GENETICS AND GENOMICS OF POLYPLOID PLANTS

EDITED BY: Jun Yang, Zhangying Wang, Yiwei Jiang and Shuizhang Fei
PUBLISHED IN: Frontiers in Plant Science

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-083-7
DOI 10.3389/978-2-88963-083-7

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

GENETICS AND GENOMICS OF POLYPLOID PLANTS

Topic Editors:
Jun Yang, Shanghai Chenshan Plant Science Research Center (CAS), China
Zhangying Wang, Guangdong Academy of Agricultural Sciences, China
Yiwei Jiang, Purdue University, United States
Shuizhang Fei, Iowa State University, United States

Citation: Yang, J., Wang, Z., Jiang, Y., Fei, S., eds. (2019). Genetics and Genomics of Polyploid Plants. Lausanne: Frontiers Media. doi: 10.3389/978-2-88963-083-7

Table of Contents

frontiers in Plant Science

EDITORIAL
published: 18 July 2019
doi: 10.3389/fpls.2019.00934

Editorial: Genetics and Genomics of Polyploid Plants

Jun Yang[1]*, *Zhangying Wang*[2]*, *Yiwei Jiang*[3]* and *Shuizhang Fei*[4]*

[1] Shanghai Chenshan Plant Science Research Center (CAS), Shanghai, China, [2] Guangdong Academy of Agricultural Sciences, Guangzhou, China, [3] Department of Agronomy, Purdue University, West Lafayette, IN, United States, [4] Department of Horticulture, Iowa State University, Ames, IA, United States

Keywords: polyploid, genetics, genomics, evolution, genotype by environment (G × E) interaction

Editorial on the Research Topic

Genetics and Genomics of Polyploid Plants

Many of the most economically important crops are polyploids or paleopolyploids. The Research Topic "Genetics and Genomics in Polyploid Plants" focuses on the genome relationships and evolution; inheritance of economically important traits; and development of technologies or methods that facilitate genetics and genomics studies in polyploidy plants. The knowledge gained through studying the evolutionary and population genetics of polyploid crops will advance our understanding of the genetic origin of polyploid crops and its impact on phenotypic variation and facilitate crop improvement. This topic contains 3 reviews, 13 original research articles and 1 method paper.

Kyriakidou et al. reviewed the current genome sequencing strategies in polyploidy plants. Generally, these strategies can also be applied to investigate other organisms with sub-genomes. Ceasar et al. discussed up-to-date genetic improvement, transcriptome analysis and quantitative trait loci mapping using the released genome sequence of finger millet (*Eleusine coracana*). Kandel et al. provided a review update on the available genetic and genomic resources for sugarcane (*Saccharum* spp.) and discussed the challenges in sugarcane biomass improvement. Sharma et al. compared *Saccharum officinarum* and *Saccharum spontaneum* genomes via sequencing of bacterial artificial chromosome libraries. Divergence time estimation suggested that both *Saccharum* spp. experienced independent polyploidization. Dong et al. developed a set of chromosome-specific cytogenetic markers in *Saccharum* spp. via pooled sequencing of bacterial artificial chromosome clones derived from haploid genome of cultivated sugarcane (*S. spontaneum*). Maulana et al. reported quantitative trait loci associated with seedling heat tolerance in winter wheat (*Triticum aestivum*). Daba et al. identified candidate genes that are putatively responsible for determination of kernel weight and kernel length in soft red winter wheat population. Ren et al. studied the genetic architecture of nitrogen-deficiency tolerance in wheat seedlings using a nested association mapping population. Deng et al. examined the chromosome configurations in four artificially constructed pentaploid wheats (BBAAD) and the chromosome instabilities in their progenies. Lu et al. presented us a finely assembled genome of *Arachis ipaensis*, the B-genome progenitor of peanut (*Arachis hypogaea*). As a result, genes related to disease resistance, drought adaptation, nitrogen fixation, and fatty acid biosynthesis were identified and discussed. Clevenger et al. introduced HAPLOSWEEP, the haplotype-based genotyping software, which is applicable in allopolyploid genome analysis and has been applied in *A. hypogaea* as an example. Jia et al. studied accuracies of three Bayes statistical methods on genomic prediction of agronomic traits in alfalfa (*Medicago sativa*). Adhikari et al. suggested that the genetic basis of winter hardiness and fall dormancy in alfalfa were independent and therefore could be improved separately in breeding programs. Wadl et al. investigated the population

OPEN ACCESS

Edited and reviewed by:
Rongling Wu,
Pennsylvania State University,
United States

***Correspondence:**
Jun Yang
jyang03@sibs.ac.cn
Zhangying Wang
wzhying@hotmail.com
Yiwei Jiang
yjiang@purdue.edu
Shuizhang Fei
sfei@iastate.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *03 May 2019*
Accepted: *03 July 2019*
Published: *18 July 2019*

Citation:
Yang J, Wang Z, Jiang Y and Fei S (2019) Editorial: Genetics and Genomics of Polyploid Plants. Front. Plant Sci. 10:934. doi: 10.3389/fpls.2019.00934

structure and genetic diversity of the hexaploid sweetpotato (*Ipomoea batatas*) accessions originating from Africa, Australia, Caribbean, Central America, Far East, North America, Pacific Islands, and South America. Taylor et al. performed a genome-wide association study, and significant signals for heading and anthesis were identified in switchgrass (*Panicum virgatum*). At the same time, they also highlighted the potential of manipulating these candidate genes in late-flowering switchgrass breeding. Guo et al. implemented genomic prediction in tetraploid perennial ryegrass (*Lolium perenne*), and gained the highest predictive ability by sequencing depth between 10 and 20 times. Yu et al. presented a hypothetical molecular regulatory network for apetalous trait through the CHR11-PLP pathway in rapeseed (*Brassica napus*).

The research articles featured in this collection covers a range of polyploid plant species including wheat, sugarcane, alfalfa, peanut, sweetpotato, switchgrass, finger millet, perennial ryegrass, and rapeseed. This research collection greatly advanced our understanding of polyploid plants. Many polyploid species are notoriously difficult to study because of their complex genomes, unique reproductive mode, and in some cases long life cycle. Rapid and continued improvements on sequencing capacity and accuracy, computing power, bioinformatics tools, phenotyping throughput and accuracy and other technologies will undoubtedly enhance genetic and genomic research in these species in near future.

AUTHOR CONTRIBUTIONS

JY drafted the manuscript. YJ, SF, and ZW edited and revised the manuscript.

ACKNOWLEDGMENTS

JY acknowledges funding support by National Key R&D Program of China (2018YFD1000700-2018YFD1000701-4), Shanghai Municipal Afforestation & City Appearance and Environmental Sanitation Administration (G182402, G192413, and G192414) and Youth Innovation Promotion Association CAS.

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

REVIEW
published: 21 November 2018
doi: 10.3389/fpls.2018.01660

Current Strategies of Polyploid Plant Genome Sequence Assembly

Maria Kyriakidou[1], *Helen H. Tai*[2], *Noelle L. Anglin*[3], *David Ellis*[3] *and Martina V. Strömvik*[1]*

[1] Department of Plant Science, McGill University, Montreal, QC, Canada, [2] Fredericton Research and Development Centre, Agriculture and Agri-Food Canada, Fredericton, NB, Canada, [3] International Potato Center, Lima, Peru

Polyploidy or duplication of an entire genome occurs in the majority of angiosperms. The understanding of polyploid genomes is important for the improvement of those crops, which humans rely on for sustenance and basic nutrition. As climate change continues to pose a potential threat to agricultural production, there will increasingly be a demand for plant cultivars that can resist biotic and abiotic stresses and also provide needed and improved nutrition. In the past decade, Next Generation Sequencing (NGS) has fundamentally changed the genomics landscape by providing tools for the exploration of polyploid genomes. Here, we review the challenges of the assembly of polyploid plant genomes, and also present recent advances in genomic resources and functional tools in molecular genetics and breeding. As genomes of diploid and less heterozygous progenitor species are increasingly available, we discuss the lack of complexity of these currently available reference genomes as they relate to polyploid crops. Finally, we review recent approaches of haplotyping by phasing and the impact of third generation technologies on polyploid plant genome assembly.

Keywords: polyploidy, plant genomics, genome assembly, third generation sequencing, reference genome

OPEN ACCESS

Edited by:
Zhangying Wang,
Guangdong Academy of Agricultural Sciences, China

Reviewed by:
Danny W.-K. Ng,
The Chinese University of Hong Kong, China
Prathima Perumal Thirugnanasambandam,
Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Australia

***Correspondence:**
Martina V. Strömvik
martina.stromvik@mcgill.ca

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *13 April 2018*
Accepted: *25 October 2018*
Published: *21 November 2018*

Citation:
Kyriakidou M, Tai HH, Anglin NL, Ellis D and Strömvik MV (2018) Current Strategies of Polyploid Plant Genome Sequence Assembly. Front. Plant Sci. 9:1660. doi: 10.3389/fpls.2018.01660

INTRODUCTION TO POLYPLOIDY

The fusion of two or more genomes within one nucleus results in polyploidy, resulting in each cell containing more than two pairs of homologous chromosomes. Polyploidy occurs in the majority of angiosperms and is important in agricultural crops that humans depend on for survival. Examples of important polyploid plants used for human food include, *Triticum aestivum* (wheat), *Arachis hypogaea* (peanut), *Avena sativa* (oat), *Musa* sp. (banana), many agricultural *Brassica* species, *Solanum tuberosum* (potato), *Fragaria ananassa* (strawberry), and *Coffea arabica* (coffee). Autopolyploidy results from whole genome duplication, while an allopolyploid is characterized by interspecific or intergeneric hybridizations followed by chromosome doubling (Doyle et al., 2008; Chen, 2010). Genome duplication (autoployploidy) can be a source of genes with novel functions leading to new phenotypes and novel mechanisms for adaptation (Crow and Wagner, 2005). Autopolyploids typically suffer from reduced fertility whereas allopolyploids have potential for heterosis or hybrid vigor (Ramsey and Schemske, 1998). Polyploidy generates great genetic, genomic, and phenotypic novelty (Soltis et al., 2016); however, the higher complexity between genotype and phenotype in polyploids compared to diploid plants makes linking genotype to phenotype a challenging task. For example, allopolyploid plant cells have complex regulatory mechanisms in order to unify gene expression between the homeologs and define their relative contributions to the final phenotype. Hence, polyploidization is one of the major forces of plant

evolution and is intimately linked to speciation and diversity (Bento et al., 2011). It is estimated that around 80% of all living plants are polyploids (Meyers and Levin, 2006), while many plant lineages including monocots (i.e., *Oryza*) and eudicots (*Arabidopsis*) have at least one paleo-polyploidy event in their history.

OVERVIEW OF THE SEQUENCING TECHNIQUES AND THEIR APPLICATIONS IN POLYPLOID PLANT GENOMES

Genome sequencing was initiated in the mid 1970's with alternative methods to determine the composition of DNA in a target cell or organism (Sanger and Coulson, 1975; Maxam and Gilbert, 1977). The first whole genome to be sequenced was that of a bacteriophage PhiX (Sanger et al., 1977a) with a genome size at 5.3 Kb. However, the revolution in sequencing technology came about when Sanger developed the chain termination or dideoxy method (Sanger et al., 1977b). This technique, now known as Sanger sequencing, was adopted by most molecular biology laboratories and was the primary method of sequencing for 30+ years allowing sequencing of fragments of approximately 800–1,000 bp.

It took over 20 years from the time the first genome of a bacteriophage was sequenced until plant biologists had a draft genome of a flowering plant. First to be sequenced was the genome of *Arabidopsis thaliana*, a small weedy plant (Arabidopsis Genome Initiative, 2000). After the release of the Arabidopsis genome sequence, economically important crops such as *Oryza sativa* (rice), *Carica papaya* (papaya), and *Zea mays* (maize) were sequenced using Sanger sequencing (International Rice Genome Sequencing Project, 2005; Ming et al., 2008; Schnable et al., 2009). Yet, of these plant genomes, only rice and Arabidopsis were sequenced using the Bacterial Artificial Chromosome (BAC) approach, and thus, are more complete genomes, whereas the others are drafts in a less completed stage (Claros et al., 2012).

The diploidized tetraploid genome of *Glycine max* (soybean) was the first polyploid plant genome released (publicly available in early 2008, Schmutz et al., 2010), followed by the tetraploid *Arabidopsis lyrata* (Hu et al., 2011) (**Table 1**). The soybean project was very costly, and the resulting assembly consisted of the largest published plant genome performed using the Sanger Whole Genome Sequencing (WGS) method. In 2011, the genome of *Jatropha curcas* (an oil-bearing tree) that has variable ploidy levels (**Table 1**), was also sequenced using the Sanger method (Sato et al., 2010). The assembly of the complex tetraploid genome of cultivated cotton—*Gossypium arboretum* (Li et al., 2014) was followed by the reference genome of wheat, derived from the assembly of the large complex genome of *Aegilops tauschii*, one of the three diploid progenitors of bread wheat (Zimin et al., 2017).

Next Generation Sequencing (NGS) technologies became commercially available in 2004 (Mardis, 2008) reducing sequencing costs and increasing massively sequencing throughputs, but also expanding the complexity of fragment assembly due to its short-sequence read output. NGS allows genome sequencing to be performed with lower DNA concentrations and thus, has applications in genome sequencing and re-sequencing, metagenomics, transcriptomics (RNA-sequencing) and even in personal genomics (personal medicine). These techniques can reduce the gap between genotype and phenotype by combining for example genomics and transcriptomics data. Some of the NGS platforms that have been employed in recent years include: 454 or pyro-sequencing (by Roche, Basel, Switzerland, with read lengths up to 700 bp), SOLiD (by Life Technologies, Carlsbad, California, 50 bp), HiSeq (by Illumina, San Diego, California, 2 × 250 bp), MiSeq (by Illumina, 2 × 300 bp) and Ion Torrent/Proton (by Life Technologies, 200 bp). NGS technologies are advantageous because, unlike Sanger sequencing, DNA cloning is not required making the process simpler, with greater adaption for a broad range of biological phenomena, and massive parallelization at decreased costs. However, NGS does suffer from some disadvantages: the short sequence length requires unique assembly algorithms, base calling is less accurate than Sanger sequencing, and the quality of NGS assemblies is lower than those made from Sanger sequence (Claros et al., 2012). Examples of polyploid plant genomes sequenced using Illumina technology are the first assembly of the hexaploid *T. aestivum* (wheat) genome (Choulet et al., 2010), and the genome of *G. hirsutum* (cotton) (Li et al., 2015). The genomes of *Brassica oleracea* (cabbage) and *B. napus* (rapeseed) (Chalhoub et al., 2014) were sequenced with a combination of 454 and Illumina technologies. A genome assembly service using only high-quality short Illumina reads is offered by NRGene's DenovoMAGIC platform (http://www.nrgene.com/technology/denovomagic/). The recently annotated allohexaploid wheat genome was constructed using DenovoMAGIC2 (International Wheat Genome Sequencing Consortium (IWGSC), 2018). The latest version; DenovoMAGIC v 3.0 promises production of long, phased scaffolds using only NGS.

