007/s00122-015-2474-4

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Xu, J., Yang, H., and Zhang, W. (2018). NADPH metabolism: a survey of its theoretical characteristics and manipulation strategies in amino acid biosynthesis. *Crit. Rev. Biotechnol.* 38, 1061–1076. doi: 10.1080/07388551.2018.1437387

Yalcin, B., Flint, J., and Mott, R. (2005). Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics* 171, 673–681. doi: 10.1534/genetics.104.028902

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, Y., Ge, F., Hou, F., Sun, W., Zheng, Q., Zhang, X., et al. (2017). Transcription factors responding to Pb stress in maize. *Genes* 8:E231. doi: 10.3390/genes8090231

Zhao, Y. Y., Wang, D. F., Wu, T. Q., Guo, A., Dong, H. S., and Zhang, C. L. (2014). Transgenic expression of a rice riboflavin synthase gene in tobacco enhances plant growth and resistance to Tobacco mosaic virus. *Can. J. Plant Pathol.* 36, 100–109. doi: 10.1080/07060661.2014.881921

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 04 January 2019
doi: 10.3389/fpls.2018.01872

Association Mapping Analysis of Fatty Acid Content in Different Ecotypic Rapeseed Using mrMLM

Mingwei Guan[1,2†], Xiaohu Huang[1,2†], Zhongchun Xiao[1,2†], Ledong Jia[1,2], Shuxian Wang[1,2], Meichen Zhu[1,2], Cailin Qiao[1,2], Lijuan Wei[1,2], Xinfu Xu[1,2], Ying Liang[1,2], Rui Wang[1,2], Kun Lu[1,2], Jiana Li[1,2] and Cunmin Qu[1,2*]*

[1] *Chongqing Engineering Research Center for Rapeseed, College of Agronomy and Biotechnology, Southwest University, Chongqing, China,* [2] *Academy of Agricultural Sciences, Southwest University, Chongqing, China*

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Guo-Bo Chen,
Zhejiang Provincial People's Hospital,
China
Chunyu Zhang,
Huazhong Agricultural University,
China
Jinjin Jiang,
Yangzhou University, China

***Correspondence:**
Jiana Li
ljn1950@swu.edu.cn
Cunmin Qu
drqucunmin@swu.edu.cn

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: *14 June 2018*
Accepted: *04 December 2018*
Published: *04 January 2019*

Citation:
Guan M, Huang X, Xiao Z, Jia L, Wang S, Zhu M, Qiao C, Wei L, Xu X, Liang Y, Wang R, Lu K, Li J and Qu C (2019) Association Mapping Analysis of Fatty Acid Content in Different Ecotypic Rapeseed Using mrMLM. Front. Plant Sci. 9:1872. doi: 10.3389/fpls.2018.01872

Brassica napus L. is a widely cultivated oil crop and provides important resources of edible vegetable oil, and its quality is determined by fatty acid composition and content. To explain the genetic basis and identify more minor loci for fatty acid content, the multi-locus random-SNP-effect mixed linear model (mrMLM) was used to identify genomic regions associated with fatty acid content in a genetically diverse population of 435 rapeseed accessions, including 77 winter-type, 55 spring-type, and 303 semi-winter-type accessions grown in different environments. A total of 149 quantitative trait nucleotides (QTNs) were found to be associated with fatty acid content and composition, including 34 QTNs that overlapped with the previously reported loci, and 115 novel QTNs. Of these, 35 novel QTNs, located on chromosome A01, A02, A03, A05, A06, A09, A10, and C02, respectively, were repeatedly detected across different environments. Subsequently, we annotated 95 putative candidate genes by BlastP analysis using sequences from *Arabidopsis thaliana* homologs of the identified regions. The candidate genes included 34 environmentally-insensitive genes (e.g., *CER4*, *DGK2*, *KCS17*, *KCS18*, *MYB4*, and *TT16*) and 61 environment-sensitive genes (e.g., *FAB1*, *FAD6*, *FAD7*, *KCR1*, *KCS9*, *KCS12*, and *TT1*) as well as genes invloved in the fatty acid biosynthesis. Among these, *BnaA08g08280D* and *BnaC03g60080D* differed in genomic sequence between the high- and low-oleic acid lines, and might thus be the novel alleles regulating oleic acid content. Furthermore, RT-qPCR analysis of these genes showed differential expression levels during seed development. Our results highlight the practical and scientific value of mrMLM or QTN detection and the accuracy of linking specific QTNs to fatty acid content, and suggest a useful strategy to improve the fatty acid content of *B. napus* seeds by molecular marker-assisted breeding.

Keywords: *Brassica napus* L., candidate genes, GWAS, mrMLM, fatty acid content

INTRODUCTION

Rapeseed (*Brassica napus* L.) is one of the most important oil crops in the world, providing not only edible vegetable oil but also its potential use in lubricants and biofuels (Saeidnia and Gohari, 2012). However, the physical, chemical, and nutritional qualities of rapeseed oil depend mainly on its fatty acid composition, which consists approximately of 60% oleic acid (C18:1), 4% palmitic acid (16:0), and 2% stearic acid (18:0) (Bauer et al., 2015; Wen et al., 2015). Rapeseed oil is considered by many nutritionists to be ideal for human nutrition and superior to many other plant oils (Zhao et al., 2008; Qu et al., 2017), as it can be heated to high temperatures without smoking (Miller et al., 1987), and reduces levels of undesirable low-density lipoprotein cholesterol in the blood plasma, reducing the risk of arteriosclerosis (Chang and Huang, 1998). Optimizing the fatty acid composition is an important breeding objective for rapeseed cultivar development.

In *B. napus*, fatty acid metabolism is influenced by both genotype and environmental factors. Efforts to improve the oil quality have yielded many high oleic acid *Brassica* lines, including *B. rapa* (Tanhuanpää et al., 1996), *B. carinata* (Velasco et al., 1997), and *B. napus* (Pleines and Friedt, 1989; Fei et al., 2012). Further, oleic acid concentrations >70% have already been achieved in rapeseed through hybrid breeding methods (Zhang et al., 2009). Fatty acid content is a typical quantitative trait controlled by multiple genes that regulate its desaturation (Wang et al., 2015; Chen et al., 2018), and numerous quantitative trait loci (QTLs) for fatty acids have been mapped to all 19 chromosomes of *B. napus*, with most being found on chromosomes A01, A02, A03, A08, A10, C03, A04, A07, A09, C01, C06, and C08 (Burns et al., 2003; Zhao et al., 2008; Liu and Li, 2014; Bauer et al., 2015; Lee et al., 2015; Teh, 2015; Wen et al., 2015; Javed et al., 2016). With the increasing availability of whole-genome-sequences and SNP array development, association mapping represents a powerful approach for dissecting the genetic basis of complex quantitative traits at high resolution, which could significantly increase the precision of estimating QTL locations (Meuwissen and Goddard, 2000). Recently, genome-wide association studies (GWAS) have been performed to detect the genetic variation associated with important agronomic traits in rapeseed using the Illumina Infinium *Brassica* 60K SNP array (Delourme et al., 2013; Li et al., 2014; Lu et al., 2014; Hatzig et al., 2015; Luo et al., 2015), including seed weight and quality (Li et al., 2014), seed oil content in a panel of 521 rapeseed accessions (Liu et al., 2016), and the composition of seven fatty acids (Qu et al., 2017). Although these studies have revealed loci for associated with fatty acid traits, no beneficial alleles have been detected within the *B. napus* accessions.

Numerous studies showed that *FATTY ACID DESATURASE 2* (*FAD2*) is the major gene responsible for the desaturation of oleic acid to linolenic acid (Hu et al., 2006; Peng et al., 2010; Yang et al., 2012), and four paralogs of *FAD2* were previously identified in *B. napus* (Scheffler et al., 1997; Yang et al., 2012). These paralogs were mainly expressed in the developing seeds, suggesting possible roles in controlling oleic acid content in *B. napus* (Xiao et al., 2008). In addition, *KCS18*, is known to play a crucial role in regulating erucic acid biosynthesis in *B. napus* (Wang et al., 2008; Wu et al., 2008; Li et al., 2014). However, the identified QTL were not cloned and undertaken for contributing to the minor fatty acids. Furthermore, the genetic basis of fatty acid synthesis is still unclear.

The multi-locus random-SNP-effect mixed linear model (mrMLM) is emerged as a powerful tool for quantitative trait nucleotide (QTN) detection and QTN effect estimation for complex traits (Wang et al., 2016; Li et al., 2017; Chang et al., 2018; Peng et al., 2018). For example, Li et al. (2017) detected 38 significantly-associated loci and identified numerous highly-promising candidate genes (e.g., *TAC1*, *SGR1*, *SGR3*, and *SGR5*), for branch angle across 472 rapeseed accessions. Zhang et al. (2018) identified 127 significant QTNs for stalk lodging resistance-related traits using mrMLM in a population of 257 maize inbred lines. As reported by Ma et al. (2018), 127 significant QTNs with maize embryonic callus regenerative capacity were identified in a population of 144 maize inbred lines, and many candidate genes were reported to relate with auxin transport, cell fate, seed germination, or embryo development, respectively. In the present study, we analyzed the fatty acid composition in 77 winter varieties, 55 spring varieties, and 303 semi-winter varieties of rapeseed grown in three environments, and genotyped all of the accessions using the high-through *Brassica* 60K SNP array (Clarke et al., 2016). Then, 32,543 SNPs from the 60K SNP array were used for genome-wide association analysis usingmrMLM. In total, 149 QTNs were identified using mrMLM, suggesting that this is an effective model for identifying candidate genes underlying complex traits. Subsequently, 95 candidate genes were annotated using BlastP against *A. thaliana* homologs, providing insight into the genetic control of fatty acid content in *B. napus*. Furthermore, novel fatty acid content-associated SNPs identified here may be useful for marker-based breeding programs aimed at improving the fatty acid content of *B. napus* seeds.

MATERIALS AND METHODS

Plant Materials

A diversity panel consisting of 55 spring, 77 winter, and 303 semi-winter rapeseed accessions (*B. napus*; **Supplementary Table S1**) was used for the association analysis. These accessions were grown in three growing seasons (2015–2016, 2016–2017, and 2017–2018) in Beibei (106.38°E, 29.84°N), Chongqing, China. Three rows of 10–12 plants per accession were established in the experimental fields with a randomized complete block design and three replications. Self-pollinated seeds were harvested from plants at complete physiological maturity and used for the fatty acid analysis.

Fatty Acid Measurement and Statistical Analysis

Seeds (200 mg) were homogenized with a pestle and extracted in 2 mL petroleum ether:ether (1:1) for 40 min, and methylated with 1 mL KOH/methanol (0.4 mol L^{-1}). The supernatants separated by adding distilled water were identified by gas-liquid chromatography on a Model GC-2010 (Shimadzu, Kyoto, Japan). Chromatographic analysis was carried out using a fused silica capillary column DB-WAX (30 m × 0.246 mm × 0.25 um) with default parameters (Qu et al., 2017). Fatty acid profiles were calculated as a percentage of total fatty acids in the seeds, and optimized with an R script of the best linear unbiased prediction (BLUP) (Merk et al., 2012). The resulting values for each accession were used in the association analysis. All experiments were performed in triplicate, and the mean, standard deviation (SD), coefficient of variation (CV), and minimum (Min) and maximum (Max) values of the oleic acid content were calculated using SPSS 15.0 (IBM Corp, Armonk, NJ, USA).

SNP Genotyping Data Acquisition and Analysis

The methods used for SNP genotyping and mapping were described in previous reports of Li et al. (2014) and Liu et al. (2016). Using the *Brassica* 60K Illumina® Infinium SNP array (Clarke et al., 2016), the genotype of each accession was generated at the National Subcenter of Rapeseed Improvement in Wuhan (Huazhong Agricultural University, Wuhan, China). The low quality SNPs (call frequency <0.9 and a minor allele frequency ≤ 0.05) were filtered in all accessions. In addition, SNPs not accurately mapped to the *B. napus* genome were excluded. The probe sequences of 52,157 SNPs were used to perform a local BlastN search against the *B. napus* "Darmor-*bzh*" reference genome (version 4.1, http://www.genoscope.cns.fr/brassicanapus/data/; Chalhoub et al., 2014) using our previously published method (Wei et al., 2015). In total, 32,543 SNPs were analyzed further.

The Q matrix of population structure was calculated by a Bayesian model-based analysis in STRUCTURE 2.1 (Pritchard et al., 2000) with published parameters of Falush et al. (2003) and Qu et al. (2017). The optimal number of *K* values ($K = 2$; **Supplementary Figure S1**) was determined using the Evanno method (Evanno et al., 2005). The Q matrix was selected as the fixed covariate in the subsequent association analysis (Gajardo et al., 2015). To visualize genetic relatedness among all genotypes, the principal component analysis (PCA) was constructed using the GCTA tool (Yang et al., 2011). The relative kinship matrix for each association panel was calculated using SPAGeDi (Hardy and Vekemans, 2002), and the negative values were defined as zero between two individuals, following the method of Yu et al. (2006).

Genome-Wide Association Analysis

The mrMLM significantly improved the power and precision of the GWAS, which was previously used in *B. napus* (Li et al., 2017). Therefore, the multi-locus GWAS method (mrMLM, https://cran.r-project.org/web/packages/mrMLM.GUI/index.html) was performed to evaluate the trait-SNP association analysis in this study (Wang et al., 2016). Moreover, the phenotypic and genotypic datasets, kinship (K), and population structure (Q) were imported into the R package mrMLM, and significantly associated SNPs were identified by mrMLM with the critical log of odds (LOD) score of 3.0 ($p = 0.0002$) (Wang et al., 2016). The QTNs were named using the nomenclature described by McCouch et al. (1997). For example, *q-C16:0-A03-1* indicated the first locus located on chromosome A03 associated with palmitic acid.

Candidate Gene Prediction

Candidate genes were identified using significant SNP markers, which were detected using mrMLM (Wang et al., 2016). The association regions and 100-kb region upstream or downstream of peak SNPs associated with fatty acid content were identified based on the physical distance of chromosomes of significant associated-trait SNPs in the *B. napus* "Darmor-*bzh*" genome (version 4.1; Chalhoub et al., 2014). Subsequently, putative candidate genes were predicted according to the annotation of the SNP-tagged genome regions and confirmed by BlastP searches against the *Arabidopsis* genome with an *E*-value $\leq$1E-10. The function of these candidate genes was further annotated using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (https://www.genome.jp/kegg/pathway.html). Highly-orthologous genes involved in fatty acid biosynthesis were analyzed further, which were defined as the environment-insensitive and -sensitive genes according to the frequency detected between the different ecological genotypes and environments. To identify the directed functional genes for fatty acid, sequences of these candidate genes, isolated from plants with higher- and lower fatty acid levels, were aligned using ClustalW (Thompson et al., 1994) implemented in Geneious 4.8.5 software (Biomatters, Auckland, New Zealand).

Analysis of the Expression Profiles of Candidate Fatty Acid-Associated Genes

Total RNA was extracted from the seeds of *B. napus* cultivar Zhongshuang No. 11 (ZS11) at 15 developmental stages (3–49 days after pollination) using the RNAprep Pure Plant Kit (Tiangen Biotech, Beijing, China), following the manufacturer's instructions. The cDNA library construction and RNA sequencing were performed using Novogene Bioinformatics Technology (Beijing, China). Transcriptome sequencing datasets were deposited in the BioProject database (BioProject ID PRJNA358784). The data were analyzed as previously described (Qu et al., 2015), and the expression profiles of the candidate genes were quantified in terms of fragments per kilobase of exon per million mapped fragments (FPKM), using Cufflinks with default parameters (Trapnell et al., 2012). These transcriptome datasets were previously deemed suitable for selecting candidate genes (Zhou et al., 2017). Candidate genes were validated using RT-qPCR analysis. Three biological replicates and three technical replicates were performed on a CFX96 Real-Time PCR system (Bio-Rad, Laboratories, Hercules, CA, USA), and the expression levels of candidate genes were calculated using the $2^{-\Delta\Delta Ct}$ method (Zhou et al., 2017). Hence, the expression values of the 106 candidate genes were normalized by Log_2 (expression values). Heatmaps of the candidate genes were drafted using HemI 1.0 (http://hemi.biocuckoo.org/). The specific primer sequences used in this study were obtained from the qPCR Primer Database (Lu et al., 2017) and are listed in **Supplementary Table S3**.

RESULTS

Phenotypic Variation and Correlation Among Different Rapeseed Genotypes

Extensive variation in fatty acid content was observed between rapeseed plants of different genotypes grown in over 3 years (**Table 1**). The content of palmitic acid, stearic acid and linolenic acid varied slightly among the different ecotypic rapeseed accession at different years. For example, the mean palmitic acid

TABLE 1 | Statistical analysis of fatty acid content in different ecological types of rapeseed grown in different environments.

Fatty acid	Env.	Min.	Max.	Mean ± SD	CV(%)	Skewness	Kurtosis	F_G	F_E
Palmitic acid	16Sp	2.93	5.44	3.92 ± 0.07	13.27	0.15	0.29	7.65**	4.82*
	17Sp	2.53	5.49	3.99 ± 0.09	17.29	0.05	−0.19		
	18Sp	3.03	5.14	4.18 ± 0.07	10.77	−0.36	0.19		
	16Win	2.77	5.06	4.04 ± 0.06	12.13	−0.35	0.18	6.81**	3.38*
	17Win	2.66	5.78	4.09 ± 0.08	16.14	0.20	−0.45		
	18Win	2.92	5.15	4.12 ± 0.05	10.44	−0.26	0.26		
	16Semi	3.19	6.10	4.89 ± 0.07	12.07	−0.80	0.64	7.10**	26.45**
	17Semi	2.86	5.29	4.38 ± 0.06	14.16	−0.75	−0.46		
	18Semi	2.69	5.04	4.20 ± 0.06	14.52	−0.93	−0.30		
Stearic acid	16Sp	0.92	2.94	1.85 ± 0.06	13.78	0.13	−0.07	5.57**	39.73**
	17Sp	1.04	1.96	1.42 ± 0.03	15.49	0.43	−0.21		
	18Sp	0.24	0.79	0.33 ± 0.03	13.64	−0.01	−0.69		
	16Win	0.70	3.98	1.86 ± 0.07	13.87	0.78	1.04	9.62**	17.99**
	17Win	0.55	3.46	1.47 ± 0.05	12.65	1.16	3.42		
	18Win	0.05	0.80	0.36 ± 0.03	58.33	0.00	−0.90		
	16Semi	0.09	1.06	0.62 ± 0.03	13.87	−0.68	0.23	8.97**	26.96**
	17Semi	1.05	2.89	2.08 ± 0.05	12.60	−0.45	−0.81		
	18Semi	0.58	2.83	1.91 ± 0.05	16.18	−0.68	−0.25		
Oleic acid	16Sp	14.49	74.02	56.37 ± 2.53	33.30	−1.22	−0.10	9.39**	309.73**
	17Sp	14.69	72.21	53.16 ± 2.65	36.98	−1.15	−0.36		
	18Sp	23.09	69.20	50.29 ± 1.86	25.33	−0.80	−0.40		
	16Win	10.97	76.03	56.49 ± 2.34	36.38	−1.27	−0.04	7.33**	179.39**
	17Win	10.65	76.19	53.73 ± 2.45	39.96	−1.08	−0.53		
	18Win	11.78	76.17	52.24 ± 1.99	30.88	−1.07	−0.20		
	16Semi	10.66	71.37	52.89 ± 1.73	27.15	−1.75	1.75	9.97**	269.64**
	17Semi	14.12	73.65	52.94 ± 1.88	34.91	−0.95	−0.77		
	18Semi	7.91	83.00	49.29 ± 2.11	40.62	−0.87	−0.77		
Linoleic acid	16Sp	11.68	26.15	17.38 ± 0.42	17.89	0.50	−0.09	8.67**	3.57*
	17Sp	10.98	28.32	18.76 ± 0.54	21.54	0.37	−0.20		
	18Sp	11.20	24.41	18.13 ± 0.45	17.21	−0.27	−0.52		
	16Win	11.34	27.61	17.13 ± 0.31	15.70	0.52	1.75	5.30**	3.21*
	17Win	10.07	26.87	17.87 ± 0.44	21.66	0.24	−0.54		
	18Win	10.92	23.10	17.31 ± 0.33	15.60	−0.05	−0.64		
	16Semi	12.66	28.07	22.34 ± 0.44	16.47	−0.72	0.10	5.57**	69.05
	17Semi	11.34	22.86	17.92 ± 0.29	15.85	−0.73	−0.51		
	18Semi	10.41	21.42	16.79 ± 0.29	16.62	−0.72	−0.64		
Linolenic acid	16Sp	4.90	10.04	7.39 ± 0.16	16.37	0.09	−0.55	6.82**	35.98**
	17Sp	6.73	12.15	9.12 ± 0.16	12.83	0.23	−0.09		
	18Sp	5.95	14.61	9.34 ± 0.22	15.95	0.51	2.45		
	16Win	2.15	11.41	7.45 ± 0.19	12.82	−0.15	0.62	7.16**	23.34**
	17Win	6.26	11.36	8.75 ± 0.12	12.11	−0.08	−0.35		
	18Win	5.60	11.13	8.79 ± 0.15	13.42	−0.03	0.02		
	16Semi	8.31	12.31	10.02 ± 0.11	9.08	0.36	−0.04	6.85**	35.01**
	17Semi	6.54	12.06	8.84 ± 0.11	12.78	0.30	−0.25		
	18Semi	6.88	11.69	8.84 ± 0.10	10.52	0.70	0.92		
Eicosenoic acid	16Sp	1.92	17.16	3.58 ± 0.72	50.00	1.20	0.09	9.27**	9.41**
	17Sp	2.31	17.03	3.10 ± 0.68	63.55	1.35	0.48		

(Continued)

TABLE 1 | Continued

Fatty acid	Env.	Min.	Max.	Mean ± SD	CV(%)	Skewness	Kurtosis	F_G	F_E
	18Sp	2.91	16.30	6.96 ± 0.54	53.30	0.92	−0.25		
	16Win	1.12	18.05	3.07 ± 0.58	166.12	1.35	0.39	4.76**	12.92**
	17Win	2.01	16.81	2.47 ± 0.51	183.00	1.54	0.98		
	18Win	2.21	16.95	6.09 ± 0.45	60.10	1.17	0.32		
	16Semi	1.01	15.54	4.51 ± 0.43	79.60	1.56	1.70	7.33**	6.77**
	17Semi	0.78	19.44	4.88 ± 0.61	122.75	0.96	−0.52		
	18Semi	0.22	22.34	7.42 ± 0.68	86.52	0.69	−0.80		
Erucic acid	16Sp	2.42	53.41	25.82 ± 4.00	69.25	0.06	−1.50	7.99**	9.618**
	17Sp	1.19	52.77	13.42 ± 1.95	101.64	1.22	0.34		
	18Sp	0.00	36.13	10.38 ± 1.53	100.96	1.26	0.28		
	16Win	3.04	52.83	32.92 ± 3.73	54.28	−0.41	−1.41	9.35**	28.95**
	17Win	0.24	40.70	10.90 ± 1.40	107.34	1.30	0.31		
	18Win	0.00	38.61	10.15 ± 1.54	123.05	1.11	−0.37		
	16Semi	0.00	37.65	5.82 ± 1.40	199.83	1.90	1.93	6.24**	35.21**
	17Semi	0.00	42.92	10.11 ± 1.57	152.92	1.11	−0.62		
	18Semi	0.00	48.58	12.15 ± 1.77	138.52	1.01	−0.74		

*Env., environment; Min., Minimum; Max., Maximum; SD, standard deviation; CV, coefficient of variation; Sp, Spring-type rapeseed; Win, Winter-type rapeseed; Semi, Semi-winter-type rapeseed; 16, 17, and 18 represent the 2016, 2017, and 2018 growing seasons in Chongqing, China, respectively. F_G and F_E: the F-values for genotypes and environments, respectively. * and **: the 0.05 and 0.01 levels of significance, respectively.*

content varied from 2.53 to 5.49% in spring rapeseed, 2.66 to 5.78% in winter rapeseed, and 2.69 to 6.10% in semi-winter rapeseed. The stearic acid content varied from 0.24 to 2.94% in spring rapeseed, 0.05 to 3.98% in winter rapeseed, and 0.09 to 2.89% in semi-winter rapeseed. The linolenic acid content varied from 4.90 to 14.61% in spring rapeseed, 2.15 to 11.41% in winter rapeseed, and 6.54 to 12.31% in semi-winter rapeseed. However, considerable quantitative variation was found for the content of oleic acid, linoleic acid, eicosenoic acid, and erucic acid. For instance, the mean oleic acid content ranged from 14.49 to 72.21% in spring rapeseed, 10.65 to 76.19% in winter rapeseed, and 7.91 to 83.00% in semi-winter rapeseed; the linoleic acid content ranged from 10.98 to 28.32%, 10.07 to 27.61%, and 10.41 to 28.07% in spring, winter, and semi-winter rapeseed, respectively, the eicosenoic acid content were ranged from 1.92 to 17.16%, 1.12 to 18.05%, and 0.22 to 22.34% in spring, winter and semi-winter rapeseed, respectively, and the erucic acid content ranged from 0 to 53.41 μmol g^{-1}, 0 to 52.83 μmol g^{-1}, and 0 to 48.58 μmol g^{-1} in spring, winter and semi-winter rapeseed, respectively (**Table 1**). Moreover, the largest CV (coefficient of variation) was found among the oleic acid, eicosenoic acid, and erucic acid content in different ecotypic rapeseed at different environments, ranging from 25.33 to 40.62, 50.00 to 183.00%, and 54.28 to 199.83%, respectively (**Table 1**), indicating that extensive variation was widely detected in the panel of accessions. In addition, the phenotypic values of fatty acid content were displayed among the ecotypic rapeseed accessions at different years (**Figures 1A–U**, **Table 1**). Of these, the palmitic acid, stearic acid, linoleic acid, and linolenic acid content were normally distributed, but the eicosenoic, oleic, and erucic acids content were skewed for three genotypic populations in different years (**Figures 1A–U**). Plants with higher oleic acid content were more common than those with lower content for each ecological type of rapeseed (**Figures 1G–I**). Analysis of variance (ANOVA) was performed among the spring, winter, and semi-winter rapeseed ecological types in different years, and showed that genotype and environment have significant effects on the fatty acid content of rapeseed (**Table 1**).