The emergence of the Third Generation Sequencing technologies consists of the most recent genome sequencing approaches, characterized by long reads. These methods have further reduced sequencing costs, simplified preparatory and sequencing methods (Schadt et al., 2010), while providing longer read lengths, typically measured in kilo bases (Kb) rather than bases (bp). While there are many upsides to this new technology, caveats include high error rates and a requirement for very high-quality DNA. However, these approaches currently look promising in meeting the challenges of sequencing and assembling large, repetitive, and complex plant genomes by the production of large quantities of long reads to help bridge difficult regions in the genome. There are currently two types of technologies included in the Third-Generation sequencing approaches: long-read sequencing and long-range scaffolding technologies (Jiao and Schneeberger, 2017).

Among the long-read sequencing technologies, the most widely used technology is the Pacific Biosciences' Single Molecule Real-Time (SMRT), with an average read length 20 Kb. For the assembly of the *Chenopodium quinoa* genome, a read length of ∼12 Kb was reported using this technology (Jarvis et al., 2017). Additionally, Illumina introduced another long-read technology, the Synthetic Long-Reads (SLR) from short-read sequencing data, with a median length of 8–10Kb (**Table 2**). However, a

TABLE 1 | Sequenced plant polyploid genomes through May 2018.

NA	Organism name	Genome size (Mb)	Current status	1st Release date in NCBI	Ploidy level	References/center
1	*Arabidopsis lyrata* subsp *lyrata*	206.823	Scaffold	2009-11-30	Tetraploid	Hu et al., 2011
2	*Glycine max*	978.972	Chromosome	2010-01-05	Allotetraploid	Schmutz et al., 2010
3	*Triticum aestivum*	15344.7	Chromosome 3B	2010-07-15	Allohexaploid	Choulet et al., 2010
4	*Solanum tuberosum*	705.934	Scaffold	2011-05-24	Autotetraploid	Potato Genome Sequencing Consortium, 2011
5	*Actinidia chinensis*	604.217	Contig	2013-09-16	Tetraploid	Huang et al., 2013
6	*Fragaria orientalis*	214.356	Scaffold	2013-11-27	Tetraploid	Hirakawa et al., 2014
7	*Fragaria x ananassa*	697.762	Scaffold	2013-11-27	Allooctaploid	Hirakawa et al., 2014
8	*Beta vulgaris*	566.55	Chromosome	2013-12-18	2n, 4n (Beyaz et al., 2013)	Dohm et al., 2014
9	*Oryza minuta*	45.1659	Chromosome	2014-04-16	Tetraploid	Oryza Chr3 Short Arm Comparative Sequencing Project
10	*Camelina sativa*	641.356	Chromosome	2014-04-17	Hexaploid	Kagale et al., 2014
11	*Brassica napus*	976.191	Chromosome	2014-05-05	Allotetraploid	Chalhoub et al., 2014
12	*Brassica oleracea* var. oleracea	488.954	Chromosome	2014-05-22	Hexaploid	NCBI
13	*Nicotiana tabacum*	3643.47	Scaffold	2014-05-29	Allotetraploid	Sierro et al., 2014
14	*Eragrostis tef*	607.318	Scaffold	2015-04-08	Allotetraploid	Cannarozzi et al., 2014
15	*Gossypium hirsutum*	2189.14	Chromosome	2015-04-29	Allotetraploid	Li et al., 2015
16	*Zoysia japonica*	334.384	Scaffold	2016-03-15	Tetraploid	Tanaka et al., 2016
17	*Zoysia matrella*	563.439	Scaffold	2016-03-15	Allotetraploid	Tanaka et al., 2016
18	*Zoysia pacifica*	397.01	Scaffold	2016-03-15	Allotetraploid	Tanaka et al., 2016
19	*Musa itinerans*	455.349	Scaffold	2016-05-21	2n, 3n hybrids (Wu et al., 2016)	South China Botanic Garden, CAS
20	*Rosa x damascena*	711.72	Scaffold	2016-06-13	Tetraploid	BIO-FD & C CO., LTD
21	*Chenopodium quinoa*	1333.55	Scaffold	2016-07-11	Tetraploid	Jarvis et al., 2017
22	*Brassica juncea* var. tumida	954.861	Chromosome	2016-07-19	Allotetraploid	Zhejiang University
23	*Hibiscus syriacus*	1748.25	Scaffold	2016-07-29	2n, 3n, 4n (Van Huylenbroeck et al., 2000)	Korea Research Institute of Science and Biotechnology (Kim et al., 2017)
24	*Gossypium barbadense*	2566.74	Scaffold	2016-10-28	Tetraploid	Huazhong Agricultural University
25	*Momordica charantia*	285.614	Scaffold	2016-12-27	2n to 6n (Kausar et al., 2015)	Urasaki et al., 2016
26	*Drosera capensis*	263.788	Scaffold	2016-12-30	Tetraploid (Rothfels and Heimburger, 1968)	Butts et al., 2016
27	*Capsella bursa-pastoris*	268.431	Scaffold	2017-01-29	Tetraploid	Lomonosov Moscow State University
28	*Saccharum* hybrid cultivar	1169.95	Contig	2017-03-03	It varies (D'Hont, 2005)	Riaño-Pachón and Mattiello, 2017
29	*Xerophyta viscosa*	295.462	Scaffold	2017-03-31	Hexaploid	Costa et al., 2017
30	*Triticum dicoccoides*	10495	Chromosome	2017-05-18	Tetraploid	WEWseq consortium
31	*Utricularia gibba*	100.689	Chromosome	2017-05-31	16-ploid	Lan et al., 2017
32	*Eleusine coracana*	1195.99	Scaffold	2017-06-08	Allotetraploid	Hittalmani et al., 2017
33	*Dioscorea rotundata*	456.675	Chromosome	2017-07-28	Tetraploid	Iwate Biotechnology Research Center
34	*Ipomoea batatas*	837.013	Contig	2017-08-26	Autohexaploid	Yang et al., 2017
35	*Echinochloa crus-galli*	1486.61	Scaffold	2017-10-23	Hexaploid	Zhejiang University
36	*Pachycereus pringlei*	629.656	Scaffold	2017-10-31	Autotetraploid	Zhou et al., 2017

(Continued)

TABLE 1 | Continued

NA	Organism name	Genome size (Mb)	Current status	1st Release date in NCBI	Ploidy level	References/center
37	*Olea europaea*	1141.15	Chromosome	2017-11-01	2n, 4n, 6n (Besnard et al., 2007)	Unver et al., 2017
38	*Monotropa hypopitys*	2197.49	Contig	2018-01-03	Hexaploid	Institute of Bioengineering, RAS
39	*Dactylis glomerata*	839.915	Scaffold	2018-01-19	Autotetraploid	Sichuan Agricultural University
40	*Panicum miliaceum*	848.309	Scaffold	2018-01-23	Allotetraploid	China Agricultural University
41	*Euphorbia esula*	1124.89	Scaffold	2018-02-06	Hexaploid	USDA-ARS
42	*Santalum album*	220.961	Scaffold	2018-02-12	2n, 4n etc (Xin-Hua et al., 2010)	Center for Cellular and Molecular Platforms
43	*Avena sativa*	67.3266	Contig	2018-02-26	Hexaploid	The Sainsbury Laboratory
44	*Panicum miliaceum*	850.677	Chromosome	2018-04-09	Tetraploid	Shanghai Center for Plant Stress Biology
45	*Arachis monticola*	2618.65	Chromosome	2018-04-23	Tetraploid	Henan Agricultural University
46	*Arachis hypogaea*	2538.28	Chromosome	2018-05-02	Allotetraploid	International Peanut Genome Initiative
47	*Artemisia annua*	1792.86	Scaffold	2018-05-08	Tetraploid	Shen et al., 2018

The release date refers to the first release of the genomes in NCBI, before any improvement of the assemblies. Some have been updated after this date.

maximum length of ∼21 Kb was achieved in a sugarcane hybrid sequencing project (Riaño-Pachón and Mattiello, 2017). SLR can be used to resolve the haplotype of individuals, which is highly desired in the case of polyploid plant genomes. Finally, Nanopore, introduced by Oxford Nanopore Technologies, can generate a median length greater than 5 Kb, however a ∼12 Kb median length was reported while sequencing the wild *Solanum pennellii* genome (Schmidt et al., 2017).

Even with the rapid progress and improvement of long-read technologies, it is still not possible to assemble a complete diploid plant genome using only NGS sequencing reads (Jiao and Schneeberger, 2017). Hence, long-range scaffolding technologies are essential for improving the contiguity of an assembly, which requires the extension of the contigs into scaffolds and eventually their alignment into chromosomes. Based on currently available sequencing technologies, additional genetic and physical maps are required. An alternative approach is based on chromosome conformation capture sequencing (Hi-C) provided by Dovetail Genomics (https://dovetailgenomics.com/) and PhaseGenomics (https://phasegenomics.com/), which creates long-range mate pair data for NGS (Lieberman-Aiden et al., 2009; van Berkum et al., 2010). The generated data can be used for phasing and scaffolding, which captures the entire eukaryotic chromosomes when they are combined with high quality draft assemblies (Sedlazeck et al., 2018). Genome phasing is the identification of the alleles in each of the chromosomes. The most recent announcement of the PhaseGenomics Biotechnology company is its collaboration with Pacific Biosciences for the release of FALCON-Phase (Kronenberg et al., 2018). FALCON-Phase tool promises to solve the haplotyping problem in diploids, by enabling the construction of fully-phased chromosome-scale assemblies by combining SMRT long reads and Hi-C data. The latest technology is from GemCode, introduced by 10X Genomics in 2015 (www.10xgenomics.com). This approach is similar to the SLR protocol of Illumina, but it can process longer fragments and it does not require as much read depth as the SLR. The average read length captured with this approach can be greater than 100 Kb (**Table 2**).

CHALLENGES OF POLYPLOID GENOME ASSEMBLY

A reference genome is a digital, linear nucleic acid sequence containing only a single set of chromosomes plus any unanchored heterozygous contigs and/or scaffolds. A reference genome is used to observe variations across different individuals within a species, to study evolution and to aid genome assembly. In the case of a polyploid genome, things become more complicated. For an allopolyploid organism, a reference genome contains the assembled DNA sequences of the ancestors subgenomes (e.g., *F. ananassa*, *B. napus*, *A. hypogaea*, *G. hirsutum*, *and T. aestivum*) in addition to any unanchored sequences that are kept in additional pseudochromosome(s) (e.g., *T. aestivum*, *S. tuberosum*), and for an autopolyploid organism the genome that went through the duplication event(s) (e.g., *S. tuberosum*) in addition to any unanchored sequences. It does not necessarily represent any allelic variation present in the individuals. When high throughput sequencing reads are

TABLE 2 | Third generation sequencing platforms.

Technology	Reads	Drawbacks	Plant assembly
PacBio	Single molecule long-reads, average length ∼ 10–18 Kb	False insertions in the raw reads, high error rate. Error correction algorithms are required	*Chenopodium quinoa* (Jarvis et al., 2017)
Oxford Nanopore	Single molecule long-reads, average length ∼ 10 Kb, max 100 Kb	Raw reads with false deletions and homopolymer errors. Requirement for error correction algorithms	*S. pennellii*, *A. thaliana, O. coaectata* (Mondal et al., 2017; Schmidt et al., 2017; Michael et al., 2018)
Illumina Synthetic Long reads	Synthetic long-reads derived from the short sequencing reads, average length ∼ 100 Kb	High rate false indels (insertions, deletions). They require good trimming, correction algorithms	*Saccharum* sp. (Riaño-Pachón and Mattiello, 2017)
10X Genomics	Linked reads derived from short-read sequences, average length ∼ 100 Kb*	Needs designed algorithms and aligners, poor resolution of locally repetitive sequences. Sparse sequencing	*Capsicum annuum* (Hulse-Kemp et al., 2018)
BioNano Genomics	Optical mapping of long, fluorescently labeled DNA fragments, average length ∼ 250 Kb	Not many algorithms available for a reliable alignment between the optical map and the genome assembly	*Brassica juncea* (Yang et al., 2016)
Hi-C	Pairs short reads with an average length ∼ 100 bp, method originally developed to study the 3D folding of the genome	Scattered sequencing with variable genomic distance between pairs	*Triticum aestivum* (International Wheat Genome Sequencing Consortium (IWGSC), 2014)

**10X Genomics is very similar to Illumina's SLR, with the difference that 10X Genomics can process more and larger fragments and the assemble of the different fragments does not necessarily depend on the sequencing coverage. Illumina's SLR system synthesizes the sequences of DNA fragment in contrast to 10x Genomics where the reads show only a part of DNA fragments. NA, not applicable.*

mapped to a reference genome, alternate alleles can be retrieved from each genomic region, based on the sequencing coverage and diversity in the individual compared to the reference. These alternate alleles for an organism can be detected and used for haplotype assembly for each of the present haplotypes. Polyploid assembly is similar to the sum of a number of problems of haplotype reconstruction (Aguiar and Istrail, 2013); hence, the computational complexity increases with higher ploidy. This means that the genome assembly of an n-ploid organism will result in the construction of n numbers of haplotypes. This is not an easy task as the knowledge of one haplotype does not automatically determine how to phase others (Motazedi et al., 2017).

Whole-genome duplication events have also been associated with genome rearrangement, atypical recombination, transposable element activation, meiotic/mitotic defects, and intron expansions and DNA deletion (Hufton and Panopoulou, 2009). The assembly of autopolyploid genomes is extremely challenging as fragments of a subgenome might be assigned to the wrong subgenome, which results in misassembled false genomes. Allopolyploids may present the same challenge, but given the greater genetic distance, resolving their subgenomes is likely less problematic during assembly. These events multiply the regular challenges of plant genome sequence assembly, such as repeat content, transposable elements, high heterozygosity, gene content and gene families of non-coding RNAs due to their repetitiveness after duplication events and the fact that their detection is crucial for proper genome annotation.

Polyploidization can lead to higher levels of heterozygosity, which can be confounded in asexually propagated plants such as potato causing greater difficulties in the identification of haplotypes. This is due to multiple alleles from the same locus being mistaken as sequences from different loci (Huang et al., 2017). This is especially problematic when using short sequence reads for genotyping or genome assembly, because the results will be highly fragmented assemblies with a total assembly size longer than expected. In addition, contigs can break at polymorphic regions or misassemblies can occur between large-scale duplications (Claros et al., 2012). This assembly problem is not unique to polyploid plants, however and can also occur in plants with segmental genome duplications.

The ploidy level of the plant genome must be carefully considered when choosing the appropriate assembly algorithm. The presence of two or more sets of genes within the same nucleus can affect the accuracy of the assembly, making it difficult to differentiate between homologs or homeologs (Claros et al., 2012). Glover et al. (2016) define homeologs as pairs of genes or chromosomes in the same species, derived by speciation but brought back to the same genome after a polyploidization event(s). Identifying functionally conserved homeologs however, provides important genetic material for crop improvement in many crops, including *Musa acuminata* (banana), *S. tuberosum* (potato), *Gossypium hirsutum* (cotton) and *T. aestivum* (wheat) (Chen and Dubcovsky, 2012; Glover et al., 2016). Examples of how polyploids also confer emergent properties are seed oil accumulation in *Brassica napus* (canola), spinnable fibers in cotton, and grain composition in wheat (Michael and VanBuren, 2015).

As mentioned above, several complex polyploidy plant genomes have been sequenced. The decreasing costs of NGS technologies led to the sequencing and assembly of a number of polyploid plant genomes using these technologies (**Table 1**). Based on NCBI database (data retrieved on the 4th of July 2018: https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/), 320 land plants, 47 of which are polyploid, have been sequenced (as of 4th of July of 2018). Of the 72 assembled in 2017, 19 are polyploid, and three were released in January 2018. Only 16 polyploid plant genomes have been assembled into

chromosomes, 26 assembled into scaffolds, and the rest (5) are still contigs (**Table 1**).

TECHNOLOGY-RELATED CHALLENGES

There are two basic approaches to genome assembly. Comparative assembly is a reference guided method that uses the sequences of already assembled related organisms, a reference genome, for guidance. *De novo* assembly targets organisms that have not been sequenced before (Pop, 2009), putting together the pieces without guidance from a prior reference genome. The two approaches are not completely mutually exclusive, because even in cases where reference genomes are available, regions that varied in the newly sequenced target genome need to be assembled *de novo*. Different approaches of guided and *de* novo genome assemblies can be found in **Figures 1**, **2**. The reference guided comparative assembly approach (**Figure 1**) can be performed in two ways: mapping short or long reads against the reference to construct a consensus (**Figures 1A,C**) or assembling the reads *de novo* and then use the reference genome to orientate the resulting contigs or scaffolds in an alignment and identify misassembled regions (**Figures 1B,D**) (Lischer and Shimizu, 2017).

The reference-based comparative assembly approach is usually used when genomes are re-sequenced, or to correct misassemblies or extend existing contigs of already assembled genomes (**Figures 1B,D**), and also for variant detection (**Figures 1A,C**) and haplotype construction. An assembled genome sequence is used as a reference and the sequenced reads are independently aligned against this sequence. Dynamic programming is used to identify the optimal alignment for the candidate positions that match the best. Structural variations (such as insertions or deletions) in the re-sequenced genome(s) tend to increase the complexity of the alignment. The resulting alignment allows the extraction of the structural variants and construction of the haplotypes.

The *de novo* genome assembly method is applied when a reference genome sequence does not exist for a closely related species. In this case, the genome sequence is constructed through overlapping sequenced reads, usually using graph-based algorithms. It is difficult to perform *de novo* genome assembly, especially when only shorter reads are available. Both single end (SE) and paired-end (PE) reads are difficult to assemble *de novo*, with SE reads being slightly more challenging (i.e., Illumina, **Figure 2A**). Long range reads can be used (**Figure 2B**), or a hybrid approach can be applied, where shorter and longer reads can be used together for a better assembly (**Figures 2C,D**). As for the assessment, there are currently no unified assembly quality metrics to assess the quality of the *de novo* generated assembly, although one value that is commonly used is the N50. The value of N50 is a weighted median for when at least 50% of the assembly is contained in contigs or scaffolds of equal or greater length.

In general, the comparative method requires less computation as the sequenced reads are aligned to a reference genome. However, significant bias can occur in the comparative genome approach, as divergent (duplicated) regions of the genome may

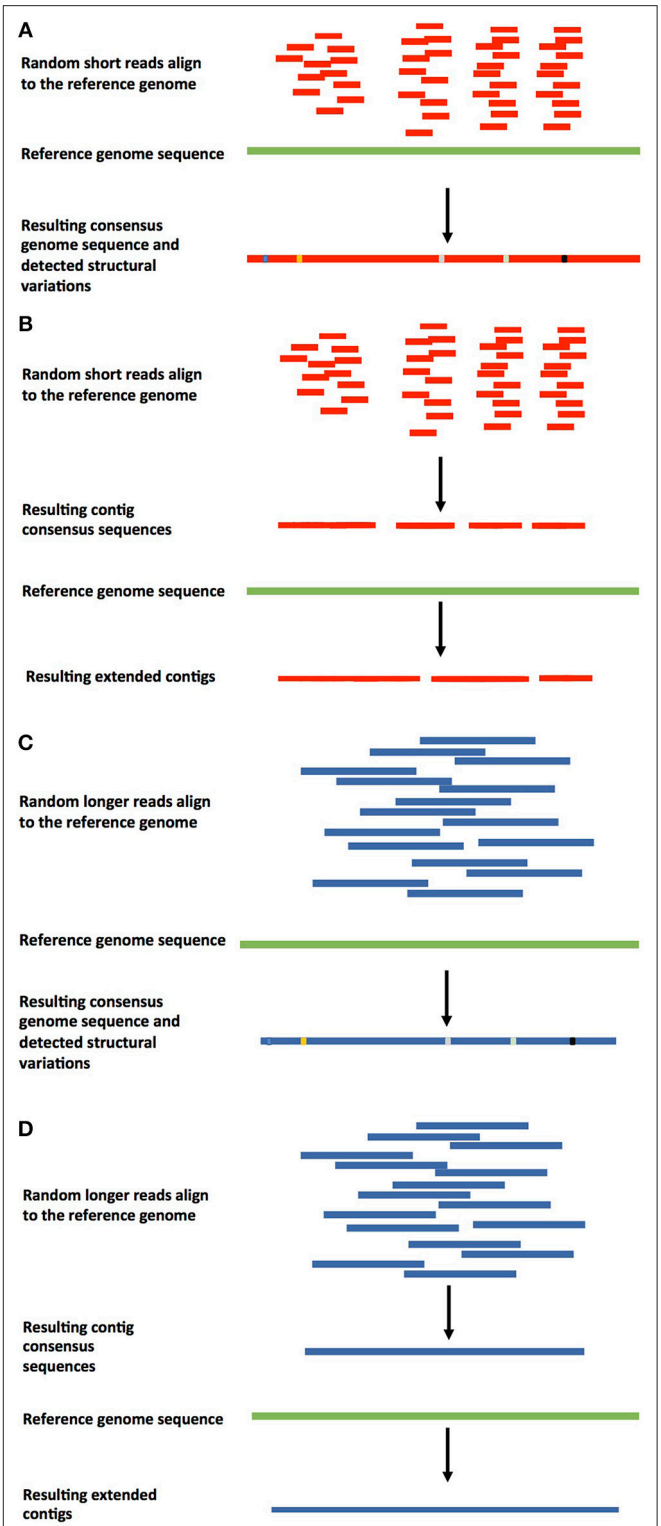

FIGURE 1 | Approaches for reference-based genome assembly. **(A)** Shorter-read guided assembly. In this method, shorter reads are aligned against the reference genome, a consensus assembly is generated, and structural variations are detected. It can also be used to detect contamination in the sequenced reads. This approach is used when genomes are

(Continued)

FIGURE 1 | re-sequenced to detect polymorphisms in individuals. **(B)** Guided *de novo* genome assembly of shorter reads. Previously *de novo* assembled shorter reads are aligned against the reference or a closely related genome to extend the existing contigs. **(C)** Longer-read guided assembly. Longer reads are aligned against the reference genome, a consensus genome assembly is constructed, and structural variations are detected. **(D)** Guided *de novo* genome assembly of longer reads. Longer reads are *de novo* assembled into contigs, which are aligned against the reference or a closely related genome to be extended.

not get reconstructed properly, and thus, may completely miss the diversity present in the newly assembled genome (Lischer and Shimizu, 2017). In contrast, the *de novo* genome assembly even for a diploid genome is classified as an "NP-hard" (non-deterministic polynomial-time) problem meaning it does not have an optimal, known solution. The genome assemblers must assemble a jigsaw puzzle of very small pieces. These pieces are the short reads (∼75–300 bp) and different assembly tools are used to resolve a best-fit assembly. However, given that it is a NP-hard problem, most assemblies are likely only an approximation of the true genome order.

The assemblers also face the challenge of the repetitive nature of plant genomes along with heterozygosity and haplotype ambiguity that frequently splits these regions into multiple contigs. A number of algorithms are used for this computation. Some of the most well-known are the overlap computation, the Greedy algorithm (Huson et al., 2002), the Eulerian path (Pevzner et al., 2001), and two classes of assembly algorithms: Overlap-Layout-Consensus (OLC) and de Brujin graph. The overlap computation within an assembly tool requires a great deal of computational time, which can be easily reduced by parallelizing the computations using multi-processor machines or servers (Pop, 2009). The complexity of the overlap computation is affected by the number of the input sequencing reads. Furthermore, the assemblers based on the Greedy algorithm give the simplest (Pop, 2009), most intuitive solution to the assembly problem, yet it is harder to prove the correctness of the algorithm even if the algorithm is correct (Pop, 2009).

The OLC, which can effectively assemble very short reads, has been one of the most successful assembly strategies. The Eulerian approach was proposed as an alternative to the OLC for the assembly of Sanger data; however, because of its sensitivity to sequencing errors it has not been extensively used (Pop, 2009). Overall, the short sequence reads need to be assembled into contigs, then the contigs need to be placed into bigger scaffolds, and finally chromosomes. Examples of tools that use the OLC algorithm in combination with other techniques is MASURCA that uses de Brujin graphs to construct mega-reads for a better assembly (Zimin et al., 2017) and BAUM that uses adaptive unique mapping to reconstruct repetitive regions (Wang et al., 2018).

De novo genome assembly is essential to capture the biological diversity within re-sequenced genomes. Yet, this task is near impossible without the use of mate-pairs, longer reads, or linked reads to provide information that can bridge these

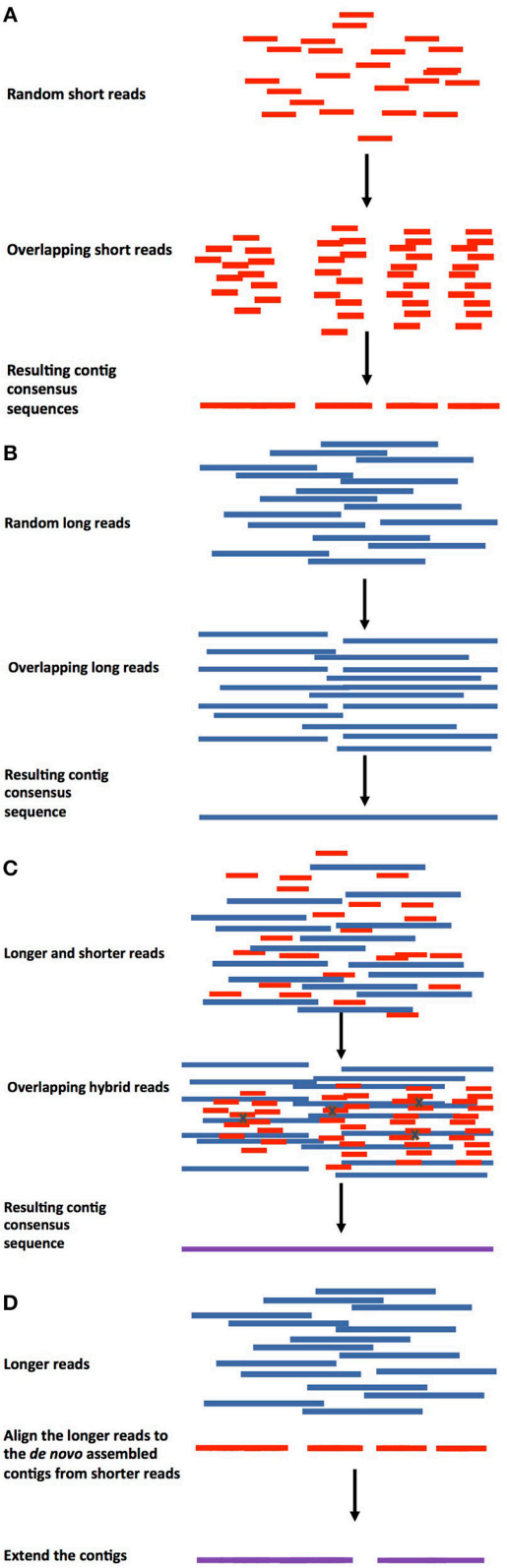

FIGURE 2 | Approaches for *de novo* assembly genome approaches. **(A)** Short read assembly. Genome assembly using only shorter reads and any assembly

(Continued)

FIGURE 2 | tool to construct contiguous sequences/contigs. **(B)** Longer reads assembly. Contig (red) assembly using longer reads (long, linked reads, optical maps) followed by scaffold assembly and gap filling. **(C)** Hybrid genome assembly. In this method, shorter reads can be assembled into contigs and the longer reads can be used for error correction (errors represented by Xs), then the corrected contigs can be assembled into scaffolds and the gaps filled. **(D)** Hybrid genome assembly using pre-assembled contigs. Longer reads are aligned against *de novo* pre-assembled contigs from shorter reads, followed by contig extension.

difficult repetitive regions. Currently, there is a lack of genome assembly and mapping algorithms specialized for polyploid genomes. These would need to be optimized for using more computational power (resources) to handle the challenge of the increased complexity and size of the data sets. Polyploid genome assemblies made from only short reads fail to capture haplotyping variation and present only a single consensus sequence of several chromosome sets. Better algorithms are necessary to minimize misassembly of paralogous and orthologous regions in polyploid plant genomes.

Sequencing errors, read length, quality values, number of reads, and coverage are important factors in assembling genomes and there is little difference in these factors/variables between diploid and polyploid plant genomes. However, because of the complex nature of polyploid genomes, there is not "a best fit" for the main assembling pipeline and not every approach is reproducible for other polyploid plant genomes. Different results can be obtained from the various algorithms used for alignment and assemblers and often genome assemblies are only an estimate of the true biological genome. It often takes a decade or longer to make improvements and corrections to the original draft release. For example, the human genome released in 2000 has gone through multiple revisions to correct errors. Furthermore, the metrics used to make comparisons tend to only focus on size which does not capture contig quality nor accuracy, and thus, there are no commonly accepted standardized methods for validation of the assemblers, which means most genomes are accepted as "draft" assemblies (Narzisi and Mishra, 2011). BUSCO (Simão et al., 2015) and QUAST (Gurevich et al., 2013) are two examples of tools that have been created in an attempt to validate the quality of an assembly.