Genome-Wide Association Analysis of Fatty Acid via mrMLM

For palmitic acid (C16:0) content, 11 QTNs were detected on chromosomes A01, A03, A04, A06, A07, A08, C01, and C03, respectively (**Table 2**). Of these, three consensus QTNs (*q-C16:0-A06*, *q-C16:0-A08-2,* and *q-C16:0-C03-2*) were commonly detected for palmitic acid among different ecotypic rapeseed and ecotypic rapeseed cultivated in different years, providing useful information for searching for candidate genes associated with palmitic acid biosynthesis. However, *q-C16:0-A01*, *q-C16:0-A04*, *q-C16:0-A06*, and *q-C16:0-A07* were mainly found in the spring-type and all 3 years, and other QTNs were detected among different ecotypic rapeseed and years (**Table 2**).

For stearic acid (C18:0) content, a total of 9 QTNs were resolved and distributed on A03, A06, A08, A09, A10, and C03, respectively (**Table 3**). Among them, two QTNs, *q-C18:0-A08,* and *q-C18:0-C03-2*, were detected in different ecotypic rapeseed and environments, and others were detected in different ecotypic rapeseed grown in at least two different environments.

For oleic acid (C18:1) content, 21 QTNs were detected and distributed throughout of the *B. napus* genome, including chromosomes A01, A02, A03, A05, A08, A09, C01, C02, C03, C04, C05, C07, C08, and C09, respectively (**Table 2**). Of these, seven QTNs (*q-C18:1-A08-3*, *q-C18:1-A08-4*, *q-C18:1-A09*, *q-C18:1-C03*, *q-C18:1-C04*, *q-C18:1-C08*, and *q-C18:1-C09*) were

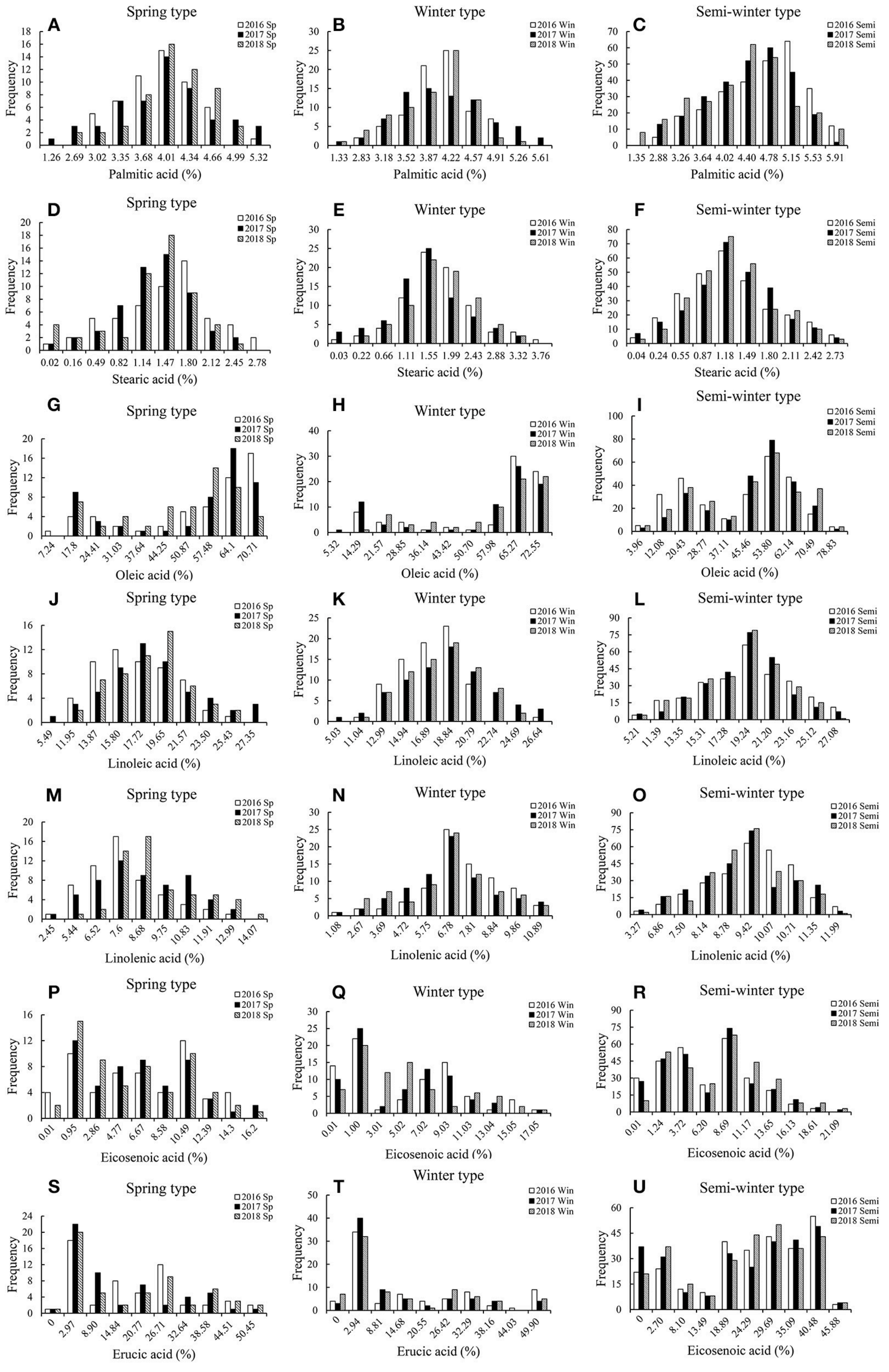


FIGURE 1 | The frequency distribution of fatty acid contents in the different rapeseed accessions grown in different environments. The percentage indicates the proportion of the total dry weight of the seed represented by fatty acid composition. Sp, Spring-type rapeseed; Win, Winter-type rapeseed; Semi, Semi-winter-type

(Continued)

FIGURE 1 | rapeseed; 16, 17, and 18 represent the 2016, 2017, and 2018 growing seasons in Chongqing, China. **(A–C)** The frequency distribution of Palmitic acid contents in Spring-type, Winter-type, and Semi-winter-type rapeseed; **(D–F)** The frequency distribution of Stearic acid contents in Spring-type, Winter-type, and Semi-winter-type rapeseed; **(G–I)** The frequency distribution of Oleic acid contents in Spring-type, Winter-type, and Semi-winter-type rapeseed; **(J–L)** The frequency distribution of Linoleic acid contents in Spring-type, Winter-type, and Semi-winter-type rapeseed; **(M–O)** The frequency distribution of Linolenic acid contents in Spring-type, Winter-type, and Semi-winter-type rapeseed; **(P–R)** The frequency distribution of Eicosenoic acid contents in Spring-type, Winter-type, and Semi-winter-type rapeseed; **(S–U)** The frequency distribution of Erucic acid contents in Spring-type, Winter-type, and Semi-winter-type rapeseed.

co-localized in the same genomic regions of A08, A09, C03, C04, C08, and C09 using mrMLM and the PCA+K model (Qu et al., 2017). These seven QTNs were considered the major candidate regions for oleic acid content.

For linoleic acid (C18:2) content, fourteen QTNs were detected and mapped on chromosomes A01, A03, A04, A06, A07, A08, A09, A10, C01, C03, C05, C07, and C09, respectively (**Table 2**). Of these, *q-C18:2-A08-2*, *q-C18:2-A09*, *q-C18:2-C03,* and *q-C18:2-C07* were identified in our previous research (Qu et al., 2017).

For linolenic acid (C18:3) content, a total 48 QTNs were found and covered almost the whole *B. napus* genome (**Table 2**). Among them, seven highly identical QTNs distributed on chromosome A01, A02, A05, A06, A08, A09, and C02, along with other minor loci could be identified in different ecotypic rapeseed accessions in at least 1 year of growth (**Table 2**).

For eicosenoic acid (C20:1) content, 30 QTNs were obtained, including seven QTNs that overlapped with previously published QTNs (Qu et al., 2017), and were distributed on chromosome A01, A04, A06, A08, C03, C07, and C09, respectively (**Table 2**). The novel loci for eicosenoic acid content also displayed marked variation among the different ecotypic rapeseeds and environments.

For erucic acid (C22:1) content, 16 QTNs were detected and mapped on chromosome A01, A02, A03, A06, A08, A09, A10, C02, C03, C05, C07, and C08, respectively (**Table 2**). Of these, two QTNs (*q-C22:1-A08* and *q-C22:1-C03*) had been widely considered as the major genetic regions of A08 and C03, consistent with the findings of published works (Li et al., 2014; Lee et al., 2015; Xu et al., 2015; Qu et al., 2017). In addition, two QTNs (*q-C22:1-A09-1* and *q-C22:1-C08-1*) associated with erucic acid content were also detected in different ecotypic rapeseed and environments (**Table 2**), indicating that mrMLM is a powerful and accurate tool to detect QTNs and estimate the effect of QTNs on complex traits.

In all, 149 QTNs associated with fatty acid content were detected using mrMLM (**Table 2**, **Supplementary Table S2**), while only 62 associated regions were detected using the PCA + K model in TASSEL 5.2.1 (Qu et al., 2017). Among these, 34 QTNs were overlapped, including the association regions on A08 and C03, which had been widely reported previously (Wang et al., 2015; Liu et al., 2016; Qu et al., 2017), indicating that these results were credible and reproducible. In addition, 115 novel loci were identified for fatty acids via mrMLM compared with MLM (PCA+K; **Table 2**), indicating that a multi-locus random effect MLM method was better able to detect QTNs of complex quantitative traits. Furthermore, of these, 30.43% novel QTNs (35/115) were repeatedly detected for fatty acid content among different ecotypic rapeseed accessions and environments, located on chromosome A01, A02, A03, A05, A06, A09, A10, and C02, respectively (bold in **Table 2**), and other novel QTNs (80/115) were found for fatty acid content in a single environment. Among these QTNs, 29 were simultaneously detected in three ecotypic rapeseed, with greater QTN variation in spring and semi-winter rapeseed (**Figure 2**, **Supplementary Table S2**). Our results provideinsight into the mechanism underlying fatty acid composition, and lay the foundation for marker assisted selection in breeding projects for improved rapeseed genotypes with high oil quality and an ideal fatty acid composition.

Identification of Candidate Genes

To predict candidate genes for loci significantly associated with fatty acid content, the reported and repeadly detected novel QTNs were used to confirm the genomic regions in the *B. napus* "Darmor v4.1" reference genome (Chalhoub et al., 2014). We identified five environment-insensitive and fifteen environment-sensitive association regions and screened for candidate genes within these regions. Subsequently, we extracted gene sequences within the GWAS peaks in candidate association regions, and identified 95 putative genes that possibly influence fatty acid content (**Table 3**). Of these, 63.16% candidate genes (60/95) were screened in the overlapping and repeatedly detected association regions, while the remaining genes (35/95) were detected on the single QTN regions (**Table 3**). Using peak SNPs on A08 and C03 (Bn-A08-p13066424 and Bn-scaff_15794_3-p89999, respectively), 9 and 4 candidate genes were selected in the association regions on each chromosome, respectively (**Table 3**). *BnaA08g11130D* and *BnaC03g65980D* are putative paralogs of 3-ketoacyl-CoA synthase 18 (*KCS18*), while *BnaA08g11140D* and *BnaC03g66040D* are putative paralogs of *KCS17* (**Table 3**), based on comparisons of the physical positions of genes associated with erucic acid traits in a GWAS (Wu et al., 2008; Li et al., 2014; Lee et al., 2015; Xu et al., 2015), indicating that there is a strong correlation between GWAS peak regions and candidate genes. In addition, the putative candidate genes were characterized and annotated, such as 3-methylcrotonyl-CoA carboxylase (*MCCB*, *BnaA08g11650D*), *TRANSPARENT TESTA*16 (*TT16*, *BnaA09g05410D,* and *BnaC02g42240D*), and *MYB4* (*BnaC03g60080D*), which might be environment-insensitive genes located on chromosome A06, A09, and C02 respectively (**Table 3**). Of these, 17 genes had been identified that might be involved in the fatty acid pathway, and 12 members were annotated for fatty acid metabolism in KEGG analysis (**Table 3**).

The expression of candidate genes in low-frequency association regions identified in this study were influenced

TABLE 2 | Quantitative trait nucleotides (QTNs) associated with fatty acid content in *B. napus* accessions grown in different environments.

QTN	Chr	SNP associated	Position (bp)	$-\log_{10}(P)$	Environment
q-C16:0-A01	A01	Bn-A01-p22085117	18711173	8.54	16Sp, 17Sp, 18Sp
q-C16:0-A03-1	A03	Bn-A03-p20145024	19008703	9.93	17Sp, 17Semi
q-C16:0-A03-2	A03	Bn-A03-p28560659	27022904	5.13	16Win, 17Sp
q-C16:0-A04	A04	Bn-A04-p14687930	15158346	6.72	16Sp, 17Sp, 18Sp
q-C16:0-A06	A06	Bn-A06-p16071214	17559622	5.68	16Sp, 17Sp, 18Sp
q-C16:0-A07	A07	Bn-A07-p10430301	11624458	7.78	17Sp, 18Sp
q-C16:0-A08-1	A08	Bn-scaff_16110_1-p214256	5357890	7.78	16Win, 17Sp, 18Sp, 18Semi
q-C16:0-A08-2	*A08*	Bn-A08-p13379983	11124385	7.70	16Win, 16Semi, 17Sp, 17Semi, 18Semi, 18Win
q-C16:0-C01	C01	Bn-A01-p12593802	16875948	15.26	17Win, 18Sp
q-C16:0-C03-1	C03	Bn-scaff_17298_1-p1471882	23560777	6.12	16Win, 17Win
q-C16:0-C03-2	C03	Bn-scaff_15794_3-p108033	55728615	8.09	16Sp,16Win, 16Semi, 17Sp, 17Semi, 18Semi
q-C18:0-A03-1	A03	Bn-A03-p21942870	20741914	6.65	16Semi, 18Win
q-C18:0-A03-2	A03	Bn-A03-p27339890	25582011	10.02	16Win, 17Sp
q-C18:0-A06	A06	Bn-A06-p6341389	5792362	5.22	17Sp, 18Sp
q-C18:0-A08	A08	Bn-A08-p10068904	8070062	5.39	16Win, 16Semi, 17Sp, 17Semi, 18Semi, 18Win
q-C18:0-A09-1	A09	Bn-A09-p3234323	3135040	5.30	16Semi, 17Sp, 18Win
q-C18:0-A09-2	A09	Bn-A09-p7329993	5542359	7.85	16Sp, 17Sp, 17Win
q-C18:0-A10	A10	Bn-A10-p13965313	13956813	11.41	16Sp, 17Sp
q-C18:0-C03-1	C03	Bn-scaff_17298_1-p779577	23106957	5.67	16Win, 17Sp, 17Win
q-C18:0-C03-2	C03	Bn-scaff_17457_1-p493971	53921047	6.00	17Semi, 18Sp
q-C18:1-A01-1	A01	Bn-A01-p2825565	2327566	8.27	16Semi, 18Sp
q-C18:1-A01-2	A01	Bn-A01-p26369651	20893064	11.54	17Sp, 18Sp
q-C18:1-A02-1	A02	Bn-A02-p10591779	7458917	7.37	16Semi, 18Sp, 18Win
q-C18:1-A02-2	A02	Bn-A02-p21061002	18906336	11.48	16Sp, 18Sp
q-C18:1-A02-3	A02	Bn-A02-p22386317	20775741	5.74	17Sp, 18Sp
q-C18:1-A03	A03	Bn-A03-p20369417	19241578	6.16	17Semi, 18Semi
q-C18:1-A05	A05	Bn-A05-p20425452	18636249	11.54	17Sp, 18Sp
q-C18:1-A08-1	A08	Bn-A08-p4077507	3476858	7.15	16Semi, 17Semi, 18Win
q-C18:1-A08-2	A08	Bn-A08-p7814432	6786988	7.38	16Semi, 17Semi, 18Semi, 18Win
q-C18:1-A08-3	A08	Bn-A08-p10068904	8070062	12.15	16Sp, 16Win, 16Semi, 17Sp, 17Win, 17Semi, 18Sp, 18Win, 18Semi
q-C18:1-A08-4	A08	Bn-A08-p12820786	10587675	13.42	16sp, 16Win, 16Semi, 17sp, 17Win, 17Semi, 18Sp, 18Semi
q-C18:1-A08-5	A08	Bn-scaff_24726_1-p33555	14029706	8.52	16Sp, 17Sp
q-C18:1-A09	A09	Bn-A09-p3051349	2971334	7.53	16Sp, 16Win, 17Semi, 18Win, 18Semi
q-C18:1-C01	C01	Bn-scaff_21015_1-p34786	32559311	13.22	16Sp, 16Semi, 18Sp
q-C18:1-C02	C02	Bn-scaff_16139_1-p1277806	45267495	8.71	16Sp, 16Win, 17Sp, 17Win, 18Semi
q-C18:1-C03	C03	Bn-scaff_15794_3-p89999	55717350	10.88	16Win, 16Sp, 16Semi, 17Sp, 17Win, 18Win
q-C18:1-C04	C04	Bn-scaff_16394_1-p1090896	32408105	6.18	17Sp, 17Semi, 18Sp
q-C18:1-C05	C05	Bn-scaff_20901_1-p1505546	2515848	11.54	16Semi, 18Sp
q-C18:1-C07	C07	Bn-scaff_16069_1-p431757	36777489	8.45	16Win, 17Sp, 18Sp, 18Win
q-C18:1-C08	C08	Bn-scaff_16361_1-p2793822	30283886	12.95	16Semi, 17Sp, 17Win, 18Sp
q-C18:1-C09	C09	Bn-scaff_16456_1-p415818	35068732	9.32	17Win, 18Sp
q-C18:2-A01	A01	Bn-A01-p4167795	3847687	7.15	17Semi, 18Sp
q-C18:2-A03	A03	Bn-A03-p20369417	19241578	7.04	17Semi, 17Sp
q-C18:2-A04	A04	Bn-A04-p14687930	15158346	8.26	17Sp, 18Sp
q-C18:2-A06	A06	Bn-A06-p22331680	21365690	5.08	17Sp, 18Sp
q-C18:2-A07	A07	Bn-A07-p14682292	22343999	6.88	17Sp, 18Win
q-C18:2-A08-1	A08	Bn-scaff_16110_1-p214256	5357890	12.18	16Semi, 17Sp, 17Semi

(Continued)

TABLE 2 | Continued

QTN	Chr	SNP associated	Position (bp)	$-\log_{10}(P)$	Environment
q-C18:2-A08-2	A08	Bn-A08-p14351709	12051686	12.18	16Semi, 17Sp, 17Semi, 18Semi, 18Win
q-C18:2-A09	A09	Bn-A09-p36112515	33233968	7.01	16Sp, 18Sp
q-C18:2-A10	A10	Bn-A10-p14179334	14175178	15.95	16Sp, 17Sp
q-C18:2-C01	C01	Bn-A08-p9268915	32559113	7.22	16Win, 18Sp
q-C18:2-C03	C03	Bn-scaff_15794_3-p108033	55728615	7.06	16Semi, 18Win, 17Semi, 18Semi
q-C18:2-C05	C05	Bn-scaff_16414_1-p863783	1091070	6.60	17Sp, 18Sp
q-C18:2-C07	C07	Bn-scaff_15705_1-p2274493	35279701	5.25	18Sp, 18Win
q-C18:2-C09	C09	Bn-scaff_18944_1-p566719	19915878	12.18	16Sp, 17Sp
q-C18:3-A01-1	A01	Bn-A01-p5243181	4826424	13.86	17Semi, 18Semi
q-C18:3-A01-2	A01	Bn-A01-p15090383	12600997	10.52	18Semi, 18Sp
q-C18:3-A01-3	A01	Bn-A01-p24431478	20229832	13.03	16Sp, 18Semi
q-C18:3-A02-1	A02	Bn-scaff_15714_1-p1537912	929885	12.25	17Win, 18Semi
q-C18:3-A02-2	A02	Bn-A02-p18101171	17261238	15.65	18Sp, 18Semi
q-C18:3-A03-1	A03	Bn-A03-p7011746	6295737	10.52	18Semi, 18Sp
q-C18:3-A03-2	A03	Bn-A03-p16162908	15257414	13.69	17Semi, 18Sp, 18Semi
q-C18:3-A03-3	A03	Bn-A03-p23609377	22177215	8.3	18Semi, 18Sp
q-C18:3-A04-1	A04	Bn-A04-p2765547	2466391	15.18	18Sp, 18Semi
q-C18:3-A04-2	A04	Bn-A04-p7629926	8963652	8.82	18Sp, 18Semi
q-C18:3-A04-3	A04	Bn-A04-p15296217	15753636	11.33	18Semi, 18Sp
q-C18:3-A05-1	A05	Bn-A05-p461633	571525	6.67	17Semi, 18Semi
q-C18:3-A05-2	A05	Bn-A05-p10939740	9532568	12.67	18Sp, 18Semi
q-C18:3-A05-3	A05	Bn-A05-p14206169	16030064	14.04	18Sp, 18Semi
q-C18:3-A06-1	A06	Bn-A06-p73924	60018	15.95	18Semi, 18Sp
q-C18:3-A06-2	A06	Bn-A06-p5535537	5007675	6.51	18Sp, 18Semi
q-C18:3-A06-3	A06	Bn-A06-p22331680	21365690	13.54	17Semi, 18Sp, 18Semi, 18Win
q-C18:3-A07-1	A07	Bn-Scaffold012966-p76	12552424	14.04	18Sp, 18Semi
q-C18:3-A07-2	A07	Bn-scaff_19937_1-p20028	21340943	8.18	17Semi, 18Semi, 18Sp
q-C18:3-A08-1	A08	Bn-A08-p2274232	1778991	10.52	18Sp, 18Semi
q-C18:3-A08-2	A08	Bn-A08-p6828857	5776774	6.55	16Sp, 17Semi, 18Semi
q-C18:3-A08-3	A08	Bn-A08-p15239790	12798553	6.43	17Semi, 18Semi, 18Win
q-C18:3-A08-4	A08	Bn-A05-p8245454	17667610	8.73	18Win, 18Semi
q-C18:3-A09-1	A09	Bn-A09-p2323366	1519271	14.04	18Sp, 18Semi
q-C18:3-A09-2	A09	Bn-A09-p24113289	23069752	10.68	17Semi, 18Semi, 18Sp
q-C18:3-A09-3	A09	Bn-A09-p31492693	29184323	14.04	18Sp, 18Semi
q-C18:3-A10-1	A10	Bn-A10-p3909275	913569	13.5	16Sp, 17Win, 18Sp, 18Semi
q-C18:3-A10-2	A10	Bn-A10-p7118112	8703408	14.56	17Semi, 18Semi
q-C18:3-A10-3	A10	Bn-A10-p16837056	16640509	13.5	16Sp, 18Semi, 18Sp, 18Win
q-C18:3-C01	C01	Bn-scaff_15838_5-p850445	3684748	14.95	16Sp, 17Semi, 18Semi
q-C18:3-C02	C02	Bn-scaff_18675_1-p230717	22324250	15.18	17Win, 16Sp, 18Semi, 18Sp
q-C18:3-C03	C03	Bn-scaff_26505_1-p5590	28729268	14.88	17Win, 18Semi
q-C18:3-C04-1	C04	Bn-scaff_16564_1-p236601	11168988	9.57	17Semi, 18Semi
q-C18:3-C04-2	C04	Bn-scaff_15779_1-p94004	30153296	13.19	16Sp, 17Semi, 18Semi
q-C18:3-C04-3	C04	Bn-scaff_16139_1-p785412	43939995	12.25	17Semi, 18Semi, 18Sp
q-C18:3-C05-1	C05	Bn-scaff_20901_1-p1719394	2295884	5.68	17Semi, 18Semi
q-C18:3-C05-2	C05	Bn-Scaffold000324-p108	8698912	7.29	18Sp, 18Win
q-C18:3-C05-3	C05	Bn-scaff_16454_1-p884909	21537168	9.06	16Sp, 18Sp
q-C18:3-C06-1	C06	Bn-scaff_17454_1-p225095	8428072	7.31	18Semi, 18Sp
q-C18:3-C06-2	C06	Bn-scaff_23957_1-p175042	30652290	15.26	16Sp, 17Semi, 18Semi, 18Sp
q-C18:3-C07-1	C07	Bn-scaff_22310_1-p321188	7983992	8.36	17Semi, 18Sp, 18Semi
q-C18:3-C07-2	C07	Bn-scaff_19106_1-p463047	18601376	10.61	17Semi, 18Semi