HOW TO ESTIMATE PLOIDY LEVEL IN PLANTS

The ploidy level in plants is normally estimated by measuring the *C*-value (amount of DNA in the unreplicated gametic nucleus) using flow cytometry (Dart et al., 2004; Eaton et al., 2004; Grundt et al., 2005; Clarindo et al., 2008; Harbaugh, 2008; reviewed by Yang et al., 2011). For example, flow cytometry was used to estimate genome content and ploidy in over 300 accessions of the Magnoliaceae family (Parris et al., 2010), in six *Olea europea* (olive) subspecies (Besnard et al., 2007), and in *B. napus* leaf tissue samples (Cousin et al., 2009). Public databases exist to capture *C*-value and ploidy levels in plants (e.g., http://data.kew.org/cvalues/). Recent tools have also been developed to infer the ploidy level using NGS data, such as ploidyNGS (Dos Santos et al., 2017), ConPADE (Margarido and Heckerman, 2015), and a pipeline using single nucleotide polymorphism (SNP) counts that was reported earlier by Yoshida et al. (2013) for the estimation of ploidy level in the plant pathogen *Phytophthora infestans*. A general approach to estimate ploidy levels using NGS is by mapping the sequenced reads to the reference genome and then counting the number of mapped reads, representing the different alleles at each position. PloidyNGS (Dos Santos et al., 2017) was implemented by automating the process of observing the frequency of the alleles by generating a histogram. It was tested on diploid and haploid *Saccharomyces cerevisiae* datasets. ConPADE (Margarido and Heckerman, 2015) was specifically designed to estimate the ploidy levels of highly polyploid plant genomes and has been tested on wheat. A weakness is its sensitivity to the quality of the mapping step as this can bias the ploidy estimation (Dos Santos et al., 2017). Finally, the pipeline by Yoshida et al. (2013) is similar in the sense that the distribution of read counts at biallelic SNPs is observed, which allowed the identification of diploid, triploid, and tetraploid *P. infestans* strains. Another recent statistical tool for ploidy estimation is nQuire (Weiß et al., 2017), which uses NGS data to distinguish between diploids, triploids and tetraploids.

Ploidy estimation tools have been reported such as EAGLE (Loh et al., 2016) and ReadSim (Schmid et al., 2006). More recent tools for the haploid assembly consist of HapCompass (Aguiar and Istrail, 2012), HaploSim (Bastiaansen et al., 2012), HapCut (Bansal and Bafna, 2008), and HapCUT2 (Edge et al., 2017). Real and simulated data were analyzed with HapCUT2 (Edge et al., 2017) and it was shown that it is more accurate and can use not only WGS, but also SMRT (www.pacb.com//smrt-science) and Hi-C data (Lieberman-Aiden et al., 2009) for haplotype assembly. SWEEP (Clevenger and Ozias-Akins, 2015) is a tool designed to filter SNPs detected in re-sequenced autopolyploid and allopolyploid crops using NGS approaches. The detected SNPs can be further used for the haplotype construction. Another NGS tool is HANDS (Mithani et al., 2013), which also can be used for auto- and allopolyploids and by aligning the sequenced reads to the reference genome(s) it can detect the subgenomes in polyploids. Longranger software by 10X Genomics can be used for phasing. It can determine which barcodes are associated with each heterozygous locus and while phasing, it can construct the organism's haplotypes. Simply, it aligns the raw reads to the sequence of both alleles to determine which allele each read represents.

HOW TO "RESOLVE" THE PLOIDY ISSUE (HOW TO REDUCE THE COMPLEXITY OF THE PROBLEM)

Genome-Related Approach

Several strategies have been adopted for the sequencing and assembly of large polyploid genomes of crop plants (Bevan et al., 2017). One approach involves the reduction of genome complexity using a natural or *in vitro* generated haploid. An

example is the sequencing of the potato genome by the Potato Genome Sequencing Consortium (2011). This genome was produced from a doubled monoploid that was homozygous for a single set of 12 chromosomes to generate a reference (The Potato Genome Sequencing Consortium, 2011). A similar approach was used for the genome assembly of the hexaploid bread wheat, *T. aestivum*. Aneuploid bread wheat lines derived from double ditelosomic stocks of a hexaploid wheat cultivar were used to sequence each individual chromosome arm (except 3B) using Illumina short-reads technology (International Wheat Genome Sequencing Consortium (IWGSC), 2014). The chromosomes were assembled *de novo,* which reduced the complexity of assembling this highly redundant genome, aiding the differentiation of genes present in multiple copies and of highly conserved homologs.

A second approach involves sequencing a diploid progenitor species to aid in the assembly of the cultivated form. Care must be taken to choose the diploid progenitors most similar to the cultivated form. The diploid genomes of progenitor species can be used to determine the origin and structure of contigs when assembling large polyploid genomes. For example, strawberry (*Fragaria* × *ananassa*) is an octoploid (2n = 8x = 56) whose origin remains controversial. One theory suggests that it was formed from a natural hybridization between two octoploids- *F. virginiana* and *F. chiloensis* (Darrow, 1966). According to Davis et al. (2007), *F. vesca*, *F. nubicola*, and *F. orientalis* are possible progenitors. To access the genetic diversity of this valuable crop, one diploid variety of *F. vesca* (2n = 2x = 14) (*F. vesca* spp. *vesca* accession Hawaii 4) was sequenced (Shulaev et al., 2011).

Oilseed rape or canola (*B. napus)* is an allopolyploid derived from two diploid species of *Brassica* that are triplicated versions of an ancestral diploid. Genome assemblies of *B. napus* were assigned to these two subgenomes using sequence assemblies from each diploid progenitor, but many sequence scaffolds showed ambiguous assignment to homeologous groups, owing to homeolog exchange and frequent gene loss (Chalhoub et al., 2014). A similar strategy was used to characterize the allotetraploid genome of peanut (*Arachis hypogaea)*, which formed from two diploid species *A. duranensis* (A genome) and *A. ipaënsis* (B genome). Essentially complete assemblies of the genomes of the progenitor species *A. duranensis* and *A. ipaënsis* were generated and shown to directly align with the genetic map of a cultivated tetraploid peanut (Bertioli et al., 2015). In the same study, synthetic long-read sequencing of the tetraploid peanut genome showed that it was 98–99% identical to the diploid genomes, with differences due to recombination of polyploid genomes involved from the sequencing of DNA from purified chromosome arms (Bertioli et al., 2015). Some of the challenges in assembling the cultivated peanut genome have been the high similarity between the two-progenitor species, a high number of transposable elements, and recent evidence of tetrasomic recombination in this allotetraploid (Bertioli et al., 2015). Lastly, upland cotton (*G. hirsutum*) is an allotetraploid that formed 1–2 Myr (million years) ago from two unknown diploid progenitor species. The genome complexity of upland cotton was reduced by sequencing highly homozygous allohaploid lines to a coverage depth of 245x with Illumina short-read sequencing reads (Li et al., 2015). A dense genetic map was used to align and correct scaffolds, which covered 96% of the estimated 2.5 Gb genome, and fluorescence *in situ* hybridization (FISH) was used to confirm a successful allotetraploid assembly.

Genome Sequencing and Algorithmic (Pipeline) Approach

There are several examples of successful *de novo* sequencing and assembly of large allopolyploid genomes of crops that use long-range alignments of sequence scaffolds to generate extended haplotypes to form distinctive homeologous pseudomolecules. Tobacco (*Nicotiana tabacum;* 2n = 4x = 48) is an allotetraploid that is derived from the diploid genomes of *N. sylvestris* and *N. tomentosiformis*. Whole-genome shotgun assemblies were aligned to physical maps to create longer super scaffolds that could be assigned directly to the progenitor genomes (Sierro et al., 2013). The polyploid genome of Indian mustard (*B. juncea*) (Yang et al., 2016) has been assembled using a combination of Illumina short reads, PacBio single molecule, real-time long sequence reads and optical maps from BioNano Genomics. The short and long reads were aligned to the maps, which directly helped in the determination of the individual molecules of tagged DNA, and dense genetic maps. The genome was almost fully represented in the assembly, which was assigned to the A genome [402 Megabase (Mb)] and the B genome (547 Mb).

Furthermore, an alternative approach to resolve polyploid complexity is by haplotyping. The process of assigning variants to a particular chromosome or defining which alleles appear together (corresponding haplotypes), is called phasing and haplotyping, respectively (Huang et al., 2017). Haplotypes can provide more information than un-phased genotypes in diverse fields, such as identifying genotype-phenotype associations and exploring genetic resistance to plant diseases. An example of this approach is the recent assembly of the hexaploid genome of sweetpotato (*Ipomoea batatas)*. The authors describe haplotype construction by applying a novel approach (Yang et al., 2017) where paired reads and mate pairs were initiatlly used for *de novo* assembly, then haplotypes were phased. Overlapping haplotypes were merged into larger haplotypes, mapping all the raw reads against the phased haplotypes. Finally, scaffolds were constructed based on the haplotypes and a consensus sequence was generated (Yang et al., 2017). This method, called "Ranbow," can be downloaded at https://www.molgen.mpg.de/ranbow. A number of algorithms/tools to resolve the haplotype of polyploid genomes exist. Some examples are HANDS (Mithani et al., 2013), SDhaP (Das and Vikalo, 2015), and HapTree (Berger et al., 2014). Haplotype construction depends on the read depth or coverage as it is necessary to have a high coverage for each homolog (5–20x per homolog), as well as an insert size of 600–800 bp (Motazedi et al., 2017). It is also important to know the nature of the plant genome and ploidy before performing haplotyping in order to select the most appropriate tool. If available, it may be better to combine various individuals or parental information for haplotyping analysis (Motazedi et al., 2017). From an algorithmic point of view, haplotyping requires a lot of memory and computation time.

Another solution is the construction of a pan-genome, which shows the variation and commonality between individuals. A pan-genome includes "completeness" as it contains the core genome shared by all the individuals sequenced, but also the genes that are absent/present in some of the re-sequenced genomes. Generally, it is a very helpful approach for breeding applications as it anchors all the known variations and phenotype information and can include wild relatives of the cultivated crop lines. It also aids in the identification of novel genes from the available germplasm that are not found in the reference genome (The Computational Pan-Genomics Consortium, 2016). Additionally, it represents the polyploid genomes and in the case of the allopolyploids, it allows the quantification of allele dosage between germplasm samples (The Computational Pan-Genomics Consortium, 2016). Pan-genome construction is even more computationally challenging in the case of polyploid plant genomes as the corresponding genotype needs to be determined by variant calling and identifying novel variants for all the haploids. Previously, a pan-genome was constructed from 18 wheat cultivars and it was shown that a large number of variable genes affected by presence/absence and variation between the genes could be associated with important agronomic traits (Montenegro et al., 2017). NRGene's (www.nrgene.com) PanMAGIC platform can be used for pangenome analysis and was applied to analyze six maize genomes (Lu et al., 2015).

THIRD GENERATION GENOMIC TECHNOLOGIES COME TO THE RESCUE

Genome assembly and scaffolding can be performed using shorter reads (Illumina data), or longer reads from either PacBio (www.pacb.com) or Oxford Nanapore (https://nanoporetech.com/), or a combination of both short and long reads. Another alternative is the assembly of linked reads from 10X genomics. Additionally, for higher contiguity, longer-range scaffolders from Dovetail (dovetailgenomics.com) and BioNano Genomics (bionanogenomics.com) can be used for the construction of physical maps using very large DNA fragments. A hybrid scaffolding approach can also be applied where longer reads are used to improve assemblies generated using short-reads or even combined wiht longer-range scaffolding data.

Even though the hexaploid wheat genome was assembled from only short reads, it is very challenging to assemble such a large and highly repetitive genome using this approach. A less complicated assembly strategy is to use long-reads to aid in the assembly of difficult portions of the genome. The most widely used long-read sequencing technology is Pacific Biosciences' Single Molecule Real-Time (SMRT) sequencing. Recently, a few polyploid plant genomes were assembled using PacBio long reads including three allotetraploid plant genomes *C. quinoa* (quinoa) (Jarvis et al., 2017), *Eleusine coracana* (finger millet) (Hatakeyama et al., 2017) and *Coffea arabica* (Arabica coffee) (Cheng et al., 2017).

As mentioned earlier, another solution to the read length issue is the ultra-long and real-time data sequencing approach by Oxford Nanopore Technologies (www.nanoporetech.com). Currently three plant genomes have been sequenced with Nanopore, a wild tomato genome *Solanum pennellii* (Schmidt et al., 2017), the genome of *A. thaliana* (Mondal et al., 2017), and most recently the genome of *Oryza coarctata* (Michael et al., 2018). Illumina's SLR technology on the other hand, has already been applied for the estimation of the haploid draft genome of the polyploid sugarcane hybrid SP80-3280 (Riaño-Pachón and Mattiello, 2017).

The long-reads can also be combined with existing short-reads for genome assembly, called hybrid genome assembly. The resulting genome assembly from short-reads needs improvement in its contiguity because the contigs need to be assembled into scaffolds. Initially, the contigs are ordered using alignments from paired-end reads, read pairs from (Bacterial Artificial Chromosome) BAC or fosmid ends, which are powerful ways to increase the contiguity and help bridge the repeats—the main reason generally for breaks in the genome assemblies. In addition, genetic and physical maps are also essential for polyploid plant genome assembly (i.e., a physical map was used in the case of the tetraploid cotton genome). Optical mapping enables the fingerprinting of large genome fragments and can be used to improve highly fragmented genome assemblies. This technology promises the improvement of scaffolding and eventually lessens the need for genetic and physical mapping (Jiao and Schneeberger, 2017).

Another new promising technology that can potentially be applied to complex, polyploid plant genomes is the 10X genomics approach. There is only one scientific report on plant research using this technology to date on a diploid pepper genome (*Capsicum annuum*) (Hulse-Kemp et al., 2018). The haplotype construction was generated to karyotype aneuploidy in a cancer study (Bell et al., 2017) and it was also used in the generation of a protocol for haplotyping human genome (Porubsky et al., 2017), making it a promising technique for polyploidy genome data. Additional techniques used by polyploid plant projects include Hi-C and chromosome-scale assembly. For example, a study is underway to detect large chromosomal rearrangements in wheat genomes (Monat et al., 2018) and another project uses chromosome scale scaffolding on the allotetraploid coffee genome (Zimin et al., 2018).

ADVANCES IN GENOMIC RESOURCES AND FUNCTIONAL TOOLS IN MOLECULAR GENETICS AND BREEDING

The advance of NGS technologies has immensely impacted the field of plant genomics in model and non-model crops alike, and it is continuously contributing to bridging the gap between genotype and phenotype. The genotype can be linked to the phenotype by Genome Wide Association studies (GWAS) and the advent of NGS has revolutionized genomic, as well as, transcriptomic (RNA-Sequencing) approaches to biology including plant genomics in model and non-model crops. Modern breeding programs combine various approaches for more efficient breeding, in parallel with the reduction of the whole breeding period (Varshney et al., 2013). These approaches

include the traditional phenotype-based selection, marker-assisted selection, and genome-assisted breeding (Varshney et al., 2013). The continuous effort in improving major crops has resulted in great genetic and genomic resources for crop traits. Some instances of databases that host these resources can be found in **Table 3**.

LACK OF COMPLEXITY OF THE CURRENTLY AVAILABLE REFERENCE GENOMES OF POLYPLOID CROPS

High quality reference genomes, gene discovery, and comparative genomics depend on the construction of a high quality *de novo* genome assembly. These assemblies are more feasible, but still not perfect using haploid and inbred species. Despite their importance to reflect the genetic information within an organism, most of the currently available polyploid and diploid plant genome assemblies do not capture the heterozygosity present. The majority of the currently available reference genomes, especially those of the polyploids, lack variation and characteristics of other individuals that are not captured or presented. This happens because the simpler genomes are sequenced first, but also due to the sequencing of diploid and less heterozygous progenitor species for the reduction of the intricacy of the polyploid assembly problem. In reality, the assembled genome is a flat DNA sequence, which shows neither the variation between homologous chromosomes,

TABLE 3 | Host-databases of various plant genetic and genomic resources.

DB name	Resources	Plants	URL
Genbank	Genomic	Various plant species	https://www.ncbi.nlm.nih.gov/genbank/
EMBL	Genomic	Various plant species	https://www.ebi.ac.uk/
DDBJ	Genomic	Various plant species	http://www.ddbj.nig.ac.jp/
UniProt	Protein and functional	Various plant species	http://www.uniprot.org/
NCBI	Genomic	Various plant species	https://www.ncbi.nlm.nih.gov/
GOLD	Genomic, metagenomics, transcriptomic	Various plant species	https://gold.jgi.doe.gov/cgi-bin/GOLD/bin/gold.cgi
Phytozome	Genomic	92 assembled and annotated plant species	https://phytozome.jgi.doe.gov/pz/portal.html
Plantgdb	Genomic, transcriptomic	27 assembled and annotated plant species	http://www.plantgdb.org/
Sol	Genomic	11 *Solanaceae* species	https://solgenomics.net/
Gramene	Genomic, genetic markers, QTLs	53 plant species	http://www.gramene.org/
MaizeGCB	Genomic, annotations, tool host	*Zea mays*	https://www.maizegdb.org/
Tair	Genetic and molecular biology data	*Arabidopsis thaliana*	https://www.arabidopsis.org/
CottonGEN	Genomic, Genetic and breeding resources	49 *Gossypium* species	https://www.arabidopsis.org/
PLEXdb	Gene expression	14 plant species	http://www.plexdb.org/
RicePro	Gene expression	*Oryza sativa*	http://ricexpro.dna.affrc.go.jp/
CerealsDB	Genetic markers	*Triticum aestivum*	http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/indexNEW.php
PeanutBase	Genome, MAS, QTLs, Germplasm	*Arachis hypogaea*	https://peanutbase.org/
SoyKb	Genetic markers, genomic resources	*Glycine max*	http://soykb.org/
SoyBase	Genetic markers, QTLs, genomic resources	*G. max*	https://soybase.org/
PGDBj	Genetic markers, QTLs, genomic resources	80 plant species	http://pgdbj.jp/
SNP-Seek	Genotype, Phenotype and Variety information	*O. sativa*	http://snp-seek.irri.org/
GrainGenes	Genome, Genetic markers, QTLs, genomic resources	*T. aestivum, Hordeum vulgare, Secale cereale, Avena sativa* etc	https://wheat.pw.usda.gov/GG3/
ASRP	small RNA	*A. thaliana*	http://asrp.danforthcenter.org/
CSRDB	small RNA	*Z. mays*	http://sundarlab.ucdavis.edu/smrnas/
BrassicaInfo	Genomic	7 *Brassica* species	http://brassica.info/
BRAD	Genomics, Genetic Markers and Maps	*Brassica*	http://brassicadb.org/brad/
Ensembl Plants	Genomic	45 plant species	http://plants.ensembl.org/index.html
Ipomoea Genome Hub	Genomic, EST	*Ipomoea batatas*	https://ipomoea-genome.org/
PGSC	Genomic, annotation	*S. tuberosum, S.chacoense*	http://solanaceae.plantbiology.msu.edu/pgsc_download.shtml
GDR	Genomics, Genetics, breeding	*Rosaceae*	https://www.rosaceae.org/analysis/266
HWG	Genomics, Transcriptomics, Genetic Markers	Forest trees and woody plants	https://www.hardwoodgenomics.org/

nor allelic variations, or structural variations. The resulting "model" reference genome is more distant than the majority of the other individuals in a species. Furthermore, genes may be missing or not annotated. A solution to this problem is the construction of pan-genomes (as described above), which show the core and the variable regions of a genome between individuals. An example of a pan-genome application is in the hexaploid bread wheat (Montenegro et al., 2017).

Even in the case of the smaller, "simpler" bacterial genomes, the submitted genomes are not complete. Despite the exponential generation of NGS data, the majority of the submitted genomes represent only draft or in scaffold format, incomplete genomes. The higher ploidy levels of the polyploid plant genomes make the situation even more difficult to handle. This leads to highly fragmented genome assemblies, with disconnected contigs of repetitive sequences. As discussed, better tools are needed that allow automatic contig assembly of (plant) genomes with many repeats and that are sensitive to ploidy levels and can handle haplotype construction. Also, to date allopolyploid plant genomes cannot be represented in an integrated assembly, rather the sub-genomes are found in separate assemblies.

CONCLUSIONS

Improving genome sequencing and assembly of polyploid plant crops will have a fundamental impact on genetic research and on plant breeding by better understanding the genomes, identifying genomic variants and relating them to economic, physiological, and morphological agronomic traits, such as higher yield, abiotic/biotic tolerance, root structure etc. Better polyploid plant genome assemblies will also aid in the study of the genotype-phenotype-environment relationship. For this, more plant polyploid-oriented algorithmic and technological (sequencing) advances are necessary. High quality reference sub-genomes in polyploid crops in addition to multiple reference genomes or a pan-genome per crop species are necessary to capture variation and to better understand these economically important genomes.

AUTHOR CONTRIBUTIONS

MK: drafted the manuscript, compiled the tables, and made the figure. MK, NA, DE, HT, and MS: designed the outline, content, and edited the manuscript.

FUNDING

The authors acknowledge funding through a Nouvelles Initiatives (Project International) grant from the Centre SÈVE (Fonds de recherche du Québec - Nature et technologies (FRQ-NT) to MS, NA, DE, and HT; the Natural Sciences and Engineering Research Council of Canada (NSERC) (Grant No. 283303) to MS; A-base funding from Agriculture and Agri-Food Canada to HT; and the McGill Department of Plant Science Graduate Excellence Fund. The authors also gratefully acknowledge the support of the CGIAR Genebank Platform.

REFERENCES

Aguiar, D., and Istrail, S. (2012). HapCompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data. *J. Comput. Biol.* 19, 577–590. doi: 10.1089/cmb.2012.0084

Aguiar, D., and Istrail, S. (2013). Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics* 29, i352–i360. doi: 10.1093/bioinformatics/btt213

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815. doi: 10.1038/35048692

Bansal, V., and Bafna, V. (2008). HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* 24, i153–i159. doi: 10.1093/bioinformatics/btn298

Bastiaansen, J. W., Coster, A., Calus, M. P., van Arendonk, J. A., and Bovenhuis, H. (2012). Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. *Genet. Select. Evol.* 44:3. doi: 10.1186/1297-9686-44-3

Bell, J. M., Lau, B. T., Greer, S. U., Wood-Bouwens, C., Xia, L. C., Connolly, I. D., et al. (2017). Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. *Nucleic Acids Res.* 45, e162–e162. doi: 10.1093/nar/gkx712

Bento, M., Gustafson, J. P., Viegas, W., and Silva, M. (2011). Size matters in triticeae polyploids: larger genomes have higher remodeling. *Genome* 54, 175–183. doi: 10.1139/G10-107

Berger, E., Yorukoglu, D., Peng, J., and Berger, B. (2014). Haptree: sa novel bayesian framework for single individual polyplotyping using ngs data. *PLoS Comput. Biol.* 10:e1003502. doi: 10.1371/journal.pcbi.1003502

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K., et al. (2015). The genome sequences of arachis duranensis and arachis ipaensis, the diploid ancestors of cultivated peanut. *Nat. Genet.* 47, 438–446. doi: 10.1038/ng.3517

Besnard, G., Garcia-Verdugo, C., Rubio de Casas, R., Treier, U. A., Galland, N., and Vargas, P. (2007). Polyploidy in the olive complex (*Olea europaea*): evidence from flow cytometry and nuclear microsatellite analyses. *Ann. Bot.* 101, 25–30. doi: 10.1093/aob/mcm275

Bevan, M. W., Uauy, C., Wulff, B. B., Zhou, J., Krasileva, K., and Clark, M. D. (2017). Genomic innovation for crop improvement. *Nature* 543, 346–354. doi: 10.1038/nature22011

Beyaz, R., Alizadeh, B., Gürel, S., Fatih Özcan, S., and Yildiz, M. (2013). Sugar beet (*Beta vulgaris* L.) growth at different ploidy levels. *Caryologia* 66, 90–95. doi: 10.1080/00087114.2013.787216

Butts, C. T., Bierma, J. C., and Martin, R. W. (2016). Novel proteases from the genome of the carnivorous plant drosera capensis: structural prediction and comparative analysis. *Proteins* 84, 1517–1533. doi: 10.1002/prot.25095

Cannarozzi, G., Plaza-Wüthrich, S., Esfeld, K., Larti, S., Wilson, Y. S., Girma, D., et al. (2014). Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (eragrostis tef). *BMC Genomics* 15:581. doi: 10.1186/1471-2164-15-581

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X., et al. (2014). Plant genetics. Early allopolyploid evolution in the post-neolithic *Brassica* napus oilseed genome. *Science* 345, 950–953. doi: 10.1126/science.1253435

Chen, A., and Dubcovsky, J. (2012). Wheat TILLING mutants show that the vernalization gene VRN1 down-regulates the flowering repressor VRN2 in leaves but is not essential for flowering. *PLoS Genet.* 8:e1003134. doi: 10.1371/journal.pgen.1003134

Chen, Z. J. (2010). Molecular mechanisms of polyploidy and hybrid vigor. *Trends Plant Sci.* 15, 57–71. doi: 10.1016/j.tplants.2009.12.003

Cheng, B., Furtado, A., and Henry, R. J. (2017). Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *Gigascience* 6, 1–13. doi: 10.1093/gigascience/gix086

Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., et al. (2010). Megabase level sequencing reveals contrasted organization and evolution

patterns of the wheat gene and transposable element spaces. *Plant Cell* 22, 1686–1701. doi: 10.1105/tpc.110.074187

Clarindo, W. R., de Carvalho, C. R., Araújo, F. S., de Abreu, I. S., and Otoni, W. C. (2008). Recovering polyploid papaya *in vitro* regenerants as screened by flow cytometry. *Plant Cell Tissue Organ Cult.* 92, 207–214. doi: 10.1007/s11240-007-9325-1

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., and Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology* 1, 439–459. doi: 10.3390/biology1020439

Clevenger, J. P., and Ozias-Akins, P. (2015). SWEEP: a tool for filtering high-quality SNPs in polyploid crops. *G3 (Bethesda, Md.)* 5, 1797–1803. doi: 10.1534/g3.115.019703

Computational Pan-Genomics Consortium (2016). Computational pan-genomics: Status, promises and challenges. *Brief. Bioinform.* 19, 118–135. doi: 10.1093/bib/bbw089

Costa, M. D., Artur, M. A., Maia, J., Jonkheer, E., Derks, M. F., Nijveen, H., et al. (2017). A footprint of desiccation tolerance in the genome of xerophyta viscosa. *Nat. Plants* 3:17038. doi: 10.1038/nplants.2017.38

Cousin, A., Heel, K., Cowling, W., and Nelson, M. (2009). An efficient high-throughput flow cytometric method for estimating DNA ploidy level in plants. *Cytometry Part A* 75, 1015–1019. doi: 10.1002/cyto.a.20816

Crow, K. D., and Wagner, G. P. (2005). What is the role of genome duplication in the evolution of complexity and diversity? *Mol. Biol. Evol.* 23, 887–892. doi: 10.1093/molbev/msj083

Darrow, G. M. (1966). *The Strawberry: History, Breeding and Physiology*. Holt, Rinehart and Winston.

Dart, S., Kron, P., and Mable, B. K. (2004). Characterizing polyploidy in arabidopsis lyrata using chromosome counts and flow cytometry. *Can. J. Botany* 82, 185–197. doi: 10.1139/b03-134

Das, S., and Vikalo, H. (2015). SDhaP: Haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics* 16:260. doi: 10.1186/s12864-015-1408-5

Davis, T. M., Denoyes-Rothan, B., and Lerceteau-Köhler, E. (2007). "Strawberry," in *Genome Mapping and Molecular Breeding in Plants IV: Fruits and Nuts*, ed C. Kole (Berlin: Springer). 189–206.