(Continued)

TABLE 2 | Continued

QTN	Chr	SNP associated	Position (bp)	$-\log_{10}(P)$	Environment
q-C18:3-C07-3	C07	Bn-scaff_16110_1-p2168404	42696563	10.29	17Semi, 18Win, 18Semi
q-C18:3-C08-1	C08	Bn-scaff_16174_1-p1445094	23687956	6.74	17Semi, 18Semi
q-C18:3-C08-2	C08	Bn-scaff_16445_1-p2523413	34371202	15.11	16Sp, 17Win, 18Sp, 18Semi
q-C18:3-C09-1	C09	Bn-scaff_20903_1-p300819	16920322	8.82	18Sp, 18Semi
q-C18:3-C09-2	C09	Bn-scaff_16297_1-p392549	23679499	6.01	18Semi, 18Sp
q-C18:3-C09-3	C09	Bn-scaff_20972_1-p160691	32990956	13.3	18Sp, 18Semi
q-C20:1-A01-1	A01	Bn-A01-p4167795	3847687	15.18	16Semi, 18Sp
q-C20:1-A01-2	A01	Bn-A01-p26369651	20893064	14.54	17Sp, 18Sp
q-C20:1-A02-1	A02	Bn-A02-p11284285	8292886	10.66	16Sp, 18Sp, 18Win
q-C20:1-A02-2	A02	Bn-A02-p21061002	18906336	13.65	17Sp, 18Sp
q-C20:1-A03-1	A03	Bn-A03-p16565487	15689355	13.20	18Semi, 18Sp
q-C20:1-A03-2	A03	Bn-A03-p27337536	25579664	14.40	16Win, 17Win, 18Sp
q-C20:1-A04	A04	Bn-A04-p14687930	15158346	9.40	17Sp, 18Sp
q-C20:1-A05-1	A05	Bn-A05-p5100352	4920148	7.95	16Win, 18Sp
q-C20:1-A05-2	A05	Bn-A05-p20425452	18636249	14.54	17Sp, 18Sp
q-C20:1-A06-1	A06	Bn-A06-p853722	1082917	6.01	16Semi, 17Sp, 17Win, 18Sp
q-C20:1-A06-2	A06	Bn-A06-p21116438	20562931	6.04	17Sp, 18Sp
q-C20:1-A07	A07	Bn-A07-p20230189	21986980	5.54	16Semi, 18Sp
q-C20:1-A08-1	A08	Bn-A08-p2711497	2151791	8.08	16Semi, 18Sp, 18Semi, 18Win
q-C20:1-A08-2	A08	Bn-A08-p13066424	10878218	13.82	16Sp, 16Semi, 16Win,17Sp, 17Semi, 18Semi, 18Sp
q-C20:1-A09-1	A09	Bn-A09-p26874249	24934319	9.82	18Sp, 18Semi, 18Win
q-C20:1-A09-2	A09	Bn-A09-p35656352	32788000	5.92	17Sp, 17Win
q-C20:1-A10-1	A10	Bn-A10-p13965313	13956813	14.54	16Sp, 17Sp, 18Sp
q-C20:1-C01	C01	Bn-scaff_17827_1-p963588	7866768	9.07	17Sp, 18Sp
q-C20:1-C02	C02	Bn-scaff_16139_1-p1267317	45277206	5.60	16Semi, 17Win
q-C20:1-C03-1	C03	Bn-scaff_17636_1-p3673	38484538	6.81	16Win, 17Win
q-C20:1-C03-2	C03	Bn-scaff_15794_3-p89999	55717350	12.34	16Sp, 16Semi, 16Win, 17Sp, 17Win, 18Win
q-C20:1-C04-1	C04	Bn-scaff_16394_1-p987099	32288653	12.72	17Sp, 17Semi, 18Sp
q-C20:1-C04-2	C04	Bn-scaff_15585_1-p276555	44214027	14.54	17Sp, 18Sp
q-C20:1-C05-1	C05	Bn-scaff_21338_1-p467919	11924522	6.42	16Win, 18Sp
q-C20:1-C05-2	C05	Bn-scaff_22099_1-p251444	24830406	13.97	16Semi, 17Sp, 18Sp
q-C20:1-C05-3	C05	Bn-A07-p541617	36934827	14.54	16Win, 17Sp, 18Sp
q-C20:1-C07	C07	Bn-scaff_16069_1-p431757	36777489	6.14	16Semi, 18Sp, 18Win
q-C20:1-C08	C08	Bn-scaff_21269_1-p121333	36981334	14.40	16Win, 17Sp, 18Sp
q-C20:1-C09-1	C09	Bn-scaff_17174_1-p62030	13393631	14.54	16Semi, 18Sp
q-C20:1-C09-2	C09	Bn-scaff_16456_1-p453404	35037965	15.18	16Win, 17Sp, 18Sp
q-C22:1-A01	A01	Bn-A01-p2825565	2327566	8.45	16Sp, 16Win, 16Semi, 18Sp
q-C22:1-A02-1	A02	Bn-scaff_16565_1-p1062007	6923746	6.36	18Sp, 18Win
q-C22:1-A02-2	A02	Bn-A02-p22386317	20775741	7.59	17Sp, 18Sp
q-C22:1-A03-1	A03	Bn-A03-p1923025	1541003	5.54	17Win, 18Win
q-C22:1-A03-2	A03	Bn-A03-p20417630	19283838	6.46	16Win, 17Semi, 18Sp
q-C22:1-A06	A06	Bn-A06-p21501350	20200573	5.11	16Sp, 18Sp
q-C22:1-A08	A08	Bn-A08-p13066424	10878218	12.38	16Sp,16Win, 16Semi, 17Sp, 17Win, 17Semi, 18Sp, 18Win, 18Semi
q-C22:1-A09-1	A09	Bn-A09-p3029767	2949844	9.43	17Sp, 17Win, 18Win, 18Semi, 18Sp
q-C22:1-A09-2	A09	Bn-A09-p27109839	25110047	5.45	17Sp, 17Win, 18Sp
q-C22:1-A10	A10	Bn-A10-p5819027	5451194	10.32	16Sp, 18Sp
q-C22:1-C02	C02	Bn-scaff_16139_1-p1051077	45458213	5.84	17Semi, 18Semi, 18Win

(Continued)

TABLE 2 | Continued

QTN	Chr	SNP associated	Position (bp)	$-\log_{10}(P)$	Environment
q-C22:1-C03	C03	Bn-scaff_17457_1-p493971	53921047	6.70	16Sp, 16Win, 16Semi, 17Sp, 17Win, 18Win
q-C22:1-C05	C05	Bn-scaff_18181_1-p1691104	6103006	8.95	16Sp, 18Sp
q-C22:1-C07	C07	Bn-scaff_15705_1-p1673044	34945104	6.66	17Win, 18Win
q-C22:1-C08-1	C08	Bn-A08-p6162660	14268534	5.89	17Win, 18Win
q-C22:1-C08-2	C08	Bn-scaff_16361_1-p2793822	30283886	8.57	17Sp, 18Sp

Chr, Chromosome; Sp, Spring-type rapeseed; Win, Winter-type rapeseed; Semi, Semi-winter-type rapeseed; 16, 17, and 18 represent the 2016, 2017, and 2018 growing seasons in Chongqing, China, respectively. QTNs with underline were overlapped with those detected by Qu et al. (2017), and QTNs with bold font were identified in at least two ecotypic rapeseed and/or environments.

by the genotype and environments, and 61 environment-sensitive genes were obtained by comparing these regions with the *B. napus* reference genome (Chalhoub et al., 2014), including *GDSL* (GDSL-like Lipase), *GAPA* (Glyceraldehyde 3-phosphate dehydrogenase A), *KCS21*, *FAD3*, *FAD7*, *FAD6* fatty acid biosynthesis 1 (*FAB1*), acyl-activating enzyme 17 (*AAE17*), long chain acyl-CoA synthetase 9 (*LACS9*), oleosin 2 (*OLEO2*), beta-ketoacyl reductase 1 (*KCR1*), and trigalactosyldiacylglycerol2 (*TGD2*) (**Table 3**). Of these, some genes (e.g., *TT16* and *TT1*) were predicted to be associated with oleic acid content in *B. napus* (Lian et al., 2017; Qu et al., 2017); however, novel loci were also identified, including *MYB67*, *OLEO2*, *KCS21*, *FAD3*, *KCR1*, *TT1*, and *TGD2* (**Table 3**). Among these genes, 12 putative gene members were identified in previous research, and 14 gene members were enriched in fatty acid pathways in the KEGG database (**Table 3**).

Oleic acid is a monounsaturated fat beneficial for human health that contributes to the nutritional and economic value of rapeseed oil. To provide insight into the genetic control of oleic acid content, we therefore aligned 95 gene sequences from the different rapeseed accessions, and identified nucleotides in the intronic regions of *BnaA08g08280D* and *BnaC03g60080D* that show significant differences between the high- and low-oleic acid lines (**Figure 3**).

Expression Patterns of Candidate Genes

We assessed the relative expression levels of the candidate genes during seed development of *B. napus* variety ZS11 (**Figures 4**, **5**), which had a high oleic acid content and low erucic acid. The expression levels of the environment-insensitive genes showed no obvious variation during seed development, but *KCS18* (*BnaA08g11130D* and *BnaC03g65980D*), and *BnaC02g42910D* showed higher expression levels during the middle stages of seed development (**Figure 4**), indicating that they might contribute to the accumulation of oleic acid during the middle stages of seed development. In addition, *TT16* (*BnaA09g05410D* and *BnaC02g42240D*) and *BnaC02g42240D* were expressed at high levels in the early stages of seed development, while other genes showed low expression levels throughout seed development (**Figure 4**).

However, we found that the expression levels of the environment-sensitive genes varied throughout seed development (**Figure 5**). For example, *OLEO2* (*BnaC04g32530D*) was highly expressed in the middle and late stages of seed development, while the expression of *KCS9* (*BnaC07g05570D*) peaked in the early and middle stages (**Figure 5**). In addition, *BnaA01g29500D* and *TT4* (*BnaA02g30320D*) were mainly expressed in the middle stages of seed development, but *KCR1* (*BnaA02g13310D*) and *TTG1* (*BnaC07g29950D*) showed high expression levels throughout seed development (**Figure 5**). Furthermore, other genes displayed different patterns of expression throughout seed development.

DISCUSSION

In *B. napus*, seeds fatty acids are mainly composed of palmitic, stearic, oleic, linoleic, linolenic eicosenoic, and erucic acids, which determine the rapeseed oil quality. Enhancing the oleic acid content and quality of rapeseed through modifying its fatty acid composition has become an important breeding goal. However, previous studies identified the effect of putative fatty acid genes and the interaction of genotype and environment on the fatty acid content of rapeseed (Zhao, 2002; Zhao et al., 2005, 2008; Wen et al., 2015). Here, we report that fatty acid content also varies significantly among different rapeseed ecological types (spring, winter, and semi-winter rapeseed varieties) and environments (**Figure 1**, **Table 1**), indicating the complexity of the biosynthetic processes underlying fatty acid content in rapeseed. Interestingly, accessions with high oleic acid content were common amongst the different rapeseed ecological types (**Figures 1G–I**), possibly because these accessions are artificially selected in breeding projects aimed at producing rapeseed with high oleic acid content. Therefore, identifying the relationship between favorable alleles and environments will be beneficial for improving the fatty acid content of rapeseed.

With the development of genome sequencing and computational technologies, the Illumina Infinium *Brassica* 60K SNP array has been developed and widely used for the genome-wide association analysis of *B. napus* as well as the analysis of some trait-associated genomic regions and candidate genes (Li et al., 2014; Qian et al., 2014; Gajardo et al., 2015; Hatzig et al., 2015; Lee et al., 2015; Wei et al., 2015; Gacek et al., 2017; Qu et al., 2017). Furthermore, the MLM (Q+K and PCA+K) was found to be a powerful model for GWAS in these previous

TABLE 3 | Summary of candidate genes associated with fatty acid biosynthesis in significant association regions.

Gene type	Chr.	Candidate gene	Annotation gene	Function description	Pathway ID and description	References
Environmental insensitive	A06	BnaA06g25470D[b]	AT2G17930	Phosphatidylinositol 3- and 4-kinase family protein with FAT domain		
	A06	BnaA06g25480D[b]	AT2G17930	Phosphatidylinositol 3- and 4-kinase family protein with FAT domain		
	A06	BnaA06g30370D[b]	AT5G48230	acetoacetyl-CoA thiolase 2 (ACAT2)	K00626: Fatty acid metabolism	
	A06	BnaA06g30420D[b]	AT5G48140	Pectin lyase-like superfamily protein		
	A06	BnaA06g30430D[b]	AT5G48100	TRANSPARENT TESTA 10 (TT10)	K05909: laccase	
	A06	BnaA06g30780D[b]	AT3G29670	HXXXD-type acyl-transferase family protein		
	A06	BnaA06g30790D[b]	AT3G29635	HXXXD-type acyl-transferase family protein		
	A06	BnaA06g30800D[b]	AT3G29635	HXXXD-type acyl-transferase family protein		
	A06	BnaA06g31040D[b]	AT3G29152	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein		
	A08	BnaA08g08190D[b]	AT2G44730	Alcohol dehydrogenase transcription factor Myb/SANT-like family protein		
	A08	BnaA08g08280D[b]	AT4G17483	alpha/beta-Hydrolases superfamily protein	K01074: Fatty acid elongation	
	A08	BnaA08g08850D[b]	AT4G18550	alpha/beta-Hydrolases superfamily protein		Kim et al., 2011
	A08	BnaA08g09510D[b]	AT4G20840	FAD-binding Berberine family protein		
	A08	BnaA08g11130D[b]	AT4G34520	3-ketoacyl-CoA synthase 18 (KCS18)	K15397: Fatty acid elongation	Wang et al., 2008; Wu et al., 2008; Li et al., 2014
	A08	BnaA08g11140D[b]	AT4G34510	3-ketoacyl-CoA synthase 17 (KCS17)	K15397: Fatty acid elongation	Tresch et al., 2012
	A08	BnaA08g11440D[b]	AT4G33790	ECERIFERUM 4 (CER4)	K13356: Cutin, suberine and wax biosynthesis	Rowland et al., 2006
	A08	BnaA08g11650D[b]	AT4G34030	3-methylcrotonyl-CoA carboxylase (MCCB)	K01969: 3-methylcrotonyl-CoA carboxylase beta subunit	Ding et al., 2012
	A08	BnaA08g11810D[b]	AT4G33355	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein		
	A09	BnaA09g02110D[a]	AT3G27660	oleosin 4 (OLEO4)		
	A09	BnaA09g05070D[b]	AT5G23970	HXXXD-type acyl-transferase family protein		
	A09	BnaA09g05410D[b]	AT5G23260	TRANSPARENT TESTA16 (TT16)		Deng et al., 2012
	A09	BnaA09g06170D[b]	AT2G25710	holocarboxylase synthase 1 (HCS1)		Tasseva et al., 2004
	A09	BnaA09g50060D[b]	AT1G06090	Fatty acid desaturase family protein		Smith et al., 2013
	A09	BnaA09g50070D[a]	AT1G06090	Fatty acid desaturase family protein		Smith et al., 2013
	A09	BnaA09g50080D[b]	AT1G06090	Fatty acid desaturase family protein		Smith et al., 2013
	C02	BnaC02g42220D[b]	AT5G23280	TCP family transcription factor		Aguilar-Martínez and Sinha, 2013
	C02	BnaC02g42240D[b]	AT5G23260	TRANSPARENT TESTA16 (TT16)		Deng et al., 2012
	C02	BnaC02g42690D[b]	AT5G63770	diacylglycerol kinase 2 (DGK2)	K00901: Glycerolipid metabolism	
	C02	BnaC02g42700D[b]	AT5G63770	Diacylglycerol kinase 2 (DGK2)		

(Continued)

TABLE 3 | Continued

Gene type	Chr.	Candidate gene	Annotation gene	Function description	Pathway ID and description	References
	C02	BnaC02g42910D[b]	AT5G64080	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein		
	C03	BnaC03g60080D[b]	AT4G38620	myb domain protein 4 (MYB4)	K09422: transcription factor MYB, plant	
	C03	BnaC03g65980D[b]	AT4G34520	3-ketoacyl-CoA synthase 18 (KCS18)	K15397: Fatty acid elongation	Wang et al., 2008; Wu et al., 2008; Li et al., 2014
	C03	BnaC03g66040D[b]	AT4G34510	3-ketoacyl-CoA synthase 17 (KCS17)	K15397: Fatty acid elongation	Tresch et al., 2012
	C03	BnaC03g66380D[b]	AT4G33790	ECERIFERUM 4 (CER4)	K13356: Cutin, suberine and wax biosynthesis	Rowland et al., 2006
Environmental sensitive	A01	BnaA01g03850D[a]	AT4G33020	ZIP9		
	A01	BnaA01g08120D[b]	AT4G28780	GDSL-like Lipase/Acylhydrolase superfamily protein		
	A01	BnaA01g26680D[b]	AT3G18570	Oleosin family protein		
	A01	BnaA01g29500D[b]	AT3G14220	GDSL-like Lipase/Acylhydrolase superfamily protein		
	A01	BnaA01g30080D[b]	AT3G13062	Polyketide cyclase/dehydrase and lipid transport superfamily protein		
	A01	BnaA01g30110D[b]	AT3G13040	myb-like HTH transcriptional regulator family protein		
	A01	BnaA01g31150D[b]	AT3G11170	Fatty acid desaturase 7 (FAD7)		Maeda et al., 2008
	A02	BnaA02g13010D[a]	AT1G67260	TCP1		
	A02	BnaA02g13270D[a]	AT1G77590	Long chain acyl-CoA synthetase 9 (LACS9)	K01897: Fatty acid biosynthesis	Jessen et al., 2015
	A02	BnaA02g13310D[a]	AT1G67730	Beta-ketoacyl reductase 1 (KCR1)	K10251: Fatty acid elongation, Biosynthesis of unsaturated fatty acids	
	A02	BnaA02g27960D[b]	AT4G11850	Phospholipase D gamma 1 (PLDGAMMA1)	K01115: Glycerophospholipid metabolism	
	A02	BnaA02g28150D[b]	AT3G26650	glyceraldehyde 3-phosphate dehydrogenase A subunit (GAPA)		
	A02	BnaA02g30320D[b]	AT5G13930	TRANSPARENT TESTA 4 (TT4)	K00660: Flavonoid biosynthesis	
	A02	BnaA02g30340D[b]	AT5G13930	TRANSPARENT TESTA 4 (TT4)	K00660: Flavonoid biosynthesis	
	A02	BnaA02g30560D[b]	AT5G49070	3-ketoacyl-CoA synthase 21 (KCS21)	K15397: Fatty acid elongation	
	A03	BnaA03g02250D[a]	AT5G07870	HXXXD-type acyl-transferase family protein		
	A03	BnaA03g02290D[a]	AT5G07920	Diacylglycerol kinase1 (DGK1)	K00901: Glycerophospholipid metabolism	
	A03	BnaA03g02360D[a]	AT5G60830	basic leucine-zipper 70 (bZIP70)		
	A03	BnaA03g02470D[a]	AT5G08330	TCP family transcription factor		
	A03	BnaA03g03980D[a]	AT5G12420	O-acyltransferase (WSD1-like) family protein		Kalscheuer and Steinbüchel, 2003
	A03	BnaA03g03990D[a]	AT5G12420	O-acyltransferase (WSD1-like) family protein		Kalscheuer and Steinbüchel, 2003
	A03	BnaA03g04000D[a]	AT5G12420	O-acyltransferase (WSD1-like) family protein		Kalscheuer and Steinbüchel, 2003
	A03	BnaA03g13590D[a]	AT2G29980	fatty acid desaturase 3 (FAD3)		Hu et al., 2006; Yang et al., 2012
	A03	BnaA03g31600D[a]	AT3G11170	fatty acid desaturase 7 (FAD7)		
	A03	BnaA03g39010D[b]	AT4G34520	3-ketoacyl-CoA synthase 18 (KCS18)		

(Continued)

TABLE 3 | Continued

Gene type	Chr.	Candidate gene	Annotation gene	Function description	Pathway ID and description	References
	A03	BnaA03g39500D[b]	AT5G23260	TRANSPARENT TESTA16 (TT16)		
	A03	BnaA03g49040D[b]	AT4G28130	diacylglycerol kinase 6 (DGK6)	K00901:Glycerolipid metabolism	
	A05	BnaA05g00690D[a]	AT2G47210	myb-like transcription factor family protein		
	A05	BnaA05g09070D[a]	AT2G34770	fatty acid hydroxylase 1 (FAH1)		Nagano et al., 2012
	A05	BnaA05g25210D[b]	AT5G40990	GDSL lipase 1 (GLIP1)		
	A05	BnaA05g25220D[b]	AT3G14225	GDSL-motif lipase 4 (GLIP4)		
	A07	BnaA07g14020D[a]	AT2G28630	3-ketoacyl-CoA synthase 12 (KCS12)		Kim et al., 2013
	A08	BnaA08g12780D[b]	AT4G30950	fatty acid desaturase 6 (FAD6)	K10255:Two-component system	
	A08	BnaA08g12800D[b]	AT4G30950	fatty acid desaturase 6 (FAD6)		
	A08	BnaA08g14190D[b]	AT4G27030	fatty acid desaturase A (FADA)		
	A08	BnaA08g14200D[b]	AT4G27030	fatty acid desaturase A (FADA)		
	A09	BnaA09g10550D[a]	AT1G62640	3-ketoacyl-acyl carrier protein synthase III (KAS III)	K00648:Fatty acid biosynthesis	Katayoon et al., 2001
	A09	BnaA09g10680D[b]	AT1G62940	acyl-CoA synthetase 5 (ACOS5)		
	A10	BnaA10g19670D[b]	AT5G13930	TRANSPARENT TESTA 4 (TT4)		
	A10	BnaA10g24560D[b]	AT5G05960	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein		
	C01	BnaC01g12360D[a]	AT4G20870	fatty acid hydroxylase 2 (FAH2)		
	C01	BnaC01g22230D[a]	AT5G49555	FAD/NAD(P)-binding oxidoreductase family protein		
	C01	BnaC01g23310D[a]	AT3G51590	lipid transfer protein 12 (LTP12)		
	C04	BnaC04g27640D[b]	AT3G53100	GDSL-like Lipase/Acylhydrolase superfamily protein		
	C04	BnaC04g32530D[b]	AT5G40420	oleosin 2 (OLEO2)		
	C05	BnaC05g02350D[a]	AT1G04220	3-ketoacyl-CoA synthase 2 (KCS2)		
	C05	BnaC05g04810D[a]	AT2G30310	GDSL-like Lipase/Acylhydrolase family protein		
	C05	BnaC05g04820D[a]	AT2G24560	GDSL-like Lipase/Acylhydrolase family protein		
	C05	BnaC05g14920D[a]	AT1G74960	fatty acid biosynthesis 1 (FAB1)	K09458: Fatty acid biosynthesis	
	C05	BnaC05g36500D[a]	AT3G16850	Pectin lyase-like superfamily protein		
	C06	BnaC06g08390D[a]	AT1G34790	transparent testa 1 (TT1)		Lian et al., 2017
	C07	BnaC07g05570D[a]	AT2G16280	3-ketoacyl-CoA synthase 9 (KCS9)	K15397: Fatty acid elongation	Kim et al., 2013
	C07	BnaC07g29950D[b]	AT5G24520	TRANSPARENT TESTA GLABRA 1 (TTG1)		
	C07	BnaC07g30210D[b]	AT5G24180	Lipase class 3-related protein		
	C07	BnaC07g41010D[a]	AT3G05970	long-chain acyl-CoA synthetase 6 (LACS6)	K01897:Fatty acid biosynthesis	Hsiao et al., 2014
	C07	BnaC07g41430D[a]	AT4G28570	Long-chain fatty alcohol dehydrogenase family protein		
	C07	BnaC07g42880D[a]	AT4G30720	FAD/NAD(P)-binding oxidoreductase family protein		
	C08	BnaC08g32310D[b]	AT3G62590	Alpha/beta-Hydrolases superfamily protein		
	C09	BnaC09g05650D[a]	AT5G62470	myb domain protein 96 (MYB96)		
	C09	BnaC09g20440D[a]	AT2G01180	Phosphatidic acid phosphatase 1 (PAP1)		
	C09	BnaC09g30320D[b]	AT5G54060	UDP-glucose:flavonoid 3-o-glucosyltransferase (UF3GT)	K17193: Anthocyanin biosynthesis	

Chr, Chromosome.