D'Hont, A. (2005). Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenet. Genome Res.* 109, 27–33. doi: 10.1159/000082378

Dohm, J. C., Minoche, A. E., Holtgräwe, D., Capella-Gutiérrez, S., Zakrzewski, F., Tafer, H., et al. (2014). The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* 505, 546–549. doi: 10.1038/nature 12817

Dos Santos, R. A., Goldman, G. H., and Riaño-Pachón, D. M. (2017). ploidyNGS: Visually exploring ploidy with next generation sequencing data. *Bioinformatics* 33, 2575–2576 doi: 10.1093/bioinformatics/btx204

Doyle, J. J., Flagel, L. E., Paterson, A. H., Rapp, R. A., Soltis, D. E., Soltis, P. S., et al. (2008). Evolutionary genetics of genome merger and doubling in plants. *Annu. Rev. Genet.* 42, 443–461. doi: 10.1146/annurev.genet.42.110807.091524

Eaton, T., Curley, J., Williamson, R., and Jung, G. (2004). Determination of the level of variation in polyploidy among kentucky bluegrass cultivars by means of flow cytometry. *Crop Sci.* 44, 2168–2174. doi: 10.2135/cropsci2004.2168

Edge, P., Bafna, V., and Bansal, V. (2017). HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* 27, 801–812. doi: 10.1101/gr.213462.116

Glover, N. M., Redestig, H., and Dessimoz, C. (2016). Homoeologs: what are they and how do we infer them? *Trends Plant Sci.* 21, 609–621. doi: 10.1016/j.tplants.2016.02.005

Grundt, H. H., Obermayer, R., and Borgen, L. (2005). Ploidal levels in the arctic-alpine polyploid draba lactea (*Brassicaceae*) and its low-ploid relatives. *Botan. J. Linn. Soc.* 147, 333–347. doi: 10.1111/j.1095-8339.2005.00377.x

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086

Harbaugh, D. T. (2008). Polyploid and hybrid origins of pacific island sandalwoods (*Santalum*, *Santalaceae*) inferred from low-copy nuclear and flow cytometry data. *Int. J. Plant Sci.* 169, 677–685. doi: 10.1086/533610

Hatakeyama, M., Aluri, S., Balachadran, M. T., Sivarajan, S. R., Patrignani, A., Grüter, S., et al. (2017). Multiple hybrid de novo genome assembly of finger millet, an orphan allotetraploid crop. *DNA Res.* 25, 39–47. doi: 10.1093/dnares/dsx036

Hirakawa, H., Shirasawa, K., Kosugi, S., Tashiro, K., Nakayama, S., Yamada, M., et al. (2014). Dissection of the octoploid strawberry genome by deep sequencing of the genomes of fragaria species. *DNA Res.* 21, 169–181. doi: 10.1093/dnares/dst049

Hittalmani, S., Mahesh, H., Shirke, M. D., Biradar, H., Uday, G., Aruna, Y., et al. (2017). Genome and transcriptome sequence of finger millet (*Eleusine coracana* (L.) Gaertn.) provides insights into drought tolerance and nutraceutical properties. *BMC Genomics* 18:465. doi: 10.1186/s12864-017-3850-z

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J., Clark, R. M., et al. (2011). The arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43, 476–481. doi: 10.1038/ng.807

Huang, M., Tu, J., and Lu, Z. (2017). Recent advances in experimental whole genome haplotyping methods. *Int. J. Mol. Sci.* 18, 1944. doi: 10.3390/ijms18091944

Huang, S., Ding, J., Deng, D., Tang, W., Sun, H., Liu, D., et al. (2013). Draft genome of the kiwifruit actinidia chinensis. *Nat. Commun.* 4:2640. doi: 10.1038/ncomms3640

Hufton, A. L., and Panopoulou, G. (2009). Polyploidy and genome restructuring: a variety of outcomes. *Curr. Opin. Genet. Dev.* 19, 600–606. doi: 10.1016/j.gde.2009.10.005

Hulse-Kemp, A. M., Maheshwari, S., Stoffel, K., Hill, T. A., Jaffe, D., Williams, S. R., et al. (2018). Reference quality assembly of the 3.5-gb genome of capsicum annuum from a single linked-read library. *Horticult. Res.* 5:4. doi: 10.1038/s41438-017-0011-0

Huson, D. H., Reinert, K., and Myers, E. W. (2002). The greedy path-merging algorithm for contig scaffolding. *J. Alter. Complement. Med.* 49, 603–615. doi: 10.1145/585265.585267

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* 436, 793–800. doi: 10.1038/nature03895

International Wheat Genome Sequencing Consortium (IWGSC) (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345:1251788. doi: 10.1126/science.12 51788

International Wheat Genome Sequencing Consortium (IWGSC) (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361:eaar7191. doi: 10.1126/science.aar7191

Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmöckel, S. M., Li, B., Borm, T. J., et al. (2017). The genome of *Chenopodium quinoa*. *Nature* 542:307–312. doi: 10.1038/nature21370

Jiao, W., and Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.* 36, 64–70. doi: 10.1016/j.pbi.2017.02.002

Kagale, S., Koh, C., Nixon, J., Bollina, V., Clarke, W. E., Tuteja, R., et al. (2014). The emerging biofuel crop camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nat. Commun.* 5:3706. doi: 10.1038/ncomms4706

Kausar, N., Yousaf, Z., Younas, A., Ahmed, H. S., Rasheed, M., Arif, A., et al. (2015). Karyological analysis of bitter gourd (*Momordica charantia* L., *Cucurbitaceae*) from southeast asian countries. *Plant Genet. Resourc.* 13, 180–182. doi: 10.1017/S147926211400077X

Kim, Y.-M., Kim, S., koo, N., Shin, A.-Y., Yeom, S.-I., Seo, E., et al. (2017). Genome analysis of *Hibiscus syriacus* provides insights of polyploidization and indeterminate flowering in woody plants. *DNA Res.* 24, 71–80. doi: 10.1093/dnares/dsw049

Kronenberg, Z. N., Hall, R. J., Hiendleder, S., Smith, T. P., Sullivan, S. T., Williams, J. L., et al. (2018). FALCON-phase: Integrating PacBio and hi-C data for phased diploid genomes. *Biorxiv [Preprint]*. doi: 10.1101/327064

Lan, T., Renner, T., Ibarra-Laclette, E., Farr, K. M., Chang, T. H., Cervantes-Perez, S. A., et al. (2017). Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. *Proc. Natl. Acad. Sci. U S A.* 114, E4435–E4441. doi: 10.1073/pnas.1702072114

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

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. doi: 10.1126/science.1181369

Lischer, H. E. L., and Shimizu, K. K. (2017). Reference-guided *de novo* assembly approach improves genome reconstruction for related species. *BMC Bioinform.* 18:474. doi: 10.1186/s12859-017-1911-6

Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., et al. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* 48, 1443–1448. doi: 10.1038/ng.3679

Lu, F., Romay, M. C., Glaubitz, J. C., Bradbury, P. J., Elshire, R. J., Wang, T., et al. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun.* 6:6914. doi: 10.1038/ncomms7914

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402. doi: 10.1146/annurev.genom.9.081307.164359

Margarido, G. R., and Heckerman, D. (2015). ConPADE: genome assembly ploidy estimation from next-generation sequencing data. *PLoS Comput. Biol.* 11:e1004229. doi: 10.1371/journal.pcbi.1004229

Maxam, A. M., and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U S A.* 74, 560–564. doi: 10.1073/pnas.74.2.560

Meyers, L. A., and Levin, D. A. (2006). On the abundance of polyploids in flowering plants. *Evolution* 60, 1198–1206. doi: 10.1111/j.0014-3820.2006.tb01198.x

Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., et al. (2018). High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat. Commun.* 9:541. doi: 10.1038/s41467-018-03016-2

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

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J. H., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* linnaeus). *Nature* 452, 991–996. doi: 10.1038/nature06856

Mithani, A., Belfield, E. J., Brown, C., Jiang, C., Leach, L. J., and Harberd, N. P. (2013). HANDS: a tool for genome-wide discovery of subgenome-specific base-identity in polyploids. *BMC Genomics* 14:653. doi: 10.1186/1471-2164-14-653

Monat, C., Padmarasu, S., Himmelbach, A., Baruch, K., Kolodziej, M. C., Wicker, T., et al. (2018) "W1033: Hi-C and chromosome-scale assembly to detect large chromosomal rearrangements in wheat genomes," in *26th PAG Conference* (San Diego, CA).

Mondal, T. K., Rawal, H. C., Gaikwad, K., Sharma, T. R., and Singh, N. K. (2017). First *de novo* draft genome sequence of oryza coarctata, the only halophytic species in the genus *Oryza*. *F1000Res* 6:1750. doi: 10.12688/f1000research.12414.2

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K., et al. (2017). The pangenome of hexaploid bread wheat. *Plant J.* 90, 1007–1013. doi: 10.1111/tpj.13515

Motazedi, E., Finkers, R., Maliepaard, C., and de Ridder, D. (2017). Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: a simulation study. *Brief. Bioinform.* 19, 387–403. doi: 10.1093/bib/bbw126

Narzisi, G., and Mishra, B. (2011). Comparing *de novo* genome assembly: the long and short of it. *PLoS ONE* 6:e19175. doi: 10.1371/journal.pone.0019175

Parris, J. K., Ranney, T. G., Knap, H. T., and Baird, W. V. (2010). Ploidy levels, relative genome sizes, and base pair composition in magnolia. *J. Am. Soc. Horticult. Sci.* 135, 533–547.

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U S A.* 98, 9748–9753. doi: 10.1073/pnas.171285098

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Brief. Bioinform.* 10, 354–366. doi: 10.1093/bib/bbp026

Porubsky, D., Garg, S., Sanders, A. D., Korbel, J. O., Guryev, V., Lansdorp, P. M., et al. (2017). Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat. Commun.* 8:1293. doi: 10.1038/s41467-017-01389-4

Potato Genome Sequencing Consortium (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189–195. doi: 10.1038/nature10158

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* 29, 467–501. doi: 10.1146/annurev.ecolsys.29.1.467

Riaño-Pachón, D. M., and Mattiello, L. (2017). Draft genome sequencing of the sugarcane hybrid *SP*80–3280. *F1000Res* 6:861. doi: 10.12688/f1000research.11859.2

Rothfels, K., and Heimburger, M. (1968). Chromosome size and DNA values in sundews (*Droseraceae*). *Chromosoma* 25, 96–103. doi: 10.1007/BF00338236

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., et al. (1977a). Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* 265, 687–695. doi: 10.1038/265687a0

Sanger, F., and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94, 441–448. doi: 10.1016/0022-2836(75)90213-2

Sanger, F., Nicklen, S., and Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U S A.* 74, 5463–5467. doi: 10.1073/pnas.74.12.5463

Sato, S., Hirakawa, H., Isobe, S., Fukai, E., Watanabe, A., Kato, M., et al. (2010). Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. *DNA Res.* 18, 65–76. doi: 10.1093/dnares/dsq030

Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Hum. Mol. Genet.* 19, R227–R240. doi: 10.1093/hmg/ddq416

Schmid, R., Schuster, S., Steel, M., and Huson, D. (2006). *Readsim-a Simulator for Sanger and 454 Sequencing*. University of Tübingen.

Schmidt, M. H., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., et al. (2017). *De novo* assembly of a new *Solanum pennellii* accession using nanopore sequencing. *Plant Cell* 29, 2336–2348. doi: 10.1105/tpc.17.00521

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* 19, 329–346 doi: 10.1038/s41576-018-0003-4

Shen, Q., Zhang, L., Liao, Z., Wang, S., Yan, T., Shi, P., et al. (2018). The genome of *Artemisia annua* provides insight into the evolution of *Asteraceae* family and artemisinin biosynthesis. *Mol. Plant* 11, 776–788 doi: 10.1016/j.molp.2018.03.015

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 43, 109–116. doi: 10.1038/ng.740

Sierro, N., Battey, J. N., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., et al. (2014). The tobacco genome sequence and its comparison with those of tomato and potato. *Nat. Commun.* 5:3833. doi: 10.1038/ncomms4833

Sierro, N., Battey, J. N., Ouadi, S., Bovet, L., Goepfert, S., Bakaher, N., et al. (2013). Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biol.* 14:R60. doi: 10.1186/gb-2013-14-6-r60

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Soltis, D. E., Visger, C. J., Marchant, D. B., and Soltis, P. S. (2016). Polyploidy: pitfalls and paths to a paradigm. *Am. J. Bot.* 103, 1146–1166. doi: 10.3732/ajb.1500501

Tanaka, H., Hirakawa, H., Kosugi, S., Nakayama, S., Ono, A., Watanabe, A., et al. (2016). Sequencing and comparative analyses of the genomes of zoysiagrasses. *DNA Res.* 23, 171–180. doi: 10.1093/dnares/dsw006

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., et al. (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 114, E9413–E9422. doi: 10.1073/pnas.1708621114

Urasaki, N., Takagi, H., Natsume, S., Uemura, A., Taniai, N., Miyagi, N., et al. (2016). Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Res.* 24, 51–58. doi: 10.1093/dnares/dsw047

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., et al. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp*. 39:e1869. doi: 10.3791/1869

Van Huylenbroeck, J., De Riek, J., and De Loose, M. (2000). Genetic relationships among *Hibiscus syriacus*, *Hibiscus sinosyriacus* and *Hibiscus paramutabilis* revealed by AFLP, morphology and ploidy analysis. *Genet. Resour. Crop Evol.* 47, 335–343. doi: 10.1023/A:1008750929836

Varshney, R. K., Mohan, S. M., Gaur, P. M., Gangarao, N., Pandey, M. K., Bohra, A., et al. (2013). Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnol. Adv.* 31, 1120–1134. doi: 10.1016/j.biotechadv.2013.01.001

Wang, A., Wang, Z., Li, Z., and Li, L. M. (2018). BAUM: Improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. *Bioinformatics* 34, 2019–2028. doi: 10.1093/bioinformatics/bty020

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., and Burbano, H. A. (2017). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *Biorxiv [Preprint]*. doi: 10.1101/143537

Wu, W., Yang, Y., He, W., Rouard, M., Li, W., Xu, M., et al. (2016). Whole genome sequencing of a banana wild relative musa itinerans provides insights into lineage-specific diversification of the *Musa* genus. *Sci. Rep.* 6:31586. doi: 10.1038/srep31586

Xin-Hua, Z., Silva, Jaime, A., Teixeira da, and Ma, G. (2010). Karyotype analysis of *Santalum album* L. *Caryologia* 63, 142–148. doi: 10.1080/00087114.2010.10589719

Yang, J., Liu, D., Wang, X., Ji, C., Cheng, F., Liu, B., et al. (2016). The genome sequence of allopolyploid brassica juncea and analysis of differential homoeolog gene expression influencing selection. *Nat. Genet.* 48, 1225. doi: 10.1038/ng.3657

Yang, J., Moeinzadeh, M., Kuhl, H., Helmuth, J., Xiao, P., Haas, S., et al. (2017). Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nat. Plants* 3, 696–703. doi: 10.1038/s41477-017-0002-z

Yang, J., Ye, C.,Cheng, Z., Tschaplinski, T. J., Wullschleger, S. D., Yin, W., et al. (2011). Genomic aspects of research involving polyploid plants. *Plant Cell Tissue Organ Cult (PCTOC)*. 104, 387–397. doi: 10.1007/s11240-010-9826-1

Yoshida, K., Schuenemann, V. J., Cano, L. M., Pais, M., Mishra, B., Sharma, R., et al. (2013). The rise and fall of the phytophthora infestans lineage that triggered the irish potato famine. *Elife* 2:e00731. doi: 10.7554/eLife.00731

Zhou, Y., Massonnet, M., Sanjak, J. S., Cantu, D., and Gaut, B. S. (2017). Evolutionary genomics of grape (*Vitis vinifera* ssp. *Vinifera)* domestication. *Proc. Natl. Acad. Sci. U.S.A.* 114, 11715–11720. doi: 10.1073/pnas.1709257114

Zimin, A., Maldonado, C. E., Yepes, M., Mockaitis, K., Moncada, P., Ganote, C., et al. (2018). "W204: chromosome scale scaffolding of the high-quality genome assemblies of the allotetraploid coffea arabica and its maternal ancestor *C. eugenioides* and validation using genetic and physical mapping data," in *26th PAG Conference* (San Diego, CA).