[a,b] *The candidate genes for fatty acid content around the isolated and overlapped QTNs, respectively.*

studies (Yu et al., 2006; Zhao et al., 2011; Xu et al., 2015; Li et al., 2016; Qu et al., 2017). In the present study, the mrMLM was employed for a GWAS, which is confirmed as a precise and powerful tool for analyzing phenotypic and genotypic information derived from numerous accessions and SNPs (Wang et al., 2016; Li et al., 2017). In the present study, 149 QTNs significantly associated with fatty acid content were identified using the mrMLM (**Table 2**, and **Supplementary Table S3**). Of these, 34 associated SNPs overlapped with those obtained using MLM (Qu et al., 2017), indicating that the association analysis was reliable; however, eight novel association regions containing 35 QTNs were simultaneously detected among different ecotypic rapeseed grown in different environments, strongly suggesting that mrMLM is more powerful to detect SNPs associated with complex traits than MLM in GWAS. In addition, 29 QTNs for fatty acids were simultaneously detected in spring, winter and semi-winter rapeseed (**Figure 2**), indicating that the orthologous genes for fatty acid might be better identified using these singinficant QTNs. Furthermore, more QTNs associated with fatty acids were identified from the sping and semi-winter type than in winter rapeseed (**Figure 2**), indicating that the fatty acids were associated with their genotype. These results might be helpful for elucidating the mechanism that determines fatty acid composition in *B. napus*.

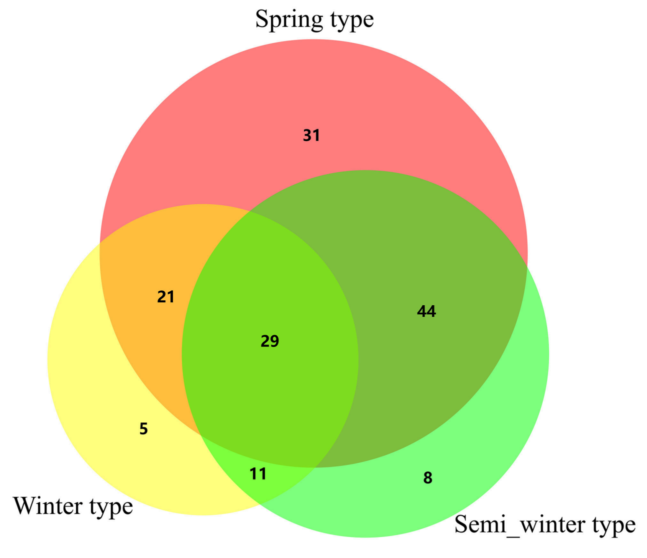
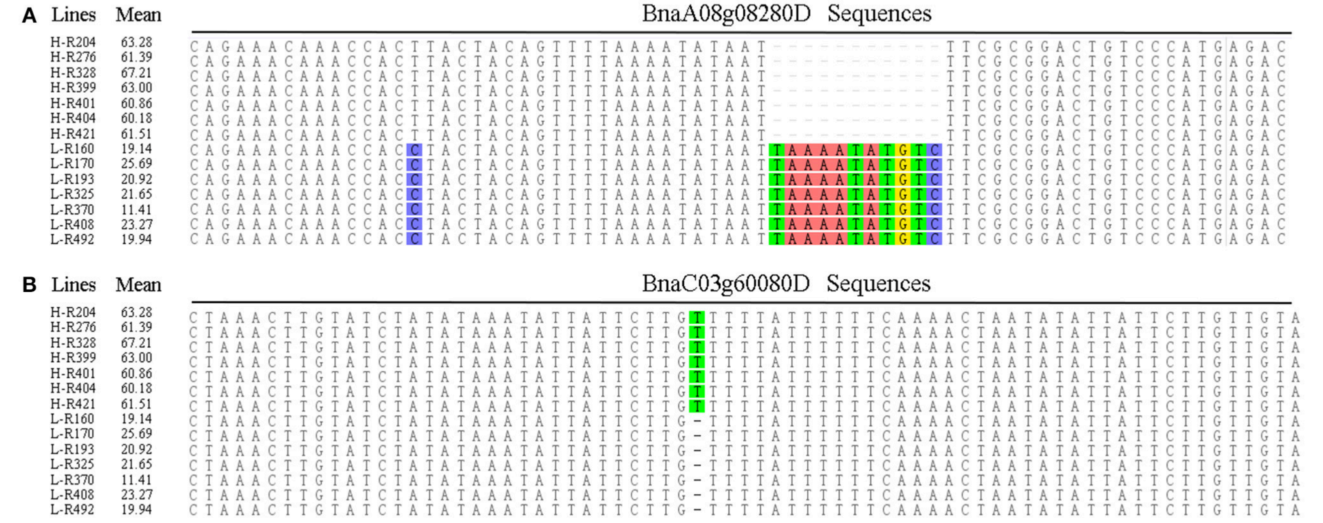


FIGURE 2 | Venn diagram analysis of QTNs for fatty acids in different ecotypic rapeseed.

In *B. napus*, fatty acid variation is controlled by multiple genes (Zhao et al., 2008; Wen et al., 2015). In this study, we categorized the candidate genes as either environment-insensitive or -sensitive genes, according to the published results and their detection frequency between the rapeseed genotypes grown in the different environments. A total of 95 candidate genes were identified with known functions in the fatty acid biosynthesis pathway. Of these, 34 were environment-insensitive genes (**Table 3**), including *KCS18*, which is known to play a crucial role in regulating erucic acid biosynthesis in *B. napus* (Wang et al., 2008; Wu et al., 2008; Li et al., 2014), and the putative *KCS18* paralogs *BnaA08g11130D* (Chromosome A08) and *BnaC03g65980D* (Chromosome C03), which are most highly expressed during the middle to late stages of seed development (**Figure 4**), suggesting that they are key genes regulating the accumulation of fatty acids in rapeseed oil. In addition, *CER4* encodes an alcohol-forming fatty acyl-coenzyme A reductase (Rowland et al., 2006), and *KCS17* is known to be involved in the biosynthesis of saturated fatty acids (Tresch et al., 2012) and would therefore be expected


```
A Lines   Mean   BnaA08g08280D Sequences
  H-R204  63.28  CAGAAACAAACCACTTACTACAGTTTTAAAATATAAT-----------TTCGCGGACTGTCCCATGAGAC
  H-R276  61.39  CAGAAACAAACCACTTACTACAGTTTTAAAATATAAT-----------TTCGCGGACTGTCCCATGAGAC
  H-R328  67.21  CAGAAACAAACCACTTACTACAGTTTTAAAATATAAT-----------TTCGCGGACTGTCCCATGAGAC
  H-R399  63.00  CAGAAACAAACCACTTACTACAGTTTTAAAATATAAT-----------TTCGCGGACTGTCCCATGAGAC
  H-R401  60.86  CAGAAACAAACCACTTACTACAGTTTTAAAATATAAT-----------TTCGCGGACTGTCCCATGAGAC
  H-R404  60.18  CAGAAACAAACCACTTACTACAGTTTTAAAATATAAT-----------TTCGCGGACTGTCCCATGAGAC
  H-R421  61.51  CAGAAACAAACCACTTACTACAGTTTTAAAATATAAT-----------TTCGCGGACTGTCCCATGAGAC
  L-R160  19.14  CAGAAACAAACCACCTACTACAGTTTTAAAATATAATTAAAATATGTCTTCGCGGACTGTCCCATGAGAC
  L-R170  25.69  CAGAAACAAACCACCTACTACAGTTTTAAAATATAATTAAAATATGTCTTCGCGGACTGTCCCATGAGAC
  L-R193  20.92  CAGAAACAAACCACCTACTACAGTTTTAAAATATAATTAAAATATGTCTTCGCGGACTGTCCCATGAGAC
  L-R325  21.65  CAGAAACAAACCACCTACTACAGTTTTAAAATATAATTAAAATATGTCTTCGCGGACTGTCCCATGAGAC
  L-R370  11.41  CAGAAACAAACCACCTACTACAGTTTTAAAATATAATTAAAATATGTCTTCGCGGACTGTCCCATGAGAC
  L-R408  23.27  CAGAAACAAACCACCTACTACAGTTTTAAAATATAATTAAAATATGTCTTCGCGGACTGTCCCATGAGAC
  L-R492  19.94  CAGAAACAAACCACCTACTACAGTTTTAAAATATAATTAAAATATGTCTTCGCGGACTGTCCCATGAGAC

B Lines   Mean   BnaC03g60080D Sequences
  H-R204  63.28  CTAAACTTGTATCTATATAAATATTATTCTTGTTTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  H-R276  61.39  CTAAACTTGTATCTATATAAATATTATTCTTGTTTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  H-R328  67.21  CTAAACTTGTATCTATATAAATATTATTCTTGTTTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  H-R399  63.00  CTAAACTTGTATCTATATAAATATTATTCTTGTTTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  H-R401  60.86  CTAAACTTGTATCTATATAAATATTATTCTTGTTTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  H-R404  60.18  CTAAACTTGTATCTATATAAATATTATTCTTGTTTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  H-R421  61.51  CTAAACTTGTATCTATATAAATATTATTCTTGTTTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  L-R160  19.14  CTAAACTTGTATCTATATAAATATTATTCTTG-TTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  L-R170  25.69  CTAAACTTGTATCTATATAAATATTATTCTTG-TTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  L-R193  20.92  CTAAACTTGTATCTATATAAATATTATTCTTG-TTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  L-R325  21.65  CTAAACTTGTATCTATATAAATATTATTCTTG-TTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  L-R370  11.41  CTAAACTTGTATCTATATAAATATTATTCTTG-TTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  L-R408  23.27  CTAAACTTGTATCTATATAAATATTATTCTTG-TTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
  L-R492  19.94  CTAAACTTGTATCTATATAAATATTATTCTTG-TTTTATTTTTTCAAAACTAATATATTATTCTTGTTGTA
```

FIGURE 3 | Multiple alignments of candidate gene sequences between the high- and low-oleic acid *B. napus* accessions. **(A)** BnaA08g08280D; **(B)** BnaC03g60080D. H-number and L-number indicates the high- and low-oleic acid *B. napus* accessions, respectively. The oleic aicd contents represents by the mean values.

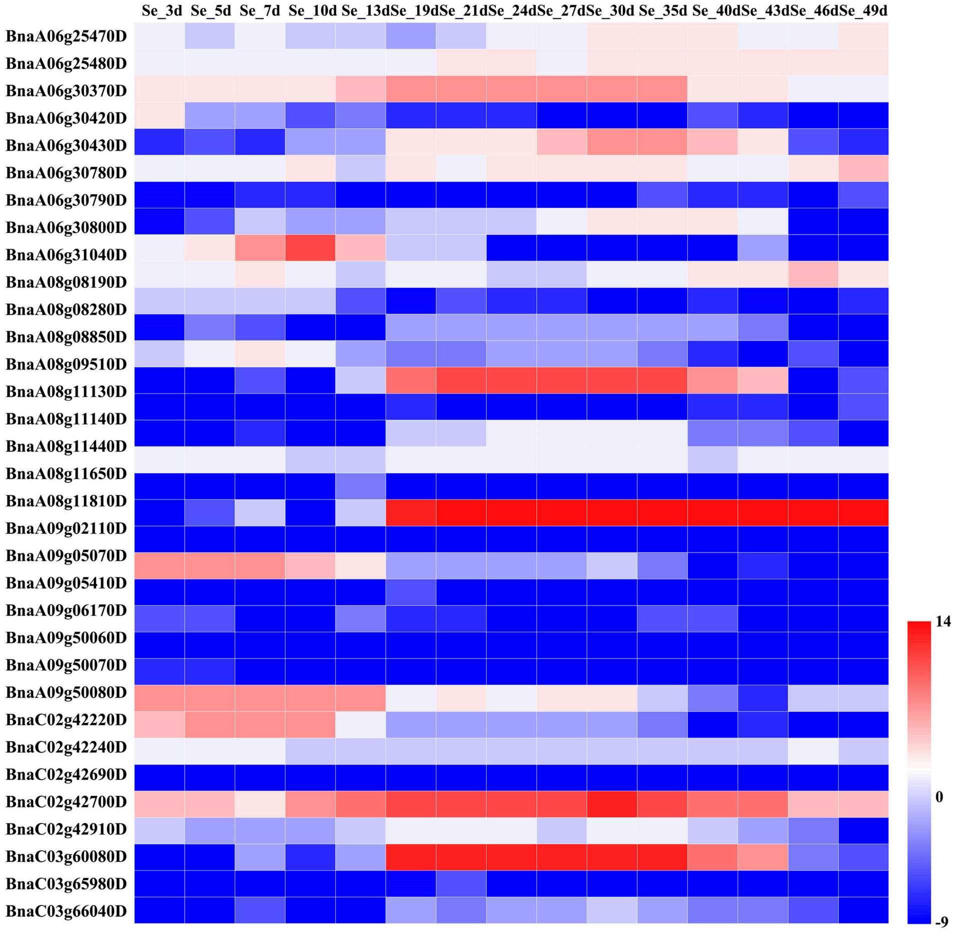


FIGURE 4 | Comparative expression analysis of environment-insensitive genes associated with fatty acid content during seed development. The abbreviations above the heatmap indicate the different developmental stages of the seeds from *B. napus* ZS11 (defined in **Supplementary Table S4**). The expression values of the candidate genes were calculated using three biological replicates with three technical replicates and normalized by Log_2 (mean expression values). The "scale" function in R was used to normalize the relative expression levels ($R = \log_2$/mean expression values). The heatmap was generated using Heatmap Illustrator 1.0 (Heml 1.0).

to be more highly expressed during seed development. We found that the putative orthologs of *KCS17* (*BnaA08g11140D* and *BnaC03g66040D*) and *CER4* (BnaC03g66380D) were associated with fatty acid content in *B. napus*, but the expression of *BnaA08g11140D* and *BnaC03g66040D*, putative *KCS17* paralogs, was downregulated during seed development (**Figure 4**), suggesting that functional segregation exists among the different paralogs of the ancestral *KCS17* gene, which has been reported previously in the *B. napus* genome (Chalhoub et al., 2014). In addition, the environment-insensitive genes, including *DGK2* (*BnaC02g42690D* and *BnaC02g42700D*), *HCS1* (*BnaA09g06170D*), *MYB4* (*BnaC03g60080D*), and *TT16* (*BnaA09g05410D* and *BnaC02g42240D*), might also been involved in fatty acid biosynthesis (Tasseva et al., 2004; Deng et al., 2012; Yang et al., 2012; Chen et al., 2013). Most of the environment-insensitive genes were steadily expressed throughout seed development (**Figure 4**), and these might be the major factors regulating oleic acid accumulation in *B. napus*. Furthermore, 12 environment-insensitive genes were enriched in the fatty acid biosynthesis pathway according to KEGG pathway analysis (**Table 3**). Importantly, the nucleotide sequences of *BnaA08g08280D* and *BnaC03g60080D* differed between the high- and low-oleic acid lines (**Figure 3**), but these differences were all located in the intronic regions. Furthermore, 61 environment-sensitive genes showed wide variation in expression during seed development (**Table 3**, **Figure 5**), and these could be divided into early, middle, and late expression genes, respectively. For example, *BnaA06g31040D* showed high expression levels in the early stages of seed development, *KCS9* (*BnaC07g05570D*) peaked at the early and middle stages, *OLEO2* (*BnaC04g32530D*) had high expression during the middle and late stages, and *BnaA01g29500D*, and *TT4* (*BnaA02g30320D*) expression markedly increased in the middle stages (**Table 3**, **Figure 5**). Several other candidate genes, including putative homologs of *FAD* (3, 6, and 7), *KCS* (9, 12, and 21), *KCR1*, and *LACS9*, were found to be associated with fatty acid biosynthesis

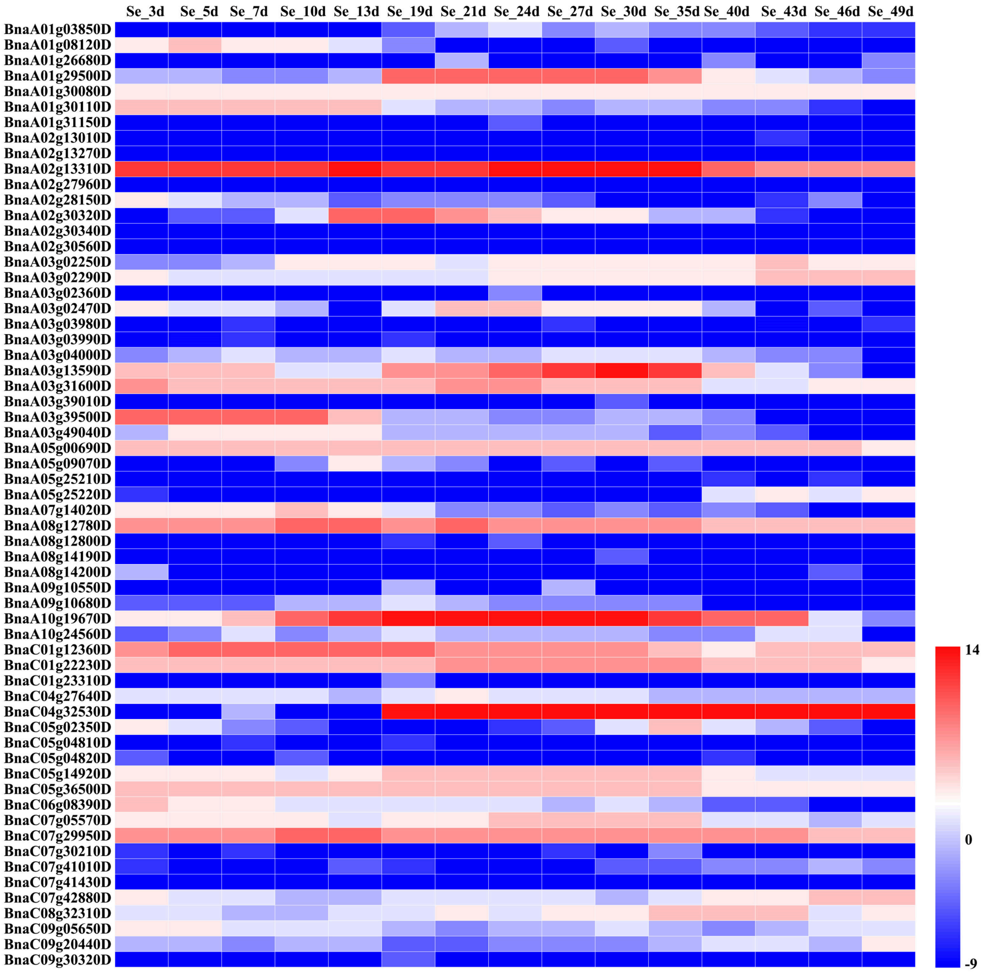


FIGURE 5 | Comparative expression analysis of environment-sensitive genes associated with the fatty acid content during seed development. The abbreviations above the heatmap indicate the different developmental stages of the seeds from *B*. *napus* ZS11 (defined in **Supplementary Table S5**). The expression values of the candidate genes were calculated using three biological replicates with three technical replicates and normalized by Log_2 (mean expression values). The "scale" function in R was used to normalize the relative expression levels ($R = log_2$/mean expression values). The heatmap was generated using Heatmap Illustrator 1.0 (Heml 1.0).

(Peng et al., 2010; Tresch et al., 2012; Yang et al., 2012; Lai et al., 2017; Shi et al., 2017), but their contribution remains to be confirmed by further studies (**Table 3**). In addition, 14 gene members were involved in the fatty acid metabolism confirmed by KEGG database analysis (**Table 3**). However, there is no clear evidence indicating that these genes control the fatty acid content in *B. napus*.

In summary, 149 QTNs for fatty acid content (including 34 reported and 115 novel loci) were detected, strongly demonstrating that mrMLM is a powerful and suitable tool for detecting QTNs for fatty acid content in rapeseed. Among these putative candidate genes, 63.16% (60/95) and 36.84% (35/95) were distributed on overlapping and isolated QTNs, respectively. Based on the pervious reports, 29 genes are involved in the fatty acid biosynthesis, and 26 gene members were enriched in the fatty acid pathway by the KEGG pathway database, indicating that mrMLM is an accurate tool to estimate the effect of QTNs on complex traits. Whether these genes exert significant regulatory effects on the fatty acid content of the seeds remains to be investigated. Hence, further studies are needed. Our findings provide useful candidate genes for the marker-assisted selection and breeding of rapeseed lines with increased oleic acid content in the seed.

AUTHOR CONTRIBUTIONS

CMQ and JL conceived and designed the experiments; MG and XH conducted the experiments; ZX, XX, LJ, and KL collected and analyzed the data; SW, and LJ made sequence alignment; YL, LW, and RW carried out the field experiments; MG, MZ, and CLQ wrote the manuscript; JL and CMQ reviewed the manuscript.

FUNDING

This work was supported by the National Key Research and Development Plan (2016YFD0101007 and 2016YFD0100202), the Department of Agriculture projects of modern agricultural technology system (CARS-12), National Science Foundation of China (31401412, 31571701), the 111 Project (B12006), Chongqing Basic Scientific and Advanced Technology Research (cstc2016shms-ztzx80010, cstc2015jcyjBX0001 and cstc2017jcyjAX0321), and Fundamental Research Funds for the Central Universities (XDJK2016B030).

ACKNOWLEDGMENTS

We would also like to thank Kathy Farquharson for critical reading of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01872/full#supplementary-material

Supplementary Figure S1 | Subpopulations in the spring, winter, and semi-winter *B. napus* accessions, determined using a principal coordinate analysis (PCA).

Supplementary Table S1 | List of rapeseed accessions used.

Supplementary Table S2 | Sum of SNPs in candidate regions for fatty acid content in *B. napus* accessions grown in different environments.

Supplementary Table S3 | Primers used to amplify candidate genes and reference genes via RT-qPCR analysis.

Supplementary Table S4 | Heatmap of the expression levels of the *B. napus* environment-insensitive genes at different stages of seed development.

Supplementary Table S5 | Heatmap of the expression levels of the *B. napus* environment-sensitive genes at different stages of seed development.

REFERENCES

Aguilar-Martínez, J. A., and Sinha, N. (2013). Analysis of the role of *Arabidopsis* class I TCP genes *AtTCP7*, *AtTCP8*, *AtTCP22*, and *AtTCP23* in leaf development. *Front. Plant Sci.* 4:406. doi: 10.3389/fpls.2013.00406

Bauer, B., Kostik, V., and Gjorgjeska, B. (2015). Fatty acid composition of seed oil obtained from different canola varieties. *Farm. Glas.* 71, 1–7.

Burns, M., Barnes, S., Bowman, J., Clarke, M., Werner, C., and Kearsey, M. (2003). QTL analysis of an intervarietal set of substitution lines in *Brassica napus*: (i) Seed oil content and fatty acid composition. *Heredity* 90, 39–48. doi: 10.1038/sj.hdy.6800176

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345, 950–953. doi: 10.1126/science.1253435

Chang, F., Guo, C., Sun, F., Zhang, J., Wang, Z., Kong, J., et al. (2018). Genome-wide association studies for dynamic plant height and number of nodes on the main stem in summer sowing soybeans. *Front. Plant Sci.* 9:e1184. doi: 10.3389/fpls.2018.01184

Chang, N. W., and Huang, P. C. (1998). Effects of the ratio of polyunsaturated and monounsaturated fatty acid to saturated fatty acid on rat plasma and liver lipid concentrations. *Lipids* 33, 481–487. doi: 10.1007/s11745-998-0231-9

Chen, F., Zhang, W., Yu, K., Sun, L., Gao, J., Zhou, X., et al. (2018). Unconditional and conditional QTL analyses of seed fatty acid composition in *Brassica napus* L. *BMC Plant Biol.* 18:49. doi: 10.1186/s12870-018-1268-7

Chen, X., Chou, H.-H., and Wurtele, E. S. (2013). Holocarboxylase synthetase 1 physically interacts with histone H3 in *Arabidopsis*. *Scientifica* 2013:983501. doi: 10.1155/2013/983501

Clarke, W. E., Higgins, E. E., Plieske, J., Wieseke, R., Sidebottom, C., Khedikar, Y., et al. (2016). A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theor. Appl. Genet.* 129, 1887–1899. doi: 10.1007/s00122-016-2746-7

Delourme, R., Falentin, C., Fomeju, B. F., Boillot, M., Lassalle, G., André, I., et al. (2013). High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus* L. *BMC Genomics* 14:120. doi: 10.1186/1471-2164-14-120

Deng, W., Chen, G., Peng, F., Truksa, M., Snyder, C. L., and Weselake, R. J. (2012). *Transparent testa 16* plays multiple roles in plant development and is involved in lipid synthesis and embryo development in canola. *Plant Physiol.* 160, 978–989. doi: 10.1104/pp.112.198713

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, 562–577. doi: 10.1111/j.1365-313X.2011.04893.x

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587. doi: 10.3410/f.1015548.197423

Fei, W., Wu, X., Li, Q., Chen, F., Hou, S., Fan, Z., et al. (2012). Genetic analysis of high oleic acid mutation materials in *Brassica napus*. *Chin. Agric. Sci. Bull.* 28, 176–180. doi: 10.3969/j.issn.1000-6850.2012.01.034

Gacek, K., Bayer, P. E., Bartkowiakbroda, I., Szala, L., Bocianowski, J., Edwards, D., et al. (2017). Genome-wide association study of genetic control of seed fatty acid biosynthesis in *Brassica napus*. *Front. Plan. Sci.* 7:e2062. doi: 10.3389/fpls.2016.02062

Gajardo, H. A., Wittkop, B., Soto-Cerda, B., Higgins, E. E., Parkin, I. A., Snowdon, R. J., et al. (2015). Association mapping of seed quality traits in *Brassica napus* L. *using GWAS and candidate QTL approaches. Mol. Breeding* 35, 1–19. doi: 10.1007/s11032-015-0340-3

Hardy, O. J., and Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* 2, 618–620. doi: 10.1046/j.1471-8286.2002.00305.x

Hatzig, S. V., Frisch, M., Breuer, F., Nesi, N., Ducournau, S., Wagner, M. H., et al. (2015). Genome-wide association mapping unravels the genetic control of seed germination and vigor in *Brassica napus*. *Front. Plant Sci.* 6, 221. doi: 10.3389/fpls.2015.00221

Hsiao, A. S., Haslam, R. P., Michaelson, L. V., Liao, P., Napier, J. A., and Chye, M. L. (2014). Gene expression in plant lipid metabolism in Arabidopsis seedlings. *PLoS ONE* 9:e107372. doi: 10.1371/journal.pone.0107372

Hu, X., Sullivan-Gilbert, M., Gupta, M., and Thompson, S. A. (2006). Mapping of the loci controlling oleic and linolenic acid contents and development of *fad2* and *fad3* allele-specific markers in canola (*Brassica napus* L.). *Theor. Appl. Genet.* 113, 497–507. doi: 10.1007/s00122-006-0315-1

Javed, N., Geng, J., Tahir, M., Mcvetty, P. B. E., Li, G., and Duncan, R. W. (2016). Identification of QTL influencing seed oil content, fatty acid profile and days to flowering in *Brassica napus* L. *Euphytica* 207, 191–211. doi: 10.1007/s10681-015-1565-2

Jessen, D., Roth, C., Wiermer, M., Fulda, M. (2015). Two activities of *Long-Chain Acyl-Coenzyme A Synthetase* are involved in lipid trafficking between the endoplasmic reticulum and the plastid in *Arabidopsis*. *Plant Physiol.* 167, 351–366. doi: 10.1104/pp.114.250365

Kalscheuer, R., and Steinbüchel, A. (2003). A novel bifunctional wax ester synthase/acyl-CoA:diacylglycerol acyltransferase mediates wax ester and triacylglycerol biosynthesis in *Acinetobacter calcoaceticus* ADP1. *J. Biol. Chem.* 278, 8075–8082. doi: 10.1074/jbc.M210533200

Katayoon, D., Patricia, E., and James, B. (2001). Overexpression of 3-ketoacyl-acyl-carrier protein synthase IIIs in plants reduces the rate of lipid synthesis 1. *Plant Physiol.* 125, 1103–1114. doi: 10.1104/pp.125.2.1103

Kim, E. Y., Seo, Y. S., and Kim, W. T. (2011). *AtDSEL*, an *Arabidopsis* cytosolic DAD1-like acylhydrolase, is involved in negative regulation of storage oil mobilization during seedling establishment. *J. Plant Physiol.* 168, 1705–1709. doi: 10.1016/j.jplph.2011.03.004

Kim, J., Jung, J. H., Lee, S. B., Go, Y. S., Kim, H. J., Cahoon, R., et al. (2013). *Arabidopsis* 3-ketoacyl-coenzyme a synthase 9 is involved in the synthesis of tetracosanoic acids as precursors of cuticular waxes, suberins, sphingolipids, and phospholipids. *Plant Physiol.* 162, 567–580. doi: 10.1104/pp.112.210450

Lai, C. P., Huang, L. M., Chen, L. F. O., Chan, M. T., and Shaw, J. F. (2017). Genome-wide analysis of GDSL-type esterases/lipases in *Arabidopsis*. *Plant Mol. Biol.* 95, 181–197. doi: 10.1007/s11103-017-0648-y

Lee, S., Jang, M. S., Jeon, E. J., Yun, K. Y., and Kim, S. (2015). "QTL Analysis for erucic acid and oleic acid content in *Brassica napus* using F2 population," in: *Plant and Animal Genome XXIII Conference* (San Diego, CA), 823.

Li, F., Chen, B., Xu, K., Gao, G., Yan, G., Qiao, J., et al. (2016). A genome-wide association study of plant height and primary branch number in Rapeseed (*Brassica napus*). *Plant Sci.* 242, 169–177. doi: 10.1016/j.plantsci.2015.05.012

Li, F., Chen, B., Xu, K., Wu, J., Song, W., Bancroft, I., et al. (2014). Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (*Brassica napus* L.). *DNA Res.* 21, 355–367. doi: 10.1093/dnares/dsu002

Li, H., Zhang, L., Hu, J., Zhang, F., Chen, B., Xu, K., et al. (2017). Genome-wide association mapping reveals the genetic control underlying branch angle in rapeseed (*Brassica napus* L.). *Front. Plant Sci.* 8:1054. doi: 10.3389/fpls.2017.01054

Lian, J., Lu, X., Yin, N., Ma, L., Jing, L., Xue, L., et al. (2017). Silencing of *BnTT1* family genes affects seed flavonoid biosynthesis and alters seed fatty acid composition in *Brassica napus. Plant Sci.* 254, 32–47. doi: 10.1016/j.plantsci.2016.10.012

Liu, L. Z., and Li, J. N. (2014). QTL mapping of oleic acid, linolenic acid and erucic acid content in *Brassica napus* by using the high density snp genetic map. *Sci. Agric. Sin.* 47, 24–32. doi: 10.3864/j.issn.0578-1752.2014.01.003

Liu, S., Fan, C., Li, J., Cai, G., Yang, Q., Wu, J., et al. (2016). A genome-wide association study reveals novel elite allelic variations in seed oil content of *Brassica napus*. *Theor. Appl. Genet.* 129, 1203–1215. doi: 10.1007/s00122-016-2697-z

Lu, G., Harper, A. L., Trick, M., Morgan, C., Fraser, F., O'neill, C., et al. (2014). Associative transcriptomics study dissects the genetic architecture of seed glucosinolate content in *Brassica napus*. *DNA Res.* 21, 613–625. doi: 10.1093/dnares/dsu024

Lu, K., Li, T., He, J., Chang, W., Zhang, R., Liu, M., et al. (2017). qPrimerDB: a thermodynamics-based gene-specific qPCR primer database for 147 organisms. *Nucleic Acids Res.* 46, D1229–D1236. doi: 10.1093/nar/gkx725

Luo, X., Ma, C., Yue, Y., Hu, K., Li, Y., Duan, Z., et al. (2015). Unravelling the complex trait of harvest index in rapeseed (*Brassica napus* L.) with association mapping. *BMC Genomics* 16:379. doi: 10.1186/s12864-015-1607-0

Ma, L., Liu, M., Yan, Y., Qing, C., Zhang, X., Zhang, Y., et al. (2018). Genetic dissection of maize embryonic callus regenerative capacity using multi-locus genome-wide association studies. *Front. Plant Sci.* 9:561. doi: 10.3389/fpls.2018.00561

Maeda, H., Sage, T. L., Isaac, G., Welti, R., and Dellapenna, D. (2008). Tocopherols modulate extraplastidic polyunsaturated fatty acid metabolism in *Arabidopsis* at low temperature. *Plant Cell* 20, 452–470. doi: 10.1105/tpc.107.054718

McCouch, S. R., Cho, Y. G., Yano, M., Paul, E., Blinstrub, M., Morishima, H., et al. (1997). Report on QTL nomenclature. *Rice Genet. Newsl.* 14, 11–13. doi: 10.1007/s10142-013-0328-1

Merk, H. L., Yarnes, S. C., Deynze, V., Tong, N., Menda, N., Mueller, L. A., et al. (2012). Trait diversity and potential for selection indices based on variation among regionally adapted processing tomato germplasm. *J. Am. Soc. Hort. Sci.* 137, 427–437.

Meuwissen, T., and Goddard, M. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* 155, 421–430.

Miller, J. F., Zimmerman, D. C., and Vick, B. A. (1987). Genetic control of high oleic acid content in sunflower oil. *Crop Sci.* 27, 923–926. doi: 10.2135/cropsci1987.0011183X002700050019x

Nagano, M., Takahara, K., Fujimoto, M., Tsutsumi, N., Uchimiya, H., and Kawai-Yamada, M. (2012). *Arabidopsis* sphingolipid fatty acid 2-hydroxylases (*AtFAH1* and *AtFAH2*) are functionally differentiated in fatty acid 2-hydroxylation and stress responses. *Plant Physiol.* 159, 1138–1148. doi: 10.1104/pp.112.199547

Peng, Q., Hu, Y., Wei, R., Zhang, Y., Guan, C., Ruan, Y., et al. (2010). Simultaneous silencing of *FAD2* and *FAE1* genes affects both oleic acid and erucic acid contents in *Brassica napus* seeds. *Plant Cell Rep.* 29, 317–325. doi: 10.1007/s00299-010-0823-y

Peng, Y., Liu, H., Chen, J., Shi, T., Zhang, C., Sun, D., et al. (2018). Genome-wide association studies of free amino acid levels by six multi-locus models in bread wheat. *Front. Plant Sci.* 9:1196. doi: 10.3389/fpls.2018.01196

Pleines, S., and Friedt, W. (1989). Genetic control of linolenic acid concentration in seed oil of rapeseed (*Brassica napus* L.). *Theor. Appl. Genet.* 78, 793–797. doi: 10.1007/BF00266660

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1111/j.1471-8286.2007.01758.x

Qian, L., Qian, W., and Snowdon, R. J. (2014). Sub-genomic selection patterns as a signature of breeding in the allopolyploid *Brassica napus* genome. *BMC Genomics* 15:1. doi: 10.1186/1471-2164-15-1170

Qu, C., Fu, F., Liu, M., Zhao, H., Liu, C., Li, J., et al. (2015). Comparative transcriptome analysis of recessive male sterility (RGMS) in sterile and fertile *Brassica napus* Lines. *PLoS ONE* 10:e0144118. doi: 10.1371/journal.pone.0144118

Qu, C., Li, J., Fu, F., Zhao, H., Lu, K., Wei, L., et al. (2017). Genome-wide association mapping and Identification of candidate genes for fatty acid composition in *Brassica napus* L. *using SNP markers. BMC Genomics* 18:232. doi: 10.1186/s12864-017-3607-8

Rowland, O., Zheng, H., Hepworth, S. R., Lam, P., Jetter, R., and Kunst, L. (2006). *CER4* encodes an alcohol-forming fatty acyl-coenzyme a reductase involved in cuticular wax production in *Arabidopsis*. *Plant Physiol.* 142, 866–877. doi: 10.1104/pp.106.086785

Saeidnia, S., and Gohari, A. R. (2012). Importance of *Brassica napus* as a medicinal food plant. *J. Med. Plants Res.* 6, 2700–2703. doi: 10.5897/JMPR11.1103

Scheffler, J. A., Sharpe, A. G., Schmidt, H., Sperling, P., Iap, P., Luhs, W., et al. (1997). Desaturase multigene families of *Brassica napus* arose through genome duplication. *Theor. Appl. Genet.* 94, 583–591. doi: 10.1007/s001220050454

Shi, J., Lang, C., Wang, F., Wu, X., Liu, R., Zheng, T., et al. (2017). Depressed expression of *FAE1* and *FAD2* genes modifies fatty acid profiles and storage compounds accumulation in *Brassica napus* seeds. *Plant Sci.* 263, 177–182. doi: 10.1016/j.plantsci.2017.07.014

Smith, M. A., Dauk, M., Ramadan, H., Yang, H., Seamons, L. E., Haslam, R. P., et al. (2013). Involvement of *Arabidopsis* ACYL-COENZYME A DESATURASE-LIKE2 (*At2g31360*) in the biosynthesis of the very-long-chain monounsaturated fatty acid components of membrane lipids. *Plant Physiol.* 161, 81–96. doi: 10.1104/pp.112.202325

Tanhuanpää, P. K., Vilkki, J. P., and Vilkki, H. J. (1996). Mapping of a QTL for oleic acid concentration in spring turnip rape (*Brassica rapa* ssp. *oleifera). Theor. Appl. Genet.* 92, 952–956. doi: 10.1007/BF00224034

Tasseva, G., De Virville, J. D., Cantrel, C., Moreau, F., and Zachowski, A. (2004). Changes in the endoplasmic reticulum lipid properties in response to low temperature in *Brassica napus*. *Plant Physiol. Biochem.* 42, 811–822. doi: 10.1016/j.plaphy.2004.10.001

Teh, L. S. (2015). *Genetic Variation and Inheritance of Phytosterol and Oil Content in Winter Oilseed Rape (Brassica napus L.)*. Available online at http://hdl.handle.net/11858/00-1735-0000-0022-5D9B-E

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680. doi: 10.1093/nar/22.22.4673

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Tresch, S., Heilmann, M., Christiansen, N., Looser, R., and Grossmann, K. (2012). Inhibition of saturated very-long-chain fatty acid biosynthesis by mefluidide and perfluidone, selective inhibitors of 3-ketoacyl-CoA synthases. *Phytochemistry* 76, 162–171. doi: 10.1016/j.phytochem.2011.12.023

Velasco, L., Fernandezmartinez, J. M., and Ade, H. (1997). Induced variability for C18 unsaturated fatty acids in Ethiopian mustard. *Can. J. Plant Sci.* 77, 91–95. doi: 10.4141/P96-025

Wang, N., Wang, Y., Tian, F., King, G. J., Zhang, C., Long, Y., et al. (2008). A functional genomics resource for *Brassica napus*: development of an EMS mutagenized population and discovery of *FAE1* point mutations by TILLING. *New Phytol.* 180, 751–765. doi: 10.1111/j.1469-8137.2008.02619.x

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wang, X., Long, Y., Yin, Y., Zhang, C., Gan, L., Liu, L., et al. (2015). New insights into the genetic networks affecting seed fatty acid concentrations in *Brassica napus*. *BMC Plant Biol.* 15:91. doi: 10.1186/s12870-015-0475-8

Wei, L., Jian, H., Lu, K., Filardo, F., Yin, N., Liu, L., et al. (2015). Genome-wide association analysis and differential expression analysis of resistance to *Sclerotinia* stem rot in *Brassica napus*. *Plant Biotechnol. J.* 14, 1368–1380. doi: 10.1111/pbi.12501

Wen, J., Xu, J., Long, Y., Xu, H., Wu, J., Meng, J., et al. (2015). Mapping QTLs controlling beneficial fatty acids based on the embryo and maternal plant genomes in *Brassica napus* L. *J. Am. Oil Chem. Soc.* 92, 541–552. doi: 10.1007/s11746-015-2618-3

Wu, G., Wu, Y., Xiao, L., Li, X., and Lu, C. (2008). Zero erucic acid trait of rapeseed (*Brassica napus* L.) results from a deletion of four base pairs in the fatty acid elongase 1 gene. *Theor. Appl. Genet.* 116, 491–499. doi: 10.1007/s00122-007-0685-z

Xiao, G., Zhang, H., Peng, Q., and Guan, C. (2008). Screening and analysis of multiple copy of oleate desaturase gene (fad2) in *Brassica napus*. *Acta Agronomica Sinica* 34, *1563–1568*. doi: 10.3321/j.issn:0496-3490.2008.09.011

Xu, J., Long, Y., Wu, J., Xu, H., Zhao, Z., Wen, J., et al. (2015). QTL identification on two genetic systems for rapeseed glucosinolate and erucic acid contents over two seasons. *Euphytica* 205, 1–11. doi: 10.1007/s10681-015-1379-2

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Yang, Q., Fan, C., Guo, Z., Qin, J., Wu, J., Li, Q., et al. (2012). Identification of *FAD2* and *FAD3* genes in *Brassica napus* genome and development of allele-specific markers for high oleic and low linolenic acid contents. *Theor. Appl. Genet.* 125, 715–729. doi: 10.1007/s00122-012-1863-1

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhang, Y., Liu, P., Zhang, X., Zheng, Q., Chen, M., Ge, F., et al. (2018). Multi-Locus genome-wide association study reveals the genetic architecture of stalk lodging resistance-related traits in maize. *Front. Plant Sci.* 9:611. doi: 10.3389/fpls.2018.00611

Zhang, Z., Xiao, G., Tan, T., and Guan, C. (2009). Advances on high oleic acid oilseed rape breeding and prospects. *Crops* 24, 1–6. doi: 10.16035/j.issn.1001-7238.2009.05.05

Zhao, J. (2002). *QTLs for Oil Content and Their Relationships to Other Agronomic Traits in an European × Chinese Oilseed Rape Population*. Germany: Diss. Grorg-Agust University of Goettingen.

Zhao, J., Becker, H. C., Zhang, D., Zhang, Y., and Ecke, W. (2005). Oil content in a European × Chinese rapeseed population: QTL with additive and epistatic effects and their genotype-environment interactions. *Crop Sci.* 45, 51–59. doi: 10.2135/cropsci2005.0051

Zhao, J., Dimov, Z., Becker, H. C., Ecke, W., and Möllers, C. (2008). Mapping QTL controlling fatty acid composition in a doubled haploid rapeseed population segregating for oil content. *Mol. Breed.* 21, 115–125. doi: 10.1007/s11032-007-9113-y

Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., et al. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* 2:467. doi: 10.1038/ncomms1467

Zhou, Y., Xu, D., Jia, L., Huang, X., Ma, G., Wang, S., et al. (2017). Genome-wide identification and structural analysis of bZIP transcription factor genes in *Brassica napus*. *Genes* 8:e288. doi: 10.3390/genes8100288

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Genome-Wide Association Studies for Pasmo Resistance in Flax (*Linum usitatissimum L.*)

Liqiang He [1,2], Jin Xiao [2], Khalid Y. Rashid [3], Zhen Yao [3], Pingchuan Li [3], Gaofeng Jia [4], Xiue Wang [2], Sylvie Cloutier [1] and Frank M. You [1,2]**

[1] Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, ON, Canada, [2] Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Agriculture, Nanjing Agricultural University/JCIC-MCP, Nanjing, China, [3] Morden Research and Development Centre, Agriculture and Agri-Food Canada, Morden, MB, Canada, [4] Crop Development Centre, University of Saskatchewan, Saskatoon, SK, Canada

OPEN ACCESS

Edited by:
Yuan-Ming Zhang,
Huazhong Agricultural University,
China

Reviewed by:
Hailong Ning,
Northeast Agricultural University,
China
Dan Zhang,
Henan Agricultural University, China

***Correspondence:**
Sylvie Cloutier
sylvie.j.cloutier@canada.ca
Frank M. You
frank.you@canada.ca

Specialty section:
This article was submitted to
Plant Breeding,
a section of the journal
Frontiers in Plant Science

Received: *07 November 2018*
Accepted: *20 December 2018*
Published: *14 January 2019*

Citation:
He L, Xiao J, Rashid KY, Yao Z, Li P, Jia G, Wang X, Cloutier S and You FM (2019) Genome-Wide Association Studies for Pasmo Resistance in Flax (Linum usitatissimum L.). Front. Plant Sci. 9:1982. doi: 10.3389/fpls.2018.01982

Pasmo is one of the most widespread diseases threatening flax production. To identify genetic regions associated with pasmo resistance (PR), a genome-wide association study was performed on 370 accessions from the flax core collection. Evaluation of pasmo severity was performed in the field from 2012 to 2016 in Morden, MB, Canada. Genotyping-by-sequencing has identified 258,873 single nucleotide polymorphisms (SNPs) distributed on all 15 flax chromosomes. Marker-trait associations were identified using ten different statistical models. A total of 692 unique quantitative trait nucleotides (QTNs) associated with 500 putative quantitative trait loci (QTL) were detected from six phenotypic PR datasets (five individual years and average across years). Different QTNs were identified with various statistical models and from individual PR datasets, indicative of the complementation between analytical methods and/or genotype × environment interactions of the QTL effects. The single-locus models tended to identify large-effect QTNs while the multi-loci models were able to detect QTNs with smaller effects. Among the putative QTL, 67 had large effects (3–23%), were stable across all datasets and explained 32–64% of the total variation for PR in the various datasets. Forty-five of these QTL spanned 85 resistance gene analogs including a large toll interleukin receptor, nucleotide-binding site, leucine-rich repeat (TNL) type gene cluster on chromosome 8. The number of QTL with positive-effect or favorite alleles (NPQTL) in accessions was significantly correlated with PR ($R^2 = 0.55$), suggesting that these QTL effects are mainly additive. NPQTL was also significantly associated with morphotype ($R^2 = 0.52$) and major QTL with positive effect alleles were present in the fiber type accessions. The 67 large effect QTL are suited for marker-assisted selection and the 500 QTL for effective genomic prediction in PR molecular breeding.

Keywords: pasmo resistance, quantitative trait loci (QTL), quantitative trait nucleotides (QTNs), fiber, linseed, core collection, flax, *Linum usitatissimum*

INTRODUCTION

Flax (*Linum usitatissimum* L.) is an important economic crop for both linseed and stem fiber. As of 2011, flax was the third largest textile fiber crop and the fifth largest oil crop in the world, with Canada being the world's largest exporter of flax seeds (You et al., 2017). Pasmo, caused by *Septoria linicola* (Speg.) Garassini, is one of the most widespread diseases threatening flax production. Infected plants show brown circular lesions on leaves and brown to black banding patterns alternating with green healthy tissue on stems. Pasmo infects flax plants from seedling to maturity, but it is most acute during ripening under high humidity and high temperature conditions. During flowering, yield losses in susceptible varieties can reach up to 75% despite fungicide application (Hall et al., 2016). Pasmo also negatively affects seed and fiber quality. Despite the slow improvements made in pasmo resistance (PR) through breeding, developing resistant varieties remains the most efficient and environmentally friendly approach to prevent yield losses caused by the disease.

Conventional breeding approaches have been widely used to incorporate genetic variations to improve agronomic traits and introduce durable resistance to diseases in flax (Soto-Cerda et al., 2014b). The availability of the latest molecular tools allows the rapid identification of genes of interest and the selection of individuals carrying favorable genes, and may well-serve to improve breeding efficiency. Development of molecular markers associated with host resistance to pathogens is paramount to marker-assisted selection (MAS), enhancing the power of selection in plant breeding by combining the advantages of high precision and reduced cost (Kumar et al., 2011). MAS for disease resistance is routinely applied for a number of plant-pathogen systems to select resistant genotypes (Miedaner and Korzun, 2012). To date, no genetic study on flax PR has been reported despite the identification of more than one million single nucleotide polymorphisms (SNPs) from a flax core collection (You et al., unpublished data) that constitute a suitable genotypic dataset to detect marker-trait associations (MTAs) through genome-wide association studies (GWAS).

GWAS commonly estimate the statistical significance of MTAs in a diverse genetic panel that can lead to the identification of causal genes underlying phenotypes. GWAS with high-throughput genotyping are advantageous over traditional biparental population analyses, such as rapid processing of large mapping populations, high abundance of molecular markers, and identification of causal loci at a higher resolution (Goutam et al., 2015; Ogura and Busch, 2015). GWAS have been successfully applied to the identification of MTAs for many important flax agronomic traits (Soto-Cerda et al., 2014a,b; Xie et al., 2017; You et al., 2018b). The effectiveness of GWAS in identifying MTAs for disease resistance traits is exemplified in wheat for fungal diseases, such as *Fusarium* head blight (FHB) (Buerstmayr et al., 2009), leaf and stem rusts (Liu et al., 2017).

In general, population structure can be represented by proportions of individuals from subpopulations, regularly called the Q matrix (Larsson et al., 2013), or alternatively principal components (PCs) (Reich et al., 2008; Stich et al., 2008; Zhang et al., 2009) derived from genome-wide molecular markers. The relationships among individuals of a population are represented by a kinship matrix (K). False positive MTAs generally result from two indirect factors: population structure and kinship among individuals (Price et al., 2006; Liu et al., 2016). Two statistical models have been widely used to reduce false positives. The first is the General Linear Model (GLM) or Q model (Price et al., 2006) in which the population structure is fitted as fixed effect. The second is the Mixed Linear Model (MLM) (Yu et al., 2006) that additionally fits kinship as random effect, hence its alternative name, the Q + K model. Theoretically, MLM methods correct the inflation from small polygenic effects, effectively controlling the population stratification bias (Wen et al., 2017); thus, some reports show that the Q + K model outperforms the independent Q and K only models (Liu et al., 2016). The computational burden of MLMs remains a major issue. Some methods have been proposed to improve computational efficiency including Efficient Mixed-Model Association (EMMA) (Kang et al., 2008) and Genome-Wide Efficient Mixed-Model Association (GEMMA) (Zhou and Stephens, 2012).