Zimin, A. V., Puiu, D., Luo, M. C., Zhu, T., Koren, S., Marcais, G., et al. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* 27, 787–792. doi: 10.1101/gr.213405.116

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: 25 September 2018
doi: 10.3389/fpls.2018.01414

Comparative Analysis of Homologous Sequences of *Saccharum officinarum* and *Saccharum spontaneum* Reveals Independent Polyploidization Events

*Anupma Sharma[1], Jinjin Song[2], Qingfan Lin[1], Ratnesh Singh[1], Ninfa Ramos[3], Kai Wang[2], Jisen Zhang[2], Ray Ming[2,4] and Qingyi Yu[1,2,5]**

[1] Texas A&M AgriLife Research and Extension Center at Dallas, Texas A&M University System, Dallas, TX, United States, [2] FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Key Laboratory of Genetics, Breeding and Multiple Utilization of Crops, Ministry of Education, Haixia Institute of Science and Technology, College of Life Science, Fujian Agriculture and Forestry University, Fuzhou, China, [3] Texas A&M AgriLife Research Center at Weslaco, Texas A&M University System, Weslaco, TX, United States, [4] Department of Plant Biology, University of Illinois at Urbana–Champaign, Urbana, IL, United States, [5] Department of Plant Pathology and Microbiology, Texas A&M University, College Station, TX, United States

OPEN ACCESS

Edited by:
Jun Yang,
Shanghai Chenshan Plant Science Research Center (CAS), China

Reviewed by:
Fernando Carlos Gómez-Merino, Colegio de Postgraduados (COLPOS), Mexico
Renato Vicentini, Universidade Estadual de Campinas, Brazil
Silvia Perea, Museo Nacional de Ciencias Naturales (MNCN), Spain
Niranjan Baisakh, Louisiana State University, United States
Wei Yao, Guangxi University, China

***Correspondence:**
Qingyi Yu
qyu@ag.tamu.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: 15 March 2018
Accepted: 06 September 2018
Published: 25 September 2018

Citation:
Sharma A, Song J, Lin Q, Singh R, Ramos N, Wang K, Zhang J, Ming R and Yu Q (2018) Comparative Analysis of Homologous Sequences of Saccharum officinarum and Saccharum spontaneum Reveals Independent Polyploidization Events. Front. Plant Sci. 9:1414. doi: 10.3389/fpls.2018.01414

Sugarcane (*Saccharum* spp. hybrids) is an economically important crop widely grown in tropical and subtropical regions for sugar and ethanol production. However, the large genome size, high ploidy level, interspecific hybridization and aneuploidy make sugarcane one of the most complex genomes and have long hampered genome research in sugarcane. Modern sugarcane cultivars are derived from interspecific hybridization between *S. officinarum* and *S. spontaneum* with 80–90% of the genome from *S. officinarum* and 10–20% of the genome from *S. spontaneum*. We constructed bacterial artificial chromosome (BAC) libraries of *S. officinarum* variety LA Purple (2n = 8x = 80) and *S. spontaneum* haploid clone AP85-441 (2n = 4x = 32), and selected and sequenced 97 BAC clones from the two *Saccharum* BAC libraries. A total of 5,847,280 bp sequence from *S. officinarum* and 5,011,570 bp from *S. spontaneum* were assembled and 749 gene models were annotated in these BACs. A relatively higher gene density and lower repeat content were observed in *S. spontaneum* BACs than in *S. officinarum* BACs. Comparative analysis of syntenic regions revealed a high degree of collinearity in genic regions between *Saccharum* and *Sorghum bicolor* and between *S. officinarum* and *S. spontaneum*. In the syntenic regions, *S. spontaneum* showed expansion relative to *S. officinarum*, and both *S. officinarum* and *S. spontaneum* showed expansion relative to sorghum. Among the 75 full-length LTR retrotransposons identified in the *Saccharum* BACs, none of them are older than 2.6 mys and no full-length LTR elements are shared between *S. officinarum* and *S. spontaneum*. In addition, divergence time estimated using a LTR junction marker and a syntenic gene shared by 3 *S. officinarum* and 1 *S. spontaneum* BACs revealed that the *S. spontaneum* intergenic region was distant to those from the 3 homologous regions in *S. officinarum*. Our results suggested that *S. officinarum* and *S. spontaneum* experienced at least two rounds of independent polyploidization in each lineage after their divergence from a common ancestor.

Keywords: sugarcane, *Saccharum*, polyploidization, genetic divergence, retrotransposon

INTRODUCTION

Sugarcane (*Saccharum* spp. hybrids) produces approximately 80% of the world's sugar production and is also an important source of biomass. Due to its high productivity, sugarcane is used as biorefineries for the production of biomass, bioenergy and biomaterials (Botha and Moore, 2014; Gómez-Merino et al., 2014). Sugarcane belongs to the genus *Saccharum* that was traditionally divided into six species, two wild species *S. spontaneum* and *S. robustum*, and four cultivated species *S. officinarum*, *S. edule*, *S. barberi*, and *S. sinense* (Zhang et al., 2013). However, as originally proposed by Irvine (1999), recent evidence based on morphological, cytological and population structure supported the classification of genus *Saccharum* into two horticultural species, *S. spontaneum* and *S. officinarum,* of which the latter one includes the other four *Saccharum* species and their interspecific hybrids (Zhang et al., 2013). *Saccharum* spp. and *Sorghum bicolor* belong to the grass tribe Andropogoneae in the subfamily Panicoideae. Within the tribe Andropogoneae, *Saccharum*, *Miscanthus*, *Erianthus*, *Narenga*, and *Sclerostachya* form a closely related interspecific breeding group - commonly known as the 'Saccharum complex.'

Saccharum officinarum (2n = 80) has high sugar content and low fiber, but poor disease resistance. *S. spontaneum* (2n = 36–128) is a low sugar, high fiber, disease-resistant species. Modern sugarcane cultivars are mainly derived from interspecific hybridization between *S. officinarum* and *S. spontaneum* to combine high sugar content from *S. officinarum* and disease resistance from *S. spontaneum*. Modern sugarcane hybrids are complex polyploids and aneuploids (2n = 80–140) and are comprised of 70–80% of chromosomes from *S. officinarum*, 10–20% from *S. spontaneum*, and 10% recombinants (D'Hont et al., 1996). The uneven progenitor genome contribution in the interspecific hybrids of sugarcane is due to a phenomenon called female restitution, wherein chromosome transmission is 2n from the female parent *S. officinarum* and n from the male parent *S. spontaneum* (Bremer, 1961).

Whole genome duplication (polyploidy) is common in plants and has been linked to rapid speciation and adaption (Otto and Whitton, 2000; Soltis et al., 2009; Van de Peer et al., 2017). Polyploids are classified as autopolyploids, allopolyploids, or segmental allopolyploids (Stebbins, 1947). Autopolyploids arise via whole genome duplication within the same species; allopolyploids arise via hybridization between two different species with concominant genome doubling; and segmental allopolyploids carry two partially differentiated genomes (Stebbins, 1947). Multiple rounds of ancient (paleo) and/or recent polyploidization events are evident in most angiosperm genomes (Soltis et al., 2009; Jiao and Paterson, 2014). Polyploidization is typically followed by genomic reorganization/fractionation that over time returns the genome to diploid state (Langham et al., 2004; Adams and Wendel, 2005). All the species in the genus *Saccharum* are polyploid and there is no related diploid or tetraploid progenitors known. Despite high ploidy, *Saccharum* species form mainly bivalents at meiosis, and display varying degrees of polysomy and preferential pairing among chromosomes. *S. robustum* shows high proportion of preferential pairing, *S. officinarum* shows some preferential pairing, *S. spontaneum* shows no preferential pairing, and the hybrids of *S. officinarum* and *S. spontaneum* display a continuous range of pairing affinities between chromosomes (D'Hont et al., 2008).

Assumption of molecular clock is useful for estimation of divergence time between species by comparing the divergence between genomic features such as genes and/or TEs. However, many factors contribute to the variation in molecular date estimates including the uncertainty in the absolute age of the evolutionary event used to calibrate the molecular clock, the use of different genes or genomic regions that may be under different selective constraints, and different methods used to estimate divergence times (Gaut et al., 1996; Gaut, 2002). The average synonymous substitution rate obtained from the grass adh1/2 alleles (6.5×10^{-9} per site per year) estimated by assuming the maize–rice divergence time of 50 million years (mys) (Gaut et al., 1996) is commonly employed to estimate the divergence time in grasses. And, a two-fold higher substitution rate of 1.3×10^{-8} mutations per site per year is commonly used to estimate the insertion time of LTR retrotransposons (Ma and Bennetzen, 2004).

The polyploidization and divergence history of *Saccharum* lineage remains poorly understood. The octaploid sugarcane genome has experienced two rounds of whole genome duplication since its divergence from sorghum, and is thus, an ideal system to study the impact of polyploidy on speciation, subgenome divergence and genomic adaption to the duplicated state (Kim et al., 2014). Recent studies have variably estimated the divergence time of sugarcane and sorghum (Jannoo et al., 2007; Wang et al., 2010; Kim et al., 2014; Vilela et al., 2017) and different models have been proposed for the type and time of polyploidy in sugarcane (Kim et al., 2014; Vilela et al., 2017). Kim et al. (2014) proposed that an allopolyploidy in the common ancestor of *Miscanthus-Saccharum* resulted in the divergence of Saccharinae and Sorghinae subtribes, and subsequent *Saccharum*-specific autopolyploidy resulted in random chromosome pairing within a group but infrequent pairing between groups. Although this scenario explains preferential pairing observed in *S. officinarum*, it does not explain no preferential pairing in *S. spontaneum*. Vilela et al. (2017) suggested that *S. officinarum* and *S. spontaneum* lineages each experienced independent autopolyploidization after their divergence. Further research is still needed to fully understand the polyploidization and divergence history of sugarcane.

The large genome size, high ploidy level, interspecific hybridization and aneuploidy make sugarcane one of the most complex genomes and have long hampered genome research in sugarcane. The two sugarcane progenitors, *S. officinarum* and *S. spontaneum* are an ideal genomic resource to infer evolutionary history of the genus *Saccharum*, as well as to study the complex mechanisms leading to the superior productivity of sugarcane cultivars. In this study, we selected and sequenced homo/homeologous BACs from *S. officinarum* and *S. spontaneum* BAC libraries, and conducted comparative analysis to assess variation in genome size, and mode and time of divergence between *Saccharum* and sorghum, and between

the modern sugarcane progenitor species, *S. spontaneum* and *S. officinarum*.

MATERIALS AND METHODS

Construction of *Saccharum officinarum* and *Saccharum spontaneum* BAC Libraries

Young leaf tissue was harvested from *Saccharum officinarum* variety LA Purple (2n = 8X = 80) and *S. spontaneum* haploid clone AP85–441 (2n = 4X = 32) and used for nuclei extraction. Nuclei was isolated following the protocol described by Ming et al. (2001). The high molecular weight DNA was extracted from nuclei and then embedded in agarose and partially digested with *Hin*d III. The fraction at approximately 120 kb was recovered and cloned into *Hin*d III linearized pSMART BAC vector (Lucigen)[1]. A total of 76,800 colonies for LA Purple and 38,400 colonies for AP85-441 were archived in 384-well plates with freezing medium. BAC clones were spotted onto high-density nylon filters (Performa II Nylon Filters, Genetix) using Q-Pix2 (Genetix) for hybridization screening.

Screening the BAC Libraries

PCR primers targeting the genes involved in sucrose, lignin, and cellulose biosynthesis pathways were designed using Primer Premier 5 software[2] and used for RT-PCR amplification. PCR products were purified using Wizard® SV Gel and PCR Clean-Up System (Promega) and used as probes to screen the BAC libraries. Hybridization screening of the BAC libraries was performed using the method described by Yu et al. (2011). High-density membranes of the BAC libraries were prehybridized in 0.5 M Na_2HPO_4, 7% SDS, 1 mM EDTA, 100 $\mu g\ ml^{-1}$ heat-denatured herring sperm DNA for at least 4 h. Probes were labeled using a random primer labeling system (NEBlot Kit, New England Biolabs). The hybridization was performed overnight at 55°C in 0.5 M Na_2HPO_4, 7% SDS, 1 mM EDTA, 100 $\mu g\ ml^{-1}$ heat-denatured herring sperm DNA with ^{32}P-labeled probes. Hybridized membranes were washed twice in 0.5 × SSPE/0.5% SDS for 10 min each time.

Verification of BAC Clones

BAC DNA was isolated using the alkaline lysis method and digested with *Hin*d III. The digested DNA samples were electrophoresed through a 0.8% agarose gel. After electrophoresis, the gel was blotted onto Amersham Hybond N+ membranes (GE Healthcare) using standard methods (Sambrook et al., 1987). Southern hybridization was performed using the method described by Yu et al. (2011).

Sequencing BAC Clones and Sequence Assembly

BAC DNA was extracted from selected BAC clones using QIAGEN Large-Construct kit (Qiagen) and used for pyrosequencing on a Roche 454 GS FLX+ Titanium platform at Texas A&M AgriLife Genomics & Bioinformatics Service. Each BAC clone was labeled with a unique multiplex identifier and every 12 BACs were pooled at equal amount and sequenced on one region of a four-gasket sequencing run.

The sequence reads were assembled using Newbler with default parameter settings. Sequence reads matching the *Escherichia coli* genome and the BAC vector were removed and trimmed. The sequence gaps were filled by primer walking and/or directly sequencing PCR products when possible.