GLM and MLMs are single-locus methods that perform one-dimensional genome scans by testing one marker at a time using stringent multiple test corrections (such as Bonferroni) as significance threshold. As such, these methods have relatively low power to detect the polygenes with small effects that underlay most quantitative traits. Thus, Multi-Locus Mixed-Model (MLMM) (Segura et al., 2012) was proposed to simultaneously test multiple markers. Alternative and powerful multi-locus methods have been proposed to identify quantitative trait nucleotides (QTNs) with small effects, such as the multi-locus random-SNP-effect Mixed Linear Model (mrMLM) (Wang et al., 2016; Li et al., 2017), the FAST multi-locus random-SNP-effect EMMA (FASTmrEMMA) (Wen et al., 2017), the polygene-background-control-based Least Angle Regression plus Empirical Bayes (pLARmEB) (Zhang et al., 2017), the Iterative modified-Sure Independence Screening EM-Bayesian LASSO (ISIS EM-BLASSO) (Tamba et al., 2017), and the integration of Kruskal–Wallis test with Empirical Bayes under polygenic background control (pKWmEB). These multi-locus methods do not rely on stringent Bonferroni correction (Ren et al., 2017); the algorithms underlying these statistical models substantially increase the statistical power and reduce Type 1 error and running time (Wang et al., 2016; Li et al., 2017; Ren et al., 2017; Tamba et al., 2017; Wen et al., 2017; Zhang et al., 2017). An additional multi-locus model, called Fixed and random model Circulating Probability Unification (FarmCPU) (Liu et al., 2016) divides the MLMM into a fixed effect model (FEM) and a random effects model (REM) and uses them iteratively. Its advantages are improved statistical power and reduction of the confounding between population structure, kinship, and QTN (Liu et al., 2016).

To find QTL associated with field PR, we performed GWAS using a diverse genetic panel of 370 accessions of the flax core collection (Diederichsen et al., 2012; Soto-Cerda et al., 2013) and 258,873 SNPs identified from this population (You et al., unpublished data). Seven multi-locus and three single-locus statistical methods were evaluated with the PR datasets from 5 consecutive years to determine the suitable statistical methods for detecting putative QTL with large or small effects and environmental stability.

MATERIALS AND METHODS

Genetic Panel for GWAS

A diverse genetic panel of 370 cultivated flax accessions from the core collection (Diederichsen et al., 2012; Soto-Cerda et al., 2013) was used. The core collection was assembled from the world collection of 3,378 flax accessions, collected from 39 countries and corresponding to 11 geographical origins defined as North America, South America, Eastern Asia, Western Asia, Southern Asia, Central, and Eastern Europe, Western Europe, Southern Europe, Northern Europe, Oceania, and Africa. This panel contained 17 landraces, 85 breeding lines, 232 cultivars, and 36 accessions of unknown improvement status that were grouped into two morphotypes: 80 fiber and 290 linseed (You et al., 2017).

Phenotyping of Pasmo Resistance and Statistical Analysis

The 391 accessions were evaluated for field PR in the same pasmo nursery from 2012 to 2016 at Agriculture and Agri-Food Canada, Morden Research and Development Center's farm, Morden, Manitoba, Canada. A type-2 modified augmented design (MAD2) (Lin and Poushinsky, 1985) was used for the field trials (You et al., 2017). Each accession was seeded during the second or third week of May every year. Approximately 200 g of pasmo-infested chopped straw from the previous growing season was spread between rows as inoculum when plants were ∼30-cm tall. A misting system was operated for 5 min every half hour for 4 weeks, except on rainy days, to ensure conidia dispersal and disease infection and development. PR was assessed on leaves and stems of all plants in a single row plot using a pasmo severity (PS) scale of 0–9 (**Table 1**). Field assessments were conducted at the early (P1) and late flowering stages (P2, 7–10 days after P1), the green boll stage (P3, 7–10 days after P2), and the early brown boll stage (P4, 7–10 days after P3). In 2014 and 2015, only the first three field assessments were conducted because early maturity of the plants did not allow for a fourth rating. A rating of 0–2 is considered resistant (R), 3–4 moderately resistant (MR), 5–6 moderately susceptible (MS), and 7–9 susceptible (S). The statistical analysis for the phenotypic data was performed as described in You et al. (2013). A total of 370 accessions that had complete phenotypic data and sequence data were used for GWAS (**Table S1**).

The variance components for pasmo severity were estimated using the linear mixed effects "lmer" model in R package "lme4." All effects of variance components were treated as random and the following linear model was used:

$$\begin{aligned} X_{ij} = \mu + G_i + Y_j + (GY)_{ij} + \varepsilon_{ij},\ i &= 1, 2, \ldots, n \ and \\ j &= 1, 2, \ldots, m, \end{aligned}$$

where n and m are the number of genotypes and years, respectively, X_{ij} is the observed pasmo severity, μ is the overall mean, G_i is the effect resulting from the ith genotype, Y_j is the effect resulting from the jth year, $(GY)_{ij}$ is the effect resulting from genotype × year (environment) interaction, and ε_{ij} is the residual error (effect resulting from the experimental error).

TABLE 1 | Criteria for field assessment of pasmo severity on a scale of 0–9.

Severity score	Criteria
0	No sign of pasmo, the most vigorous
1	<10 leaf area or/and stem area affected by pasmo
2	10–20% leaf area or/and stem area affected by pasmo
3	21–30% leaf area or/and stem area affected by pasmo
4	31–40% leaf area or/and stem area affected by pasmo
5	41–50% leaf area or/and stem area affected by pasmo
6	51–60% leaf area or/and stem area affected by pasmo
7	61–70% leaf area or/and stem area affected by pasmo
8	71–80% leaf area or/and stem area affected by pasmo
9	>80% leaf area or/and stem area affected by pasmo

Assessment of pasmo severity is based on all plants in a single row plot.

Resequencing and SNP Discovery of the Core Collection

Genotyping by sequencing (GBS) methodology was employed to genotype all individuals of the core collection. The Illumina HiSeq 2000 platform (Illumina Inc., San Diego, USA) was used to generate 100-bp paired-end reads with ∼15.5 × genome equivalents of the reference genome. All reads from each individual of the population were aligned to the scaffold sequences of the flax reference genome (Wang et al., 2012) using BWA v0.6.1(Jo and Koh, 2015) with base-quality Q score in Phred scale >20 and other default parameters. The alignment file for each individual was used as input for SNP discovery using the software package SAMtools (Li et al., 2009). All variants were further filtered to get a set of high-quality SNPs as previously described (Kumar et al., 2012). The coordinates of SNPs were then converted to the chromosome scale of the flax pseudomolecules v2.0 upon its release (You et al., 2018a). All procedures were implemented in the AGSNP pipeline (You et al., 2011, 2012) and its updated GBS version (Kumar et al., 2012). The detected SNPs were further filtered with minor allele frequency (MAF) > 0.05 and SNP genotyping rate ≥ 60%. To minimize contribution from regions of extensive strong linkage disequilibrium (LD), a single SNP was retained per 200-kb window when pairwise correlation coefficients (r^2) among neighboring SNPs were >0.8 (International HapMap, 2005; Huang et al., 2010), resulting in a total of 258,873 SNPs. Missing SNPs (on average 14.13% of a missing data rate) were imputed using Beagle v.4.2 with default parameters (Browning and Browning, 2007).

Genome-Wide Association Study and Validation

GWAS analyses were conducted separately for combinations of the 5 individual years and the 5-years average datasets with 10 single- and multi-locus methods (**Table 2**). Kinship

TABLE 2 | Statistical methods used for GWAS.

Statistical model	Q matrix or PCs	Threshold for QTNs	GWAS software	References
GLM	First six PCs	Bonferroni correction	MVP v1.0.1	Price et al., 2006
MLM	First six PCs	Bonferroni correction	MVP v1.0.1	Yu et al., 2006
FarmCPU	First six PCs	Bonferroni correction	MVP v1.0.1	Liu et al., 2016
GEMMA	None needed	Bonferroni correction	GEMMA v0.96	Zhou and Stephens, 2012
mrMLM	From Frappe	LOD $\geq$ 3	mrMLM v3.0	Wang et al., 2016
FASTmrEMMA	From Frappe	LOD $\geq$ 3	mrMLM v3.0	Wen et al., 2017
ISIS EM-BLASSO	From Frappe	LOD $\geq$ 3	mrMLM v3.0	Tamba et al., 2017
pLARmEB	From Frappe	LOD $\geq$ 3	mrMLM v3.0	Zhang et al., 2017
pKWmEB	From Frappe	LOD $\geq$ 3	mrMLM v3.0	Ren et al., 2017
FASTmrMLM	From Frappe	LOD $\geq$ 3	mrMLM v3.0	https://cran.r-project.org/web/packages/mrMLM/index.html

The kinship matrix used for each method was calculated using the module implemented in the corresponding software. PC, principal component; QTN, quantitative trait nucleotide.

genetic relationship matrices were estimated using the protocol suggested by each GWAS software package. The population structure of the 370 accessions was estimated using Frappe (http://med.stanford.edu/tanglab/software/frappe.html) or PCs as determined by principal component analysis (PCA) using MVP in the R package (https://github.com/XiaoleiLiuBio/MVP). Using Frappe, the 370 accessions of the flax core collection were grouped into five sub-populations that corresponded to two major morphotypes (fiber and oil) and different geographical regions (**Table S1**).

For GLM, MLM and FarmCPU, the first six PCs, accounting for 33.04% of the total variation, were chosen as covariates to measure population structure (**Figure S1**). GEMMA was also compared with the regular MLM methods because it does not require a Q matrix. The threshold of significant associations for all four of these methods was determined by a critical *P*-value ($\alpha = 0.05$) subjected to Bonferroni correction, i.e., the corrected *P*-value $= 1.93 \times 10^{-7}$ (0.05/258,873 SNPs). GWAS analyses for the GLM, MLM, and FarmCPU were performed using the R package MVP (https://github.com/XiaoleiLiuBio/MVP) and for GEMMA using the GEMMA software (https://github.com/genetics-statistics/GEMMA). The additional six multi-locus methods were conducted with default parameters using the R package mrMLM (https://cran.r-project.org/web/packages/mrMLM/index.html) (**Table 2**). Because these six methods are implemented in the same mrMLM R package and developed by the same research team, we refer to them as "mrMLM methods." A log of odds (LOD) score of three was used to detect robust association signals for these six methods.

After putative QTNs were identified, we performed QTN analysis to obtain sets of QTNs/QTL. The procedure is summarized in **Figure 1**. First, we tested the significance of the difference in PS between two alleles of a QTN (henceforth called QTN effect) in all accessions. Statistically significant differences served to validate the QTNs. Wilcox non-parametric tests were performed using the R function *wilcox.test* to remove the non-significant QTNs at a 5% probability level. The direction (positive or negative) of QTN effects were subsequently assessed. Only QTNs with consistent effect directions in all PS datasets were considered valid and were retained. Such QTNs were grouped into QTL by calculating linkage disequilibrium (D′) between pairs of QTNs on the same chromosomes using plink v1.9 (https://www.cog-genomics.org/plink2). Neighboring QTNs with $D' > 0.8$ were grouped into the same QTL (Twells et al., 2003; Grassmann et al., 2017). For each such defined QTL, the QTN of the largest average R^2 over all datasets was chosen as a representative or tag for the QTL. R^2 were calculated based on simple regressions of QTNs on PS because they represent the proportion of the total variation of PS explained by the QTNs/QTL.

Statistically stable QTL were those significant across all six PS datasets. Multiple regressions of all stable QTL were fitted to each of the six PS datasets using a forward stepwise regression to select QTL with significant contributions to PS. Six regression models were obtained for the six PS datasets. Only QTL existing in at least three regression models were considered to be statistically stable with relatively large effects.

To test QTL effect additivity, the number of QTL with positive-effect or favorite alleles (NPQTL) in all accessions was tallied. A QTL with positive-effect or favorite allele (PQTL) in a given accession was called if this accession possessed a positive effect allele for that QTL. In the case of the PS trait (PS rating is opposite to resistance), alleles with positive signs are associated with lower PR. A simple regression of NPQTL on PS in the population was calculated. Correlations of NPQTL with PS in the six PS datasets were calculated using the R function "*cor*."

Resistance Gene Analogs (RGAs) Co-localized With QTL

A total of 1,327 RGAs have been identified in the flax pseudomolecule (You et al., 2018a). To predict candidate resistance genes co-localized with QTL, the RGAs within 200 kb of a QTL's flanking region were considered.

Evaluation of the Flax Core Collection

The extreme pasmo resistant and susceptible accession subsets and all 370 accessions were evaluated based on the identified

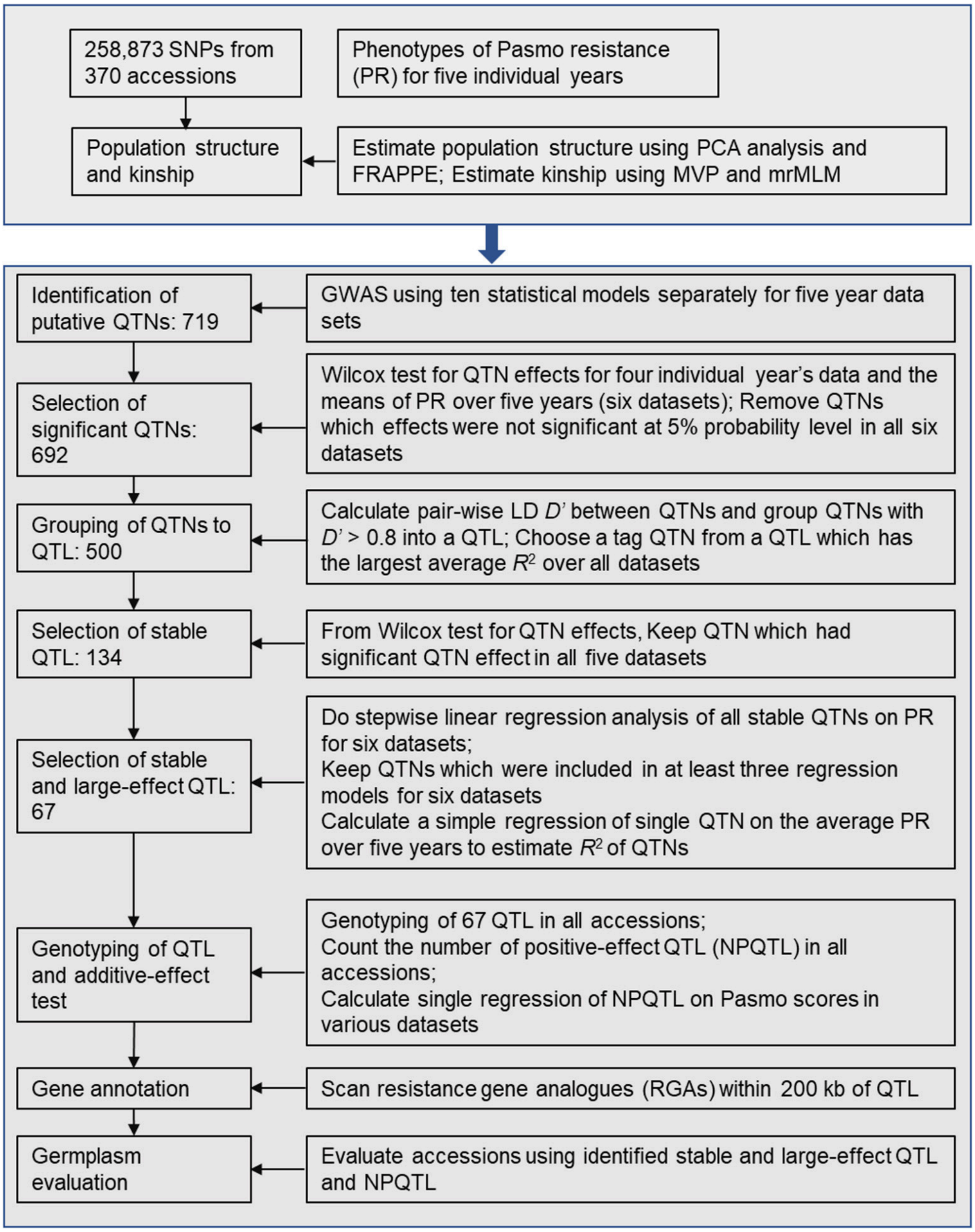

258,873 SNPs from 370 accessions
Phenotypes of Pasmo resistance (PR) for five individual years
Population structure and kinship
Estimate population structure using PCA analysis and FRAPPE; Estimate kinship using MVP and mrMLM
Identification of putative QTNs: 719
GWAS using ten statistical models separately for five year data sets
Selection of significant QTNs: 692
Wilcox test for QTN effects for four individual year's data and the means of PR over five years (six datasets); Remove QTNs which effects were not significant at 5% probability level in all six datasets
Grouping of QTNs to QTL: 500
Calculate pair-wise LD D' between QTNs and group QTNs with $D' > 0.8$ into a QTL; Choose a tag QTN from a QTL which has the largest average R^2 over all datasets
Selection of stable QTL: 134
From Wilcox test for QTN effects, Keep QTN which had significant QTN effect in all five datasets
Selection of stable and large-effect QTL: 67
Do stepwise linear regression analysis of all stable QTNs on PR for six datasets;
Keep QTNs which were included in at least three regression models for six datasets
Calculate a simple regression of single QTN on the average PR over five years to estimate R^2 of QTNs
Genotyping of QTL and additive-effect test
Genotyping of 67 QTL in all accessions;
Count the number of positive-effect QTL (NPQTL) in all accessions;
Calculate single regression of NPQTL on Pasmo scores in various datasets
Gene annotation
Scan resistance gene analogues (RGAs) within 200 kb of QTL
Germplasm evaluation
Evaluate accessions using identified stable and large-effect QTL and NPQTL

FIGURE 1 | Diagram of stable and large-effect quantitative trait loci (QTL) associated with flax pasmo resistance. NPQTL: the number of QTL with positive-effect alleles.

stable and large-effect QTL. Two extreme subsets of 23 resistant (R) and 23 susceptible (S) were selected based on PS ratings. Two-dimensional cluster analysis for accessions and the QTL were performed. The Euclidean distances between accessions or between QTL were calculated based on QTL genotypes (positive alleles as 1 and alternate alleles as 2) using the "*dist*" function with the "*euclidean*" method in R. The Ward algorithm in the function "*hclust*" of the R package stats was employed for hierarchical cluster analysis. Dendrograms and heat maps were created with the R package Complexheatmap.

RESULTS

Pasmo Resistance

PS ratings increased with growth stage and peaked at the final evaluation stage every year (**Figure 2**; **Table S2**), supporting the adoption of the final stage data for analysis. Significant correlations of PS ratings across years were observed (**Table 3**) but were largely affected by year and genotype × year interaction (**Table S3**). The variance of genotype × year interaction accounted for 24.23% of the total variation of PS (**Table S4**). Thus, datasets from all individual years and the 5-years average were used for GWAS analyses.

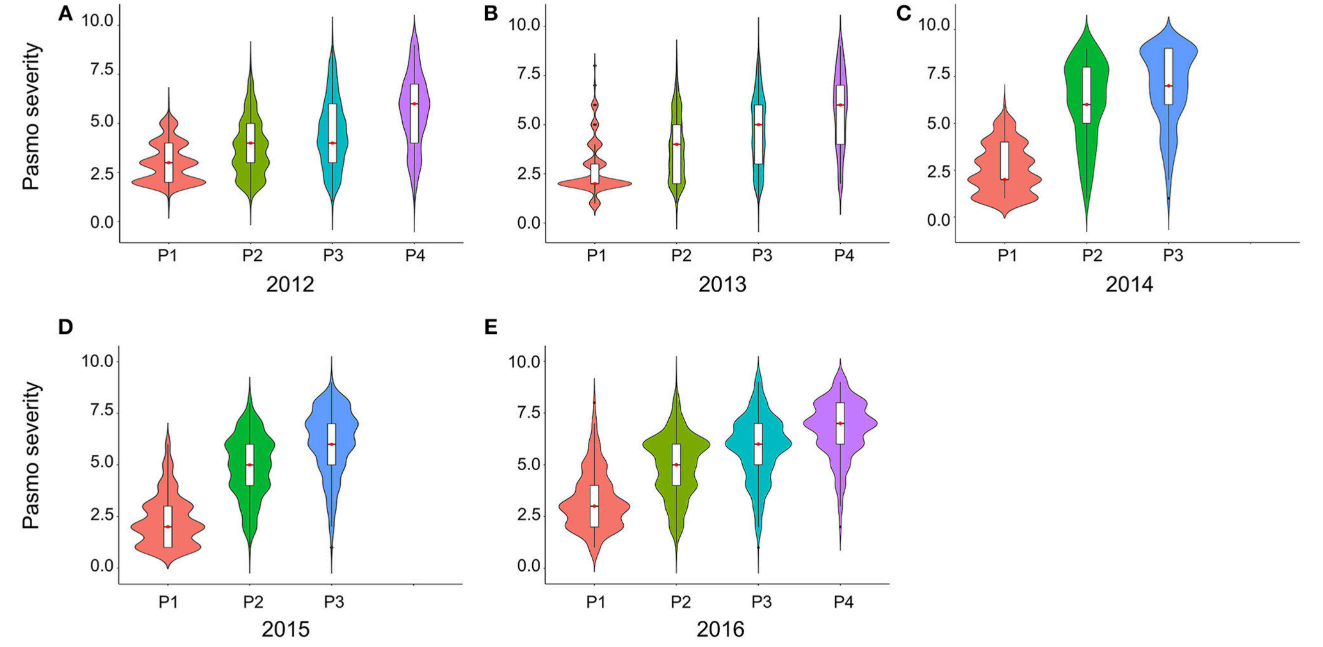


FIGURE 2 | Violin plots of pasmo severity observed in various growth stages in 2012 **(A)**, 2013 **(B)**, 2014 **(C)**, 2015 **(D)**, and 2016 **(E)**.

TABLE 3 | Spearman correlation coefficients of pasmo severity between years (2012–2016).

	2012	2013	2014	2015	2016
2012	1	0.56**	0.54**	0.44**	0.62**
2013		1	0.51**	0.49**	0.60**
2014			1	0.76**	0.70**
2015				1	0.70**
2016					1

**Significant at 1% probability level.*

TABLE 4 | Comparison of QTN/QTL identification for different statistical models.

Statistical model	No. of QTL identified	No. of QTNs identified	No. of non-significant QTNs	Average R^2 (%)	R^2 range (%)
GLM	209	346	2	5.57	0.48–15.02
MLM	1	1	0	15.02	N/A
GEMMA	6	6	0	11.13	3.59–15.02
mrMLM	97	99	7	2.75	0.36–15.02
FASTmrEMMA	60	62	2	2.82	0.25–6.92
ISIS EM-BLASSO	97	98	8	2.91	0.29–12.68
pLARmEB	118	120	8	2.69	0.22–12.68
pKWmEB	95	95	5	2.93	0.25–12.68
FASTmrMLM	125	125	3	2.69	0.25–15.02
FarmCPU	22	22	3	5.09	0.42–15.02

Identification of PR QTL

A total of 719 putative QTNs were identified using the three single-locus and seven multi-locus methods for the six PS datasets (**Table S4**). To further statistically check the significance of QTNs, a Wilcox non-parametric test was conducted for the six datasets separately. A total of 27 QTNs were removed because they were not significant in all six datasets. The remaining 692 QTNs were merged into a total of 500 QTL based on the linkage disequilibrium D' criteria between contiguous QTNs. GLM detected multiple significant QTNs in the same QTL region while in most cases a QTN corresponded to a QTL for other single-locus and multi-locus methods (**Table 4**). Tag QTNs were selected to represent each QTL for downstream analyses.

Of the three single-locus methods, MLM identified only one QTL ($R^2 = 15.02\%$) while GEMMA detected six with an average R^2 of 11.13%. GLM identified the largest number of QTL (209) or QTNs (346) of all methods and these had relatively large effects with an average R^2 of 5.57%, ranging from 0.48 to 15.02%.

QTL differed depending on the statistical methods. QTL detected by at least two methods accounted for a small proportion of overall QTL (**Tables S4, S5**). The mrMLM methods detected more common QTL than the other methods, e.g., 45 QTL in common with pLARmEB and FASTmrMLM and 32 with ISIS EM-BLASSO and pKWmEB (**Table 5**). Multi-locus methods detected more small-effect QTL than the single-locus methods (**Table 4**). Six mrMLM methods could identify more QTL with smaller effects (average R^2 of 2.80%) than FarmCPU (average R^2 of 5.09%) owing to the high stringency of the Bonferroni correction used in FarmCPU. Generally, QTL with

TABLE 5 | Number of common QTNs and QTL identified by any two statistical models.

Statistical model	GLM	MLM	GEMMA	FASTmrEMMA	FASTmrMLM	ISIS EM-BLASSO	mrMLM	pKWmEB	pLARmEB
MLM	1 (1)								
GEMMA	6 (6)	1 (1)							
FASTmrEMMA	8 (5)	0 (0)	0 (0)						
FASTmrMLM	17 (12)	1 (1)	2 (2)	26 (18)					
ISIS EM-BLASSO	18 (15)	0 (0)	3 (3)	13 (9)	33 (23)				
mrMLM	18 (11)	1 (1)	5 (5)	17 (14)	33 (27)	18 (13)			
pKWmEB	24 (17)	0 (0)	3 (3)	17 (12)	34 (30)	34 (32)	27 (19)		
pLARmEB	20 (13)	0 (0)	2 (2)	26 (15)	50 (45)	31 (20)	29 (24)	32 (27)	
FarmCPU	13 (13)	1 (1)	2 (2)	1 (1)	3 (3)	5 (4)	7 (6)	5 (4)	5 (4)

Values in parentheses are the number of QTL corresponding to QTNs.