Sugarcane Repeat Database and Estimation of Repeat Content

We used both de novo and structure-based approaches to identify high-copy number repeats in the 475 sugarcane BACs, including the BACs assembled in this study and 378 sugarcane BACs downloaded from GenBank. The BACs downloaded from GenBank included 2 BACs of AP85-441 (*S. spontaneum*), 4 BACs of LA Purple (*S. officinarum*), and 372 BACs of the modern sugarcane cultivar R570 (an interspecific hybrid between *S. officinarum* and *S. spontaneum*) (**Supplementary Table S1**). The TEdenovo pipeline from the REPET package (Flutre et al., 2011) and RepeatModeler (Smit and Hubley, 2008) were used to *de novo* predict sugarcane repeats by an all-by-all comparison with default parameters. Among the de novo identified repeats that were classified as chimeric or SSR by the TEdenovo, those with less than 10 copies (at 80% coverage threshold) in the sugarcane BACs and those with matches to repeat-masked plant CDS sequences were filtered. Finally, we used ProtExcluder[3] to remove protein coding genes from repeat library by mapping putative repeats against the plant protein database where transposon proteins were excluded[4]. In addition, LTR_finder (Xu and Wang, 2007) was used to predict full-length LTR retrotransposon and TRIMs. MITE_hunter (Han and Wessler, 2010) was used to generate consensus representative sequences for sugarcane MITEs. All repeats were combined and clustered using VSEARCH (Rognes et al., 2016). The consensus sequences obtained from VSEARCH were then annotated using the RepeatClassifier script of the RepeatModeler package by comparison to the Repbase database (Jurka et al., 2005). The final non-redundant repeat database was made using CD-Hit-EST (Li and Godzik, 2006) at 80% sequence identity. The full-length LTR representatives were classified by comparing their RT domains to the ones of the classified sugarcane LTR retrotransposons (Domingues et al., 2012) and to the Gypsy Database 2.0 (Llorens et al., 2011). The repeat content of the *Saccharum* BACs was estimated by RepeatMasker (Smit et al., 1996) using the custom sugarcane repeat database.

Gene Model Prediction and Annotation

We used MAKER (Cantarel et al., 2008) to annotate genes in the assembled *Saccharum* BACs. The gene models were predicted based on the combined available evidence based on matches to the repeat database, EST/cDNA, and

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

[2] http://www.premierbiosoft.com/primerdesign/

[3] http://www.hrt.msu.edu/uploads/535/78637/CRL_Scripts1.0.tar.gz

[4] http://www.hrt.msu.edu/uploads/535/78637/alluniRefprexp070416.gz

TABLE 1 | Summary of repeat content of *Saccharum officinarum* and *Saccharum spontaneum* BACs.

Element	*S. officinarum* BACs		*S. spontaneum* BACs	
	5,848,270 bp		5,012,466 bp	
	Masked (bp)	Masked (%)	Masked (bp)	Masked (%)
Interspersed repeats				
DNA transposons				
Unknown	2865	0.05	7135	0.14
MULE-MuDR	45154	0.77	56697	1.13
PIF-Harbinger	109620	1.87	115483	2.30
TcMar-Stowaway	42664	0.73	55349	1.10
CMC-EnSpm	32453	0.55	34989	0.70
hAT (unclassified)	3085	0.05	2834	0.06
hAT-Ac	16941	0.29	8824	0.18
hAT-Tag1	1149	0.02	6434	0.13
hAT-Tip100	6316	0.11	3330	0.07
Helitron	1958	0.03	3119	0.06
Retroelements				
LTRs				
Unknown	3801	0.06	5160	0.10
Copia (unclassified)	39635	0.68	18973	0.38
Copia-Ale	51928	0.89	90752	1.81
Copia-Ang	109938	1.88	80897	1.61
Copia-Iva	23128	0.40	22273	0.44
Copia-Max	754158	12.90	451646	9.01
Copia-Tor	34550	0.59	21755	0.43
Gypsy (unclassified)	**346401**	**5.92**	**142809**	**2.85**
Gypsy-Ath	63989	1.09	117340	2.34
Gypsy-Crm	42722	0.73	21850	0.44
Gypsy-Del	**816169**	**13.96**	**366879**	**7.32**
Gypsy-Rei	43787	0.75	29932	0.60
Gypsy-Tat	292738	5.01	278733	5.56
LINE/L1	10443	0.18	8146	0.16
LINE/RTE-BovB	35478	0.61	12619	0.25
SINE/tRNA	861	0.01	1292	0.03
Unknown	45898	0.78	42075	0.84
Total interspersed repeats	**2977829**	**50.92**	**2007325**	**40.05**
Simple sequence repeats				
Low complexity	7814	0.13	7660	0.15
Satellite	10594	0.18	12551	0.25
Simple repeat	109561	1.87	47434	0.95
Total masked	**3105798**	**53.11**	**2074970**	**41.40**

Bold values mark large differences in repeat content between the two sugarcane progenitors.

proteins, as well as predictions by *ab initio* gene prediction programs. The repeats database included the MIPS Repeat Element Database (mips-REdat)[5] (Nussbaumer et al., 2013), the Repbase repeat database[6] (Jurka et al., 2005) and the sugarcane repeats identified in this study. The transcript evidence included five RNAseq assemblies and the in-house sugarcane ESTs. The protein evidence included the plant protein database from the ProtExcluder package and plant proteins downloaded from Phytozome (Goodstein et al., 2012). Gene predictors, SNAP (Korf, 2004) using *O. sativa* hmm parameter and AUGUSTUS (Stanke et al., 2006) using maize hmm parameter, were run within MAKER on both masked and unmasked sequence and gene models with the best AED score per locus was selected. Gene models with evidence support (AED score > 1) or PFAM domains with default parameters in InterProScan were selected. The gene models were then annotated based on homology to the UniRef90 protein database (Suzek et al., 2007).

[5]ftp://ftpmips.helmholtz-muenchen.de/plants/REdat/

[6]http://www.girinst.org/

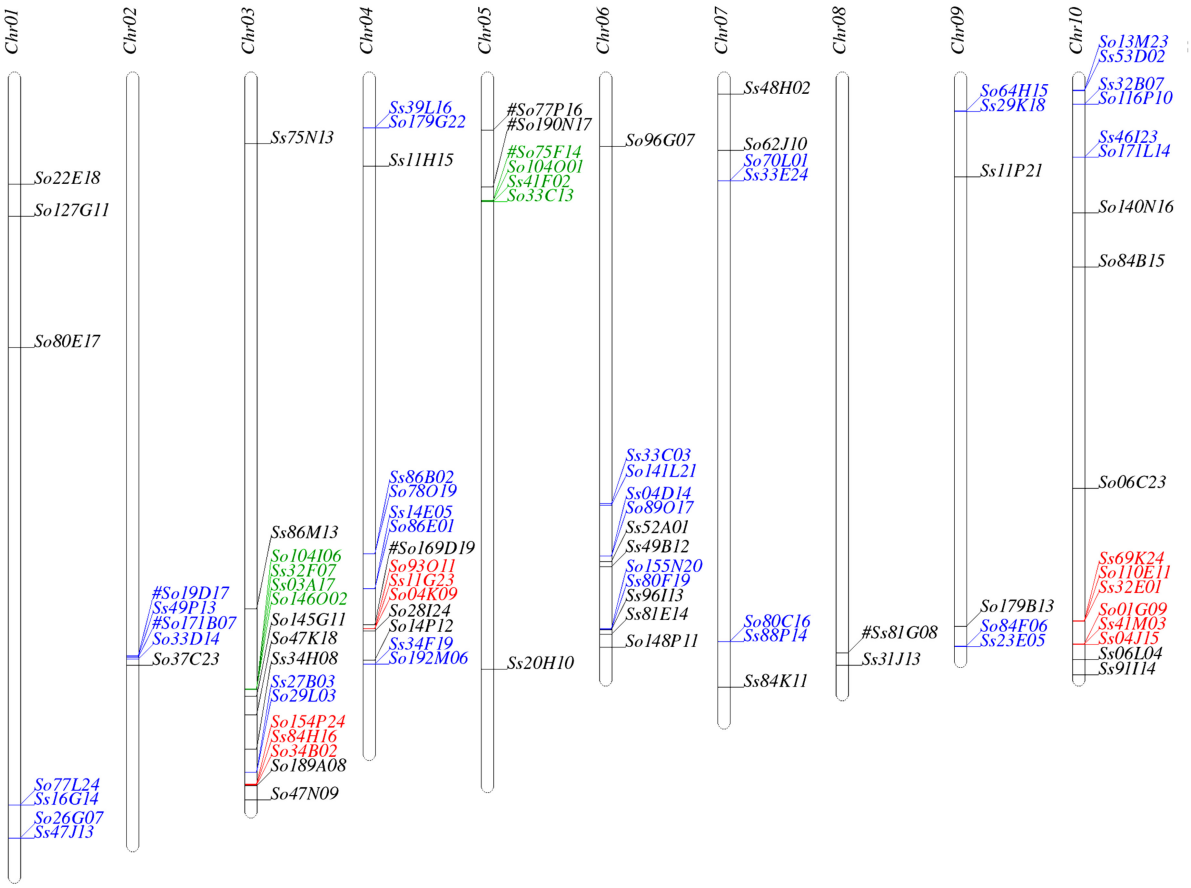

FIGURE 1 | Mapping the *Saccharum* BACs on sorghum chromosomes. Homologous groups of two, three, and four BACs are shown in blue, red, and green, respectively. #: These homologous BACs were identified by Blast.

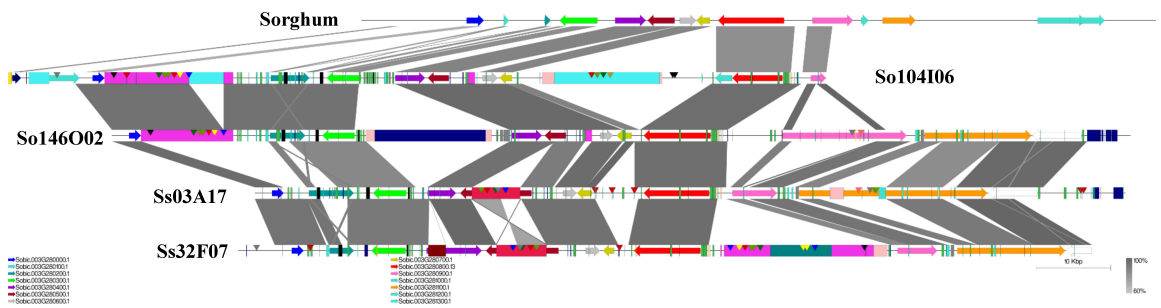

FIGURE 2 | A schematic representation of a syntenic region between *Saccharum* and sorghum and between *S. officinarum* and *S. spontaneum*. The color-coded arrows represent genes, rectangles represent repeats, and conserved domains in transposable elements are represented by pointers. The blast similarity between annotated genic regions is shown by connectors in gray color gradient. A high degree of co-linearity is shared between *Saccharum* and sorghum and between *S. officinarum* and *S. spontaneum*. The large TEs are shared by homologous regions within the same species but not by the ones from different *Saccharum* species.

Estimation of Insertion Time of Full Length LTR Retrotransposon Elements

The full-length LTR retrotransposons were identified based on full-length matches to the LTR consensus sequences using BLAST. The pairwise alignment between 5′ and 3′ LTR of each copy was generated by BLAST2seq. Pairwise alignments were conducted to estimate the number of base substitutions per site based on the Kimura 2-parameter model using MEGA7 (Kumar et al., 2016). The divergence time was estimated using the mutation rate of 1.3×10^{-8} mutations per site per year (Ma and Bennetzen, 2004). We used junctions formed at the LTR insertion sites as markers

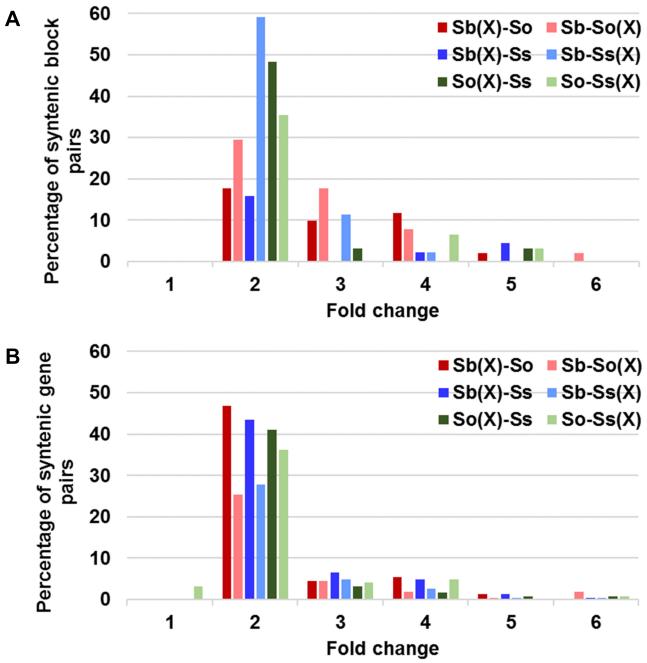

FIGURE 3 | Relative size expansion between *Saccharum* and sorghum and between *Saccharum officinarum* and *Saccharum spontaneum* in syntenic blocks **(A)** and syntenic gene pairs **(B)**. X: expansion.

(Luce et al., 2006) to identify shared insertion sites between and within *S. officinarum* and *S. spontaneum*. Up to 2 kb of the shared TE sequence (smaller than 2 kb in case of truncation) at the junction site was used for estimation of sequence divergence between paired BACs using the mutation rate of 1.3×10^{-8} mutations per site per year (Ma and Bennetzen, 2004).

Identification of Syntenic Gene Pairs and Calculation of the *Ka*/*Ks* Values

The BAC sequences were uploaded to COGE. SynMap2 at CoGe (Lyons and Freeling, 2008) was used to identify syntenic gene pairs between sorghum and *Saccharum* species (*S. officinarum* and *S. spontaneum*), and between *S. officinarum* and *S. spontaneum*. The homologous gene pairs were identified using discontinuous MegaBLAST algorithm and *e*-value less than 0.001. Relative gene order was used to compute chains of syntenic genes using DAGchainer (Haas et al., 2004), allowing a maximum distance of 30 genes and minimum number of 2 aligned gene pairs. A coverage depth ratio of 1 sorghum to 8 sugarcane genes was used. The pairwise CDS alignments for the syntenic gene pairs were generated using MACSE (Ranwez et al., 2011), and the rate of synonymous (*Ks*) and non-synonymous (*Ka*) substitutions for each syntenic gene pair was calculated using the Nei–Gojobori model in MEGA 7.0 (Kumar et al., 2016). The *Ks* values were converted to divergence times using the average synonymous substitution rate of the grass adh1/2 alleles (6.5×10^{-9} per site per year) estimated by assuming the maize–rice divergence time of 50 mys (Gaut et al., 1996).

Visualization of Orthologous BACs

The orthologous BACs were visualized using EasyFig (Sullivan et al., 2011). The repeat regions were lower case masked to allow BLAST extension from genes into neighboring shared ancestral repeats and suppress cross matches