TABLE 6 | QTNs/QTL for pasmo severity identified using phenotypic data from 5 consecutive years and their mean with ten statistical models.

Dataset	No. of QTL identified	No. of QTNs identified	No. of non-significant QTNs	Average R^2 (%)	R^2 range (%)
Mean	240	362	4	5.26	0.28–15.02
2012	82	98	12	3.68	0.22–15.02
2013	92	100	8	3.26	0.32–15.02
2014	114	138	7	4.7	0.42–15.02
2015	65	72	4	2.86	0.27–15.02
2016	85	100	3	4.46	0.39–15.02

TABLE 7 | Number of common QTNs/QTL identified from any two datasets.

Dataset	mean	2012	2013	2014	2015
2012	33 (30)				
2013	28 (25)	6 (4)			
2014	62 (52)	16 (11)	11 (10)		
2015	18 (14)	3 (2)	5 (2)	6 (3)	
2016	45 (36)	10 (6)	7 (6)	14 (11)	7 (7)

Values in parentheses are the number of QTL corresponding to QTNs.

large effects were identified by both GLM and mrMLM methods (**Table S4**).

QTL also differed across individual year datasets (**Tables 6**, **7**, **S5**) but most (240) were identified from the mean dataset which comprised two to four times more QTL than the individual year datasets (**Table 6**). This is indicative of strong gene × environment interactions and reinforces the representability of the mean dataset for QTL identification.

Validation of PR QTL

Additional validation was warranted for the identified QTNs/QTL. Of the 500 QTL, 134 were detected in all six PS datasets and explained 27.4–60.9% of the total variation; they are considered stable. By construction of forward stepwise multiple regression models, 67 out of the 134 stable QTL were detected in at least three regression models; they explained 31.5–64.2% of the total variation in individual datasets: this is comparable or slightly greater than that of the 134 QTL set. The 67 QTL subset represents the non-redundant and large effect QTL as each of them could explain 3.3–23.4% of the total variation (**Table 8**). QTL with similar contributions but that were highly correlated and/or that had small effects were excluded.

The tally of the PQTL in the 370 accessions ranged from 3 to 60 per accession (**Table S6**). NPQTL were compared between two extreme subsets of 23 resistant (R) and 23 susceptible (S), respectively. Notably, all accessions of the R group belong to the fiber type and those of the S group were oilseed type. The R group, with an average PS rating of 3.2, contained an average of 42.5 PQTL per accession ranging from 14 to 60; the S group, with an average PS rating of 8.3, averaged only 9.4 PQTL per accession (**Figure 3**). Significant negative correlations between NPQTL and PS were observed in all six datasets ($r = -0.45$ to -0.74) (**Figures 4A–F**), with the highest negative correlation being with the mean PS rating dataset ($r = -0.74$) (**Figure 4F**).

Association of PR and Its QTL With Flax Morphotype

A significant correlation between PS and morphotype ($r = 0.49$, $p < 0.00001$) was observed, showing that fiber accessions were more resistant to pasmo (**Figure 5A**). NPQTL were also significantly correlated with morphotypes ($r = -0.65$, $p < 0.00001$) (**Figure 5B**). Chi-square tests of independence were performed to identify PQTL alleles specifically belonging to a morphotype. For each QTL, the positive-effect allele was assigned a value of 0 and the alternate allele, a value of 1. Similarly, fiber type accessions were assigned 0 and linseed accessions 1. The chi-square test results indicated that most PQTL alleles were significantly associated with fiber type accessions (**Table S7**). For eight (8, 13, 14, 17, 21, 54, 55, and 63) of the 67 major QTL, between 80 and 100% of the PQTL were present in the fiber accessions; this was particularly acute for QTL 43 and 44 that were almost exclusive to the fiber germplasm. For the remaining 57 QTL, the PQTL were detected in fiber accessions (11–63 out of 80 fiber accessions) but were also found in many linseed accessions.

TABLE 8 | Stable and large-effect QTL associated with pasmo resistance.

QTL	Tag QTN	Chr	Pos	SNP	Effect	R^2	Gene/annotation
1	Lu1-9232234	1	9232234	G/A	−0.91	16.17	
2	Lu1-28707496	1	28707496	G/A	−0.54	5.70	*Lus10006052*/RLK, *Lus10006056*/RLK, *Lus10006057*/RLK, *Lus10006067*/RLK
3	Lu2-3803775	2	3803775	C/T	0.36	3.32	
4	Lu3-19643168	3	19643168	G/A	−1.97	12.82	*Lus10008221*/TNL, *Lus10008222*/TNL, *Lus10008230*/RLP
5	Lu3-20781286	3	20781286	A/C	−1.83	14.63	
6	Lu3-22688547	3	22688547	C/G	−0.89	8.98	*Lus10033291*/RLK
7	Lu4-37769	4	37769	G/A	−1.36	11.23	
8	Lu4-13306407	4	13306407	A/G	0.89	4.58	*Lus10015729*/RLK
9	Lu4-13779313	4	13779313	T/C	−1.73	13.72	
10	Lu4-14576826	4	14576826	A/G	0.42	7.99	*Lus10041509*/RLK, *Lus10041512*/TM-CC
11	Lu4-14615685	4	14615685	A/T	−0.65	10.85	*Lus10041509*/RLK, *Lus10041512*/TM-CC
12	Lu4-14738243	4	14738243	G/T	−0.61	12.64	
13	Lu4-17204590	4	17204590	C/A	0.64	5.17	*Lus10004040*/RLK, *Lus10009107*/TNL, *Lus10009108*/TX, *Lus10009109*/NBS, *Lus10020794*/TM-CC
14	Lu4-17214936	4	17214936	G/T	0.70	5.81	*Lus10004040*/RLK, *Lus10009107*/TNL, *Lus10009108*/TX, *Lus10009109*/NBS, *Lus10020779*/CNL, *Lus10020794*/TM-CC
15	Lu5-1554121	5	1554121	T/A	−0.67	7.75	*Lus10004719*/TNL, *Lus10004726*/CNL, *Lus10004727*/TN
16	Lu5-1650980	5	1650980	C/G	−0.81	6.61	*Lus10004719*/TNL, *Lus10008486*/RLK, *Lus10008491*/RLK
17	Lu5-3575865	5	3575865	C/G	−0.49	9.64	
18	Lu5-4604607	5	4604607	A/G	−0.56	6.58	*Lus10034787*/TM-CC, *Lus10034790*/RLK, *Lus10034795*/RLK
19	Lu5-4858045	5	4858045	C/T	−1.87	12.83	
20	Lu5-13500692	5	13500692	G/A	−1.40	11.9	*Lus10029802*/RLK, *Lus10029810*/TX
21	Lu6-2081466	6	2081466	T/C	0.68	8.30	*Lus10017611*/RLK
22	Lu6-5837358	6	5837358	C/T	−1.18	9.36	
23	Lu6-14738507	6	14738507	C/T	−2.01	13.34	*Lus10014441*/RLP
24	Lu6-15455712	6	15455712	A/G	−1.42	9.63	*Lus10021003*/RLK, *Lus10021022*/RLK
25	Lu6-15506450	6	15506450	A/G	−1.81	12.62	*Lus10021022*/RLK
26	Lu7-2452981	7	2452981	C/T	−0.53	6.30	*Lus10012159*/RLK
27	Lu7-2453965	7	2453965	T/C	−0.56	7.03	*Lus10012159*/RLK
28	Lu7-2491132	7	2491132	G/A	−0.56	8.05	*Lus10012159*/RLK
29	Lu8-14317356	8	14317356	A/T	−0.98	14.32	*Lus10016612*/RLP, *Lus10016620*/RLK
30	Lu8-15830073	8	15830073	C/T	−0.82	8.48	
31	Lu8-15837449	8	15837449	A/T	−1.20	8.24	
32	Lu8-15841885	8	15841885	T/C	−1.15	8.35	
33	Lu8-15963249	8	15963249	A/G	−1.70	14.22	
34	Lu8-16366918	8	16366918	C/T	−1.38	10.90	*Lus10022340*/RLK, *Lus10022345*/RLK, *Lus10022351*/CNL
35	Lu8-17270785	8	17270785	C/G	−1.08	9.59	*Lus10000591*/TM-CC
36	Lu8-17749357	8	17749357	G/A	−1.23	10.16	*Lus10011039*/RLP, *Lus10011064*/RLP
37	Lu8-18251174	8	18251174	G/A	−1.45	10.38	*Lus10007812*/TNL, *Lus10007813*/TNL, *Lus10007814*/TNL, *Lus10007821*/TNL, *Lus10007822*/TNL, *Lus10007823*/OTHER, *Lus10007825*/TNL, *Lus10007826*/TNL, *Lus10007828*/TNL, *Lus10007829*/OTHER, *Lus10007830*/NL, *Lus10007831*/TNL, *Lus10007836*/TNL, *Lus10007852*/TX
38	Lu8-18447612	8	18447612	T/C	−1.41	11.66	*Lus10007790*/TNL, *Lus10007795*/TM-CC, *Lus10007808*/TNL, *Lus10007809*/NL, *Lus10007810*/TNL, *Lus10007811*/TNL, *Lus10007812*/TNL, *Lus10007813*/TNL, *Lus10008540*/RLK
39	Lu8-23104696	8	23104696	C/A	−1.80	16.53	*Lus10018470*/TX
40	Lu8-23142500	8	23142500	T/C	−1.56	13.34	*Lus10018459*/RLK, *Lus10018470*/TX
41	Lu9-1258326	9	1258326	T/A	−1.62	16.01	
42	Lu9-1430465	9	1430465	G/C	−0.69	10.76	*Lus10004333*/RLK
43	Lu9-1896658	9	1896658	G/A	−1.94	17.12	
44	Lu9-4333365	9	4333365	C/A	−2.22	23.39	*Lus10040315*/TM-CC
45	Lu9-6270376	9	6270376	A/G	−0.81	14.34	*Lus10031043*/RLK, *Lus10031058*/TM-CC

(Continued)

TABLE 8 | Continued

QTL	Tag QTN	Chr	Pos	SNP	Effect	R^2	Gene/annotation
46	Lu9-15527375	9	15527375	G/A	−1.07	6.76	
47	Lu9-16348319	9	16348319	C/T	−0.37	4.64	
48	Lu9-19857367	9	19857367	G/A	−1.70	12.67	*Lus10011917*/RLK
49	Lu10-8700793	10	8700793	A/G	−0.53	12.10	*Lus10039958*/RLP
50	Lu11-3330783	11	3330783	A/T	−1.11	7.09	*Lus10042097*/TM-CC
51	Lu12-474480	12	474480	C/T	0.51	8.33	*Lus10020016*/CNL
52	Lu12-1621325	12	1621325	T/A	−1.90	9.41	*Lus10023391*/RLK
53	Lu12-2719326	12	2719326	C/T	−0.62	9.90	*Lus10006971*/TM-CC
54	Lu12-5552631	12	5552631	C/A	0.92	7.10	
55	Lu12-5795458	12	5795458	A/G	0.54	9.67	*Lus10037786*/TM-CC
56	Lu12-5819991	12	5819991	C/G	0.35	6.90	*Lus10037786*/TM-CC
57	Lu12-15686833	12	15686833	A/G	−1.65	13.90	
58	Lu12-16056974	12	16056974	A/C	−1.26	11.26	*Lus10043083*/RLK
59	Lu12-16358216	12	16358216	G/A	0.32	4.25	*Lus10027856*/RLP
60	Lu13-1919638	13	1919638	G/A	−1.55	13.67	*Lus10026845*/TX
61	Lu13-2016767	13	2016767	C/G	−0.38	5.12	*Lus10026845*/TX
62	Lu13-11860250	13	11860250	G/A	−0.96	9.65	
63	Lu13-13051094	13	13051094	G/C	0.68	7.96	
64	Lu13-14299019	13	14299019	A/G	0.39	8.28	*Lus10034637*/RLK, *Lus10034642*/RLK
65	Lu15-976617	15	976617	T/A	−1.65	16.08	*Lus10011216*/TX, *Lus10011223*/RLK, *Lus10011229*/TM-CC
66	Lu15-995626	15	995626	T/A	−0.44	6.27	*Lus10011216*/TX, *Lus10011223*/RLK, *Lus10011229*/TM-CC
67	Lu15-8714776	15	8714776	C/G	−0.98	15.04	

RLK, receptor-like protein kinase; RLP, receptor-like protein; TM-CC, transmembrane coiled-coil protein; NBSl, nucleotide-binding site domain; LRR, leucine-rich repeat; Toll/interleukin-1 receptor-like domain; CNL, CC–NBS–LRR; TNL, TIR-NBS-LRRs; TN, TIR–NBS; TX, TIR–unknown.

Evaluation of the Flax Core Collection With QTL

Based on the 67 core QTL of the flax collection, bi-dimensional cluster analyses were conducted using tag QTNs as representatives of the QTL. The 370 accessions grouped into four clusters (**Figure 6**). Cluster 1 with 269 accessions and Cluster 2 with 35 were mostly susceptible to pasmo (PS ratings of 6.6 ± 1.0 and 6.5 ± 1.1, respectively). Most accessions (243) of Cluster 1 and all accessions of Cluster 2 were linseed type. Cluster 3 comprised 40 moderately susceptible (PS ratings of 5.0 ± 1.1) accessions including 11 of linseed. Cluster 4 contained 26 accessions, of which, 25 were fiber type and only one was a linseed; they were resistant to pasmo (PS ratings of 3.7 ± 1.1). The number of PQTL were 14.2 ± 4.0, 14.2 ± 1.7, 27.7 ± 5.8, and 47.1 ± 6.7 for Clusters 1–4, respectively. The 26 resistant accessions of Cluster 4 represent an important germplasm for PR breeding. This resistant germplasm included 14 of the 23 accessions in **Figure 3** and CN101114, CN101115, CN101119, CN101237, CN101241, CN101296, CN101367, CN101395, CN101396, CN18987, CN98150, and CN98903.

The 67 QTL were clustered into four sub-groups. Group 1 included 13 QTL widely distributed across the germplasm (68.27% of the accessions) but with relatively low QTL effects (average R^2 of 8.31%, ranging from 3.32 to 10.85%) (**Figure 6**; **Table S7**). Groups 2 and 3 contained 7 and 11 QTL, respectively. Present in 31.08% of the accessions, these QTL had an average R^2 of 9.23%, ranging from 4.64 to 16.17%. The 36 QTL of Group 4 had an average R^2 of 11.93%, ranging from 6.61 to 23.39% and contributing to the majority of the PR. These QTL were mostly found in the resistant accessions of Cluster 4 which amounts to a mere 9.70% of the germplasm. CN101367 with 43 QTL and CN19001 with 49 are good examples of resistant germplasm.

Resistance Gene Analogs Co-localized With QTL

Among the 67 stable and large-effect QTL, 45 co-localized with 85 RGAs within the pre-defined 200 Kb QTL flanking window. Four types of RGAs were harbored at these loci: receptor like protein (RLP), receptor like kinase (RLK), nucleotide-binding site (NBS) coding genes (including TNL, TX, CNL, NL, TN, NBS, and others), and transmembrane- coiled coil protein (TM-CC) (Sekhwal et al., 2015). The majority of the RGAs were RLKs with 36.47%, followed by TNLs with 22.35% (**Figure 7**), including a TNL cluster associated with QTL 37 and 38 on chromosome 8 (**Table 8**). Additional TNL-type RGA clusters co-localized with QTL 13 and 14 on chromosome 4 and QTL 15 and 16 on chromosome 5. An RLP gene (*Lus10039958*) located only 56 bp downstream of QTN Lu10-8700793 (QTL 49) exemplifies close linkage between the RGA and the QTL identified in this study. A TNL gene (*Lus10007812*) located 99 Kb downstream of QTN Lu8-18251174 (QTL 37) was the farthest RGA from its QTL.

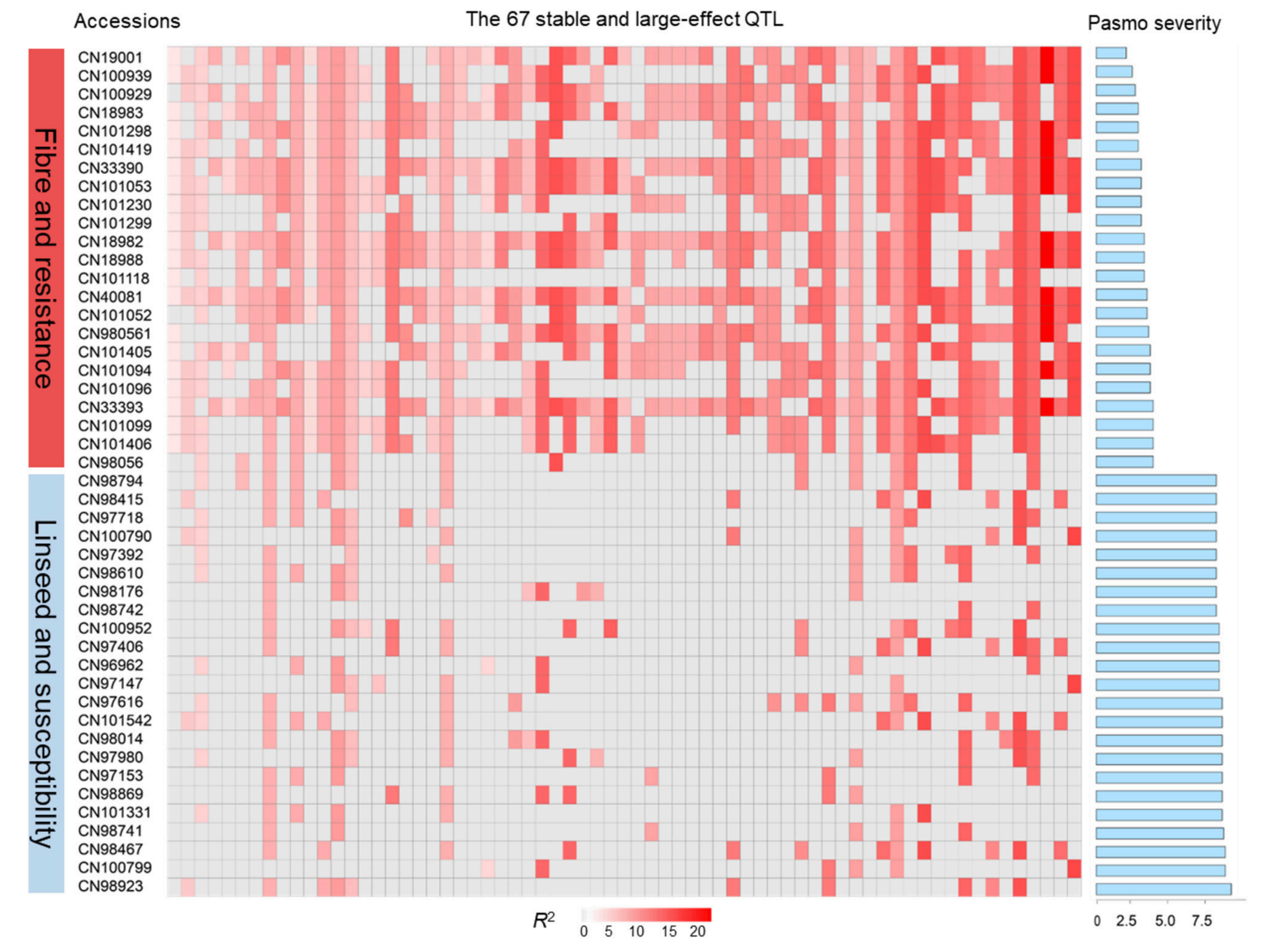


FIGURE 3 | QTL genotypes, pasmo resistance and severity, and morphotypes of two extreme subsets representing 23 resistant and 23 susceptible accessions, respectively.

DISCUSSION

Pasmo Resistance in the Core Collection

Pasmo is widespread throughout all flax growing regions (Halley et al., 2004), but no flax cultivars are truly resistant to pasmo (Diederichsen et al., 2008). Evaluation of pasmo disease response revealed a range of resistance levels in the core collection (You et al., 2017) and, consistent with previous observations, no immune or highly resistant flax varieties were identified. However, some accessions displayed a relatively high level of resistance to pasmo and had high NPQTL (**Figure 3**; **Table S1**). For example, CN19001, a fiber type from the Netherlands, and CN101367, a linseed accession from Georgia, had respective 5-years average PS ratings of 2.0 and 1.8 and possessed 49 and 43 NPQTL, respectively. These accessions of fiber and linseed lineages are good parents for improvement of flax resistance through direct hybridization with elite varieties. Moderately resistant and moderately susceptible lines accounted for 6.49–21.35 and 20.81–42.16% of all accessions in the association panel depending on the years, respectively. Due to the quantitative nature of the disease, this germplasm also holds potential in breeding through the pyramiding of QTL with smaller effects, a strategy that has been successful in improving FHB resistance in wheat (Buerstmayr et al., 2009).

The fiber accessions were generally more resistant to pasmo than the linseed accessions, not surprisingly considering that fiber flax is cultivated for its stem fibers whose quality is greatly affected by the disease. From flowering to maturity, the dark brown to black bands that appear on the stems of infected plants can reduce the quality of the fiber (Colhoun and Muskett, 1943). The relatively higher level of resistance of the fiber type is likely a reflection of artificial selection and possibly independent domestication of the fiber flax lineage (Fu et al., 2012). The transfer of PR from fiber to linseed types can be considered, particularly in schemes where faster recovery of the recurrent linseed types can be achieved by marker-assisted backcrossing for example.

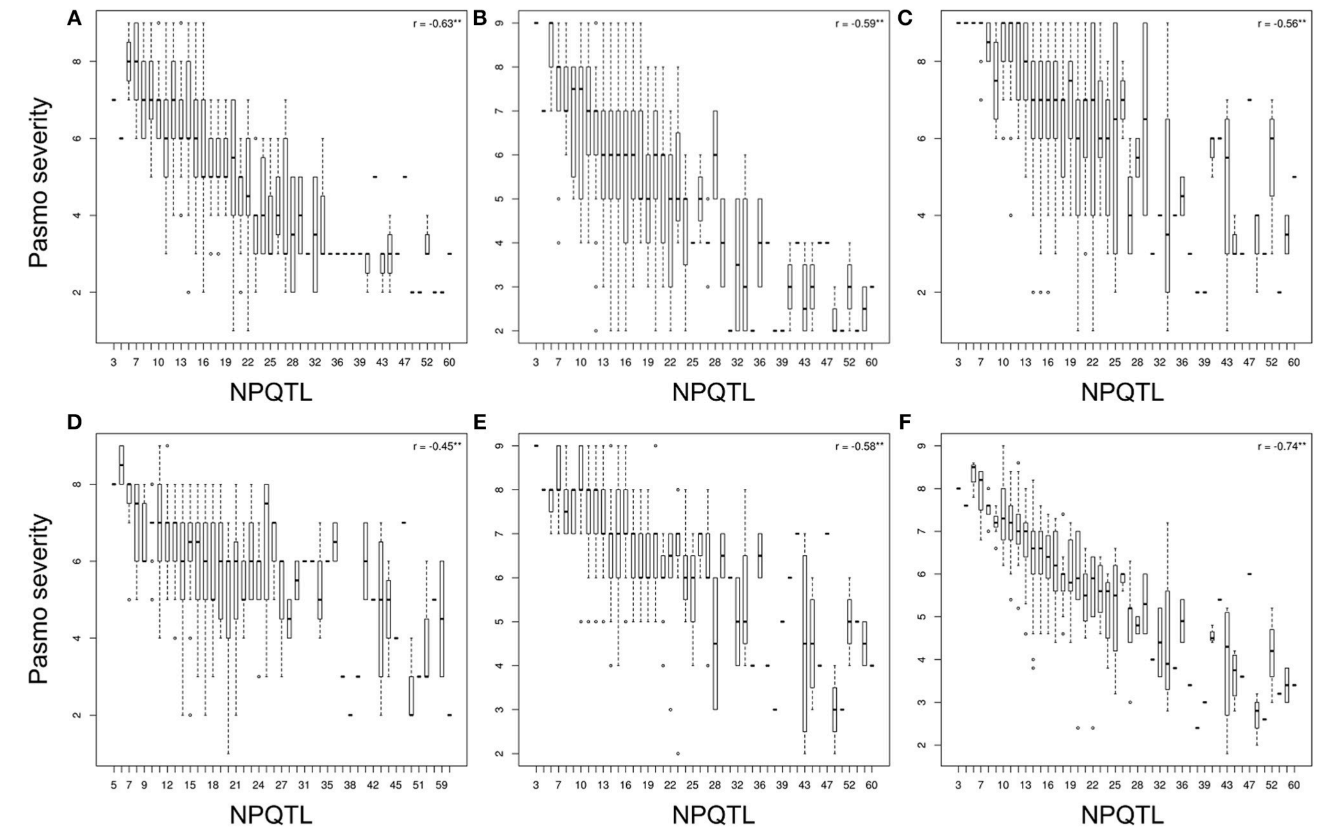


FIGURE 4 | Box plots of the number of QTL with positive-effect alleles (NPQTL) in relation to pasmo severity during 2012-2016 **(A–E)** and average pasmo severity across 5 years **(F)**; ** indicates statistical significance at the 1% probability level.

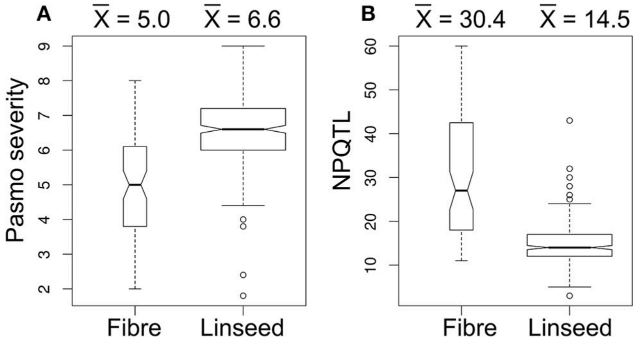


FIGURE 5 | Box plots of pasmo severity **(A)** and the number of QTL with positive-effect alleles (NPQTL) **(B)** in relation to morphotypes.

Pasmo resistance levels also varied significantly among genotypes from different geographical origins (You et al., 2017). Rainfall accumulation from June to August was significantly and positively associated with pasmo incidence and severity (Halley et al., 2004). Therefore, natural selection might be the main evolutionary pressure resulting in geographic variation. Accessions from India and Pakistan were the most susceptible of the core collection; this is not surprising considering that the environmental conditions of the flax growing regions of India are not conducive to the disease development (Diederichsen et al., 2008). On the other hand, accessions from Europe were the most resistant, a reflection of the fiber type predominance of the European germplasm (You et al., 2017) that have historically been under higher selection pressure for PR. North America appears to have the largest proportion of moderately susceptible and susceptible accessions (63 and 55) of the diversity panel, in agreement with its almost exclusive linseed germplasm (You et al., 2017). The most resistant Canadian linseed breeding line, CN101536, is only moderately resistant with an average PS of 4.4. Therefore, the incorporation of PR from linseed accession CN101367 from Georgia, as earlier noted, would benefit the improvement of PR in linseed. Interestingly, the East Asian mixed fiber and linseed germplasm is globally moderately resistant, in agreement with the long history of domestication for PR (Millam et al., 2005).

The PS of moderately resistant and susceptible accessions varied considerably across years, indicating a strong genotype × environment interaction. The low but significant correlations between the phenotypic data from any 2 years suggest the presence of interactions. In addition, the variance component analysis showed that the genotype × environment interaction accounted for a large proportion of the total variation. The

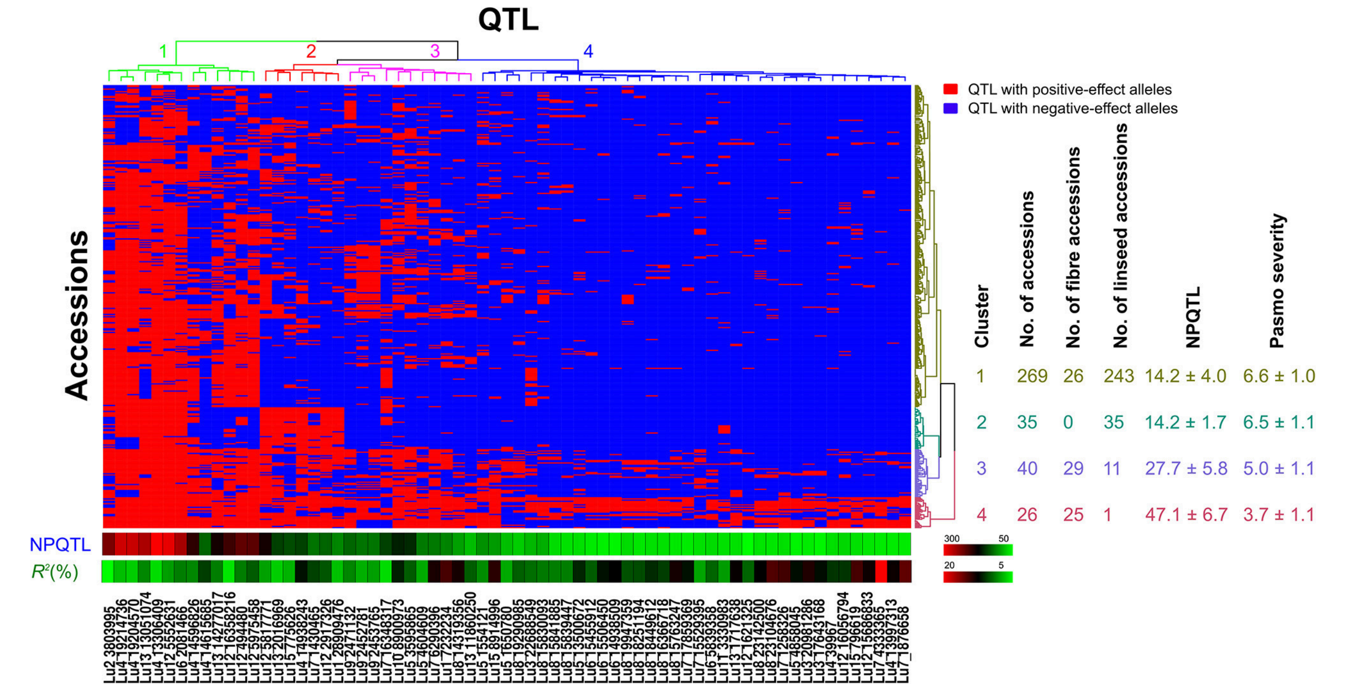


FIGURE 6 | Cluster analysis of the association panel based on a set of 67 stable large-effect QTL. The accessions were grouped into four clusters and the QTL were assigned to four sub-groups. Tag QTNs of QTL were chosen for analysis. QTL with positive-effect alleles (PQTL) in the accessions are indicated in red; blue indicates the absence of PQTL. NPQTL, the number of QTL with positive-effect alleles.

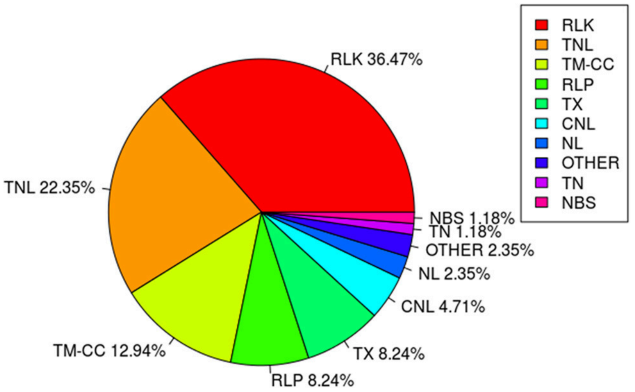


FIGURE 7 | Class distribution of resistance gene analogs (RGAs) located within 200 Kb flanking regions of QTL.

interaction partially resulted in different QTL identified in datasets from different years. QTL contribution to PR might marginally differ from year to year, stressing the need for multi-environment phenotyping to identify environment-specific QTL.

Methods Comparison

In this study, few QTL were detected by the single-locus methods MLM and GEMMA, likely as a consequence of the stringent corrected probability threshold (1.93×10^{-7}); the third single-locus method GLM and the multi-locus methods identified greater numbers of QTL. The numbers of QTL identified by GLM and mrMLM methods were similar while FarmCPU detected comparatively fewer. The phenotypic variance explained by the QTL (R^2) is also an important criterion of comparison. Both single and multi-locus methods identified some QTL with large effects (**Table 4**). However, most small-effect QTL were detected only when multi-locus methods were used. Although few QTL were common between methods, a large proportion of common QTL was observed among mrMLM methods. Thus, the complementarity between different methods is significant, and, in light of our results, the combined utilization of various statistical models is highly recommended for the identification of all potential QTL with both large and small effects.

Evaluation of Pasmo QTL in the Core Collection and Breeding Applications

Identification of QTL associated with PR can potentially facilitate their incorporation into elite germplasm, especially in North America where linseed is the main type for production (You et al., 2017). Several large-effect QTL/tag QTNs were noted, including QTL 44/Lu9-4333365 (R^2 = 23.39%), QTL 43/Lu9-1896658 (R^2 = 17.12%), QTL 39/Lu8-23104696 (R^2 = 16.53%), and QTL 1/Lu1-9232234 (R^2 = 16.17%). These were mostly present in resistant accessions as PQTL (**Table 8**) and they hold potential for MAS.

Although the large-effect QTL may be useful for MAS, a large number of small-effect QTL would be beneficial for genomic prediction (GP). GP using genome-wide markers to predict breeding values of target traits is a promising alternative method to MAS for low heritability traits including PR (Lipka et al., 2015; Poland and Rutkoski, 2016). Compared to conventional phenotypic selection, GP can accelerate genetic gains for early selection (Newell and Jannink, 2014). The accuracy and efficiency of GP models for flax PS were evaluated with three sets of QTL: 500 (SNP-500QTL), 134 (SNP-134QTL), and 67 (SNP-67QTL) which are developed in this study (He et al., 2018). The GP model built with SNP-500QTL achieved a prediction accuracy of 0.92 while the use of 134 and 67 QTL yielded accuracies of 0.75 and 0.76, respectively (He et al., 2018). The similar accuracies of the two smaller sets were expected because SNP-67QTL is essentially a non-redundant set of SNP-134QTL. These predictions serve as additional validation of the QTL identified herein and simultaneously illustrate the effectiveness of prediction models that include a full complement of large and small-effect QTL including environment-specific QTL.

Candidate Genes for Pasmo Resistance

Functional annotation of the QTL identified herein revealed 85 RGAs co-located with 45 large-effect QTL. Of them, two RGAs, *Lus10031043*/RLK and *Lus10020016*/CNL corresponding to QTL 45/Lu9-6270375 and QTL 51/Lu12-474480, respectively, may be associated with two orthologous resistance genes in Arabidopsis (Xiang et al., 2008; Saijo et al., 2009). The RLK gene *Lus10031043* is an ortholog to *AT5G20480.1* in Arabidopsis, which encodes a predicted leucine-rich repeat receptor kinase (LRR-RLK) and functions as the receptor for bacterial pathogen-associated molecular patterns (PAMPs) EF-Tu (EFR). The LRR-RLK EFR recognizes the bacterial epitopes elf18, derived from elongation factor-Tu, and triggers the plant's immune response (Saijo et al., 2009). The *Pseudomonas syringae* effector *AvrPto* is reported to bind receptor kinases, including Arabidopsis EFR (LRR-RLK EFR), to inhibit plant PAMP-triggered immunity and, to subsequently trigger strong immune responses (Xiang et al., 2008). The flax CNL gene *Lus10020016* (*RPM1*) is orthologous to *RPM1* (*AT3G07040.1*) in Arabidopsis. *RPM1* contains an N-terminal tripartite nucleotide binding site and a C-terminal tandem array of leucine-rich repeats and confers resistance to *P. syringae* strains that carry the avirulence genes *avrB* and *avrRpm1* (https://www.arabidopsis.org/). The *RPM1* gene enables dual specificity to pathogens expressing either of two unrelated *P. syringae* avirulence genes (Grant et al., 1995). The above findings hint at *Lus10031043* and *Lus10020016* as potential candidate genes for PR.

CONCLUSION

Using 10 statistical methods, a total of 500 QTL, including 67 stable and large-effect QTL and many additional small effect and environment-specific QTL were identified for PS, using a diversity panel of 370 flax accessions genotyped with 258,873 genome-wide SNPs and phenotyped in the field during 5 consecutive years. The large number of QTL identified in this study illustrates the complex genetic basis for PR in flax through a demonstration of its quantitative genetic nature and its sensitivity to environments. Multi-locus methods were able to detect small-effect QTL whereas the single-locus methods tended to identify fewer QTL of large effect. Various statistical methods identified mainly different sets of QTL, indicating the value of employing different statistical methods and multiple environment phenotypic data to capture the most comprehensive set of QTL: large, small and environment-specific. Combined utilization of multiple statistical methods is advantageous to identify QTL with small effects for traits with a complex genetic base and low heritability. The high correlation observed between PS and flax morphotype indicated that the fiber germplasm contains most of the PS QTL and constitutes an important genetic resource for flax PR breeding. The 67 large-effect QTL have potential in MAS and the 500 QTL set can be exploited for effective GP for improving pasmo resistance.

AUTHOR CONTRIBUTIONS

FY, SC, and KR conceived and designed the study. KR implemented the phenotyping of pasmo resistance. SC performed next-generation sequencing. FY conducted SNP identification and analysis. FY, LH, SC, ZY, PL, JX, and XW performed data analysis. LH, FY, and ZY prepared tables and figures. FY, LH, and JX drafted the manuscript. All authors reviewed and edited the manuscript.

FUNDING

This work was part of the Total Utilization Flax GENomics (TUFGEN) project funded by Genome Canada and other stakeholders, the A-base project (J-001004) funded by Agriculture and Agri-Food Canada. This research was also supported by Western Grains Research Foundation (WGRF), Canada-China science and technology and innovation action plan (2017ZJGH0106002) and China Scholarship Council.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01982/full#supplementary-material

REFERENCES

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi: 10.1086/521987

Buerstmayr, H., Ban, T., and Anderson, J. A. (2009). QTL mapping and marker-assisted selection for *Fusarium* head blight resistance in wheat: a review. *Plant Breed.* 128, 1–26. doi: 10.1111/j.1439-0523.2008.01550.x

Colhoun, J., and Muskett, A. E. (1943). Disease of flax. *Nature* 151:223. doi: 10.1038/151223b0

Diederichsen, A., Kusters, P. M., Kessler, D., Bainas, Z., and Gugel, R. K. (2012). Assembling a core collection from the flax world collection maintained by plant gene resources of Canada. *Genet. Resour. Crop Evol.* 60, 1479–1485. doi: 10.1007/s10722-012-9936-1

Diederichsen, A., Rozhmina, T. A., and Kudrjavceva, L. P. (2008). Variation patterns within 153 flax (*Linum usitatissimum* L.) genebank accessions based on evaluation for resistance to fusarium wilt, anthracnose and pasmo. *Plant Genet. Resour.* 6, 22–32. doi: 10.1017/S1479262108913897

Fu, Y. B., Diederichsen, A., and Allaby, R. G. (2012). Locus-specific view of flax domestication history. *Ecol. Evol.* 2, 139–152. doi: 10.1002/ece3.57

Goutam, U., Kukreja, S., Yadav, R., Salaria, N., Thakur, K., and Goyal, A. K. (2015). Recent trends and perspectives of molecular markers against fungal diseases in wheat. *Front. Microbiol.* 6:861. doi: 10.3389/fmicb.2015.00861

Grant, M. R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., et al. (1995). Structure of the arabidopsis RPM1 gene enabling dual specificity disease resistance. *Science* 269, 843–846.

Grassmann, F., Heid, I. M., Weber, B. H., and International, A. M. D. G.C. (2017). Recombinant haplotypes narrow the ARMS2/HTRA1 association signal for age-related macular degeneration. *Genetics* 205, 919–924. doi: 10.1534/genetics.116.195966

Hall, L. M., Booker, H., Siloto, R. M. P., Jhala, A. J., and Weselake, R. J. (2016). "Flax (Linum usitatissimum L.)," in *Industrial Oil Crops*, eds T. A. McKeon, D. G. Hayes, D. F. Hidebrand, and R. J. Weselake (Urbana, IL: AOCS Press), 157–194.

Halley, S., Bradley, C. A., Lukach, J. R., McMullen, M., Knodel, J. J., Endres, G. J., et al. (2004). Distribution and severity of pasmo on flax in North Dakota and evaluation of fungicides and cultivars for management. *Plant Dis.* 88, 1123–1126. doi: 10.1094/PDIS.2004.88.10.1123

He, L., Xiao, J., Rashid, K. Y., Jia, G., Li, P., Yao, Z., et al. (2018). Evaluation of genomic prediction for pasmo resistance in flax. *[Preprints] 2018*. 2018110623. doi: 10.20944/preprints201811.0623.v1

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967. doi: 10.1038/ng.695

International HapMap, C. (2005). A haplotype map of the human genome. *Nature* 437, 1299–1320. doi: 10.1038/nature04226

Jo, H., and Koh, G. (2015). Faster single-end alignment generation utilizing multi-thread for BWA. *Biomed. Mater. Eng.* 26 (Suppl 1), S1791–1796. doi: 10.3233/BME-151480

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., et al. (2008). Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709–1723. doi: 10.1534/genetics.107.080101

Kumar, J., Choudhary, A. K., Solanki, R. K., and Pratap, A. (2011). Towards marker-assisted selection in pulses: a review. *Plant Breed.* 130, 297–313. doi: 10.1111/j.1439-0523.2011.01851.x

Kumar, S., You, F. M., and Cloutier, S. (2012). Genome wide SNP discovery in flax through next generation sequencing of reduced representation libraries. *BMC Genomics* 13:684. doi: 10.1186/1471-2164-13-684

Larsson, S. J., Lipka, A. E., and Buckler, E. S. (2013). Lessons from *Dwarf8* on the strengths and weaknesses of structured association mapping. *PLoS Genet.* 9:e1003246. doi: 10.1371/journal.pgen.1003246

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. doi: 10.1093/bioinformatics/btp352

Li, H., Zhang, L., Hu, J., Zhang, F., Chen, B., Xu, K., et al. (2017). Genome-wide association mapping reveals the genetic control underlying branch angle in rapeseed (*Brassica napus* L.). *Front. Plant Sci.* 8:1054. doi: 10.3389/fpls.2017.01054

Lin, C. S., and Poushinsky, G. (1985). A modified augmented design (type 2) for rectangular plots. *Can. J. Plant Sci.* 65, 743–749. doi: 10.4141/cjps85-094

Lipka, A. E., Kandianis, C. B., Hudson, M. E., Yu, J. M., Drnevich, J., Bradbury, P. J., et al. (2015). From association to prediction: statistical methods for the dissection and selection of complex traits in plants. *Curr. Opin. Plant Biol.* 24, 110–118. doi: 10.1016/j.pbi.2015.02.010

Liu, W., Maccaferri, M., Chen, X., Laghetti, G., Pignone, D., Pumphrey, M., et al. (2017). Genome-wide association mapping reveals a rich genetic architecture of stripe rust resistance loci in emmer wheat (*Triticum turgidum* ssp. *dicoccum). Theor. Appl. Genet.* 130, 2249–2270. doi: 10.1007/s00122-017-2957-6

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12:e1005767. doi: 10.1371/journal.pgen.1005767

Miedaner, T., and Korzun, V. (2012). Marker-assisted selection for disease resistance in wheat and barley breeding. *Phytopathology* 102, 560–566. doi: 10.1094/PHYTO-05-11-0157

Millam, S., Obert, B., and Pret'ová, A. (2005). Plant cell and biotechnology studies in *Linum usitatissimum* – a review. *Plant Cell Tiss. Organ Cult.* 82, 93–103. doi: 10.1007/s11240-004-6961-6

Newell, M. A., and Jannink, J. L. (2014). Genomic selection in plant breeding. *Methods Mol. Biol.* 1145, 117–130. doi: 10.1007/978-1-4939-0446-4_10

Ogura, T., and Busch, W. (2015). From phenotypes to causal sequences: using genome wide association studies to dissect the sequence basis for variation of plant development. *Curr. Opin. Plant Biol.* 23, 98–108. doi: 10.1016/j.pbi.2014.11.008

Poland, J., and Rutkoski, J. (2016). Advances and challenges in genomic selection for disease resistance. *Annu. Rev. Phytopathol.* 54, 79–98. doi: 10.1146/annurev-phyto-080615-100056

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. doi: 10.1038/ng1847

Reich, D., Price, A. L., and Patterson, N. (2008). Principal component analysis of genetic data. *Nat. Genet.* 40, 491–492. doi: 10.1038/ng0508-491

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2017). pKWmEB: integration of Kruskal-Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120, 208–218. doi: 10.1038/s41437-017-0007-4

Saijo, Y., Tintor, N., Lu, X., Rauf, P., Pajerowska-Mukhtar, K., Häweker, H., et al. (2009). Receptor quality control in the endoplasmic reticulum for plant innate immunity. *EMBO J.* 28, 3439–3449. doi: 10.1038/emboj.2009.263

Segura, V., Vilhjalmsson, B. J., Platt, A., Korte, A., Seren, U., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

Sekhwal, M. K., Li, P., Lam, I., Wang, X., Cloutier, S., and You, F. M. (2015). Disease resistance gene analogs (RGAs) in plants. *Int. J. Mol. Sci.* 16, 19248–19290. doi: 10.3390/ijms160819248

Soto-Cerda, B., Diederichsen, A., Ragupathy, R., and Cloutier, S. (2013). Genetic characterization of a core collection of flax (*Linum usitatissimum* L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. *BMC Plant Biol.* 13:78. doi: 10.1186/1471-2229-13-78

Soto-Cerda, B. J., Duguid, S., Booker, H., Rowland, G., Diederichsen, A., and Cloutier, S. (2014a). Association mapping of seed quality traits using the Canadian flax (*Linum usitatissimum* L.) core collection. *Theor. Appl. Genet.* 127, 881–896. doi: 10.1007/s00122-014-2264-4

Soto-Cerda, B. J., Duguid, S., Booker, H., Rowland, G., Diederichsen, A., and Cloutier, S. (2014b). Genomic regions underlying agronomic traits in linseed (*Linum usitatissimum* L.) as revealed by association mapping. *J. Integr. Plant Biol.* 56, 75–87. doi: 10.1111/jipb.12118

Stich, B., Mohring, J., Piepho, H. P., Heckenberger, M., Buckler, E. S., and Melchinger, A. E. (2008). Comparison of mixed-model approaches for association mapping. *Genetics* 178, 1745–1754. doi: 10.1534/genetics.107.079707

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Comput. Biol.* 13:e1005357. doi: 10.1371/journal.pcbi.1005357

Twells, R. C., Mein, C. A., Phillips, M. S., Hess, J. F., Veijola, R., Gilbey, M., et al. (2003). Haplotype structure, LD blocks, and uneven recombination within the LRP5 gene. *Genome Res.* 13, 845–855. doi: 10.1101/gr.563703

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6:19444. doi: 10.1038/srep19444

Wang, Z., Hobson, N., Galindo, L., Zhu, S., Shi, D., McDill, J., et al. (2012). The genome of flax (Linum usitatissimum) assembled de novo from short shotgun sequence reads. *Plant J.* 72, 461–473. doi: 10.1111/j.1365-313X.2012.05093.x

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

Xiang, T., Zong, N., Zou, Y., Wu, Y., Zhang, J., Xing, W., et al. (2008). Pseudomonas syringae effector *AvrPto* blocks innate immunity by targeting receptor kinases. *Curr. Biol.* 18, 74–80. doi: 10.1016/j.cub.2007.12.020

Xie, D., Dai, Z., Yang, Z., Sun, J., Zhao, D., Yang, X., et al. (2017). Genome-wide association study identifying candidate genes influencing important agronomic traits of flax *(Linum usitatissimum* L.) using SLAF-seq. *Front. Plant Sci.* 8:2232. doi: 10.3389/fpls.2017.02232

You, F. M., Deal, K. R., Wang, J., Britton, M. T., Fass, J. N., Lin, D., et al. (2012). Genome-wide SNP discovery in walnut with an AGSNP pipeline updated for SNP discovery in allogamous organisms. *BMC Genomics* 13:354. doi: 10.1186/1471-2164-13-354

You, F. M., Duguid, S. D., Thambugala, D., and Cloutier, S. (2013). Statistical analysis and field evaluation of the type 2 modified augmented design (MAD) in phenotyping of flax (Linum usitatissimum) germplasms in multiple environments. *Aust. J. Crop Sci*. 7, 1789–1800.

You, F. M., Huo, N., Deal, K. R., Gu, Y. Q., Luo, M. C., McGuire, P. E., et al. (2011). Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics* 12:59. doi: 10.1186/1471-2164-12-59

You, F. M., Jia, G., Xiao, J., Duguid, S. D., Rashid, K. Y., Booker, H. M., et al. (2017). Genetic variability of 27 traits in a core collection of flax (*Linum usitatissimum* L.). *Front. Plant Sci.* 8:1636. doi: 10.3389/fpls.2017.01636

You, F. M., Xiao, J., Li, P., Yao, Z., Gao, J., He, L., et al. (2018a). Chromosome-scale pseudomolecules refined by optical, physical, and genetic maps in flax. *Plant J.* 95, 371–384. doi: 10.1111/tpj.13944

You, F. M., Xiao, J., Li, P., Yao, Z., Jia, G., He, L., et al. (2018b). Genome-wide association study and selection signatures detect genomic regions associated with seed yield and oil quality in flax. *Int. J. Mol. Sci.* 19:2303. doi: 10.3390/ijms19082303

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* 118, 517–524. doi: 10.1038/hdy.2017.8

Zhang, L., Li, J., Pei, Y. F., Liu, Y., and Deng, H. W. (2009). Tests of association for quantitative traits in nuclear families using principal components to correct for population stratification. *Ann. Hum. Genet.* 73(Pt. 6), 601–613. doi: 10.1111/j.1469-1809.2009.00539.x

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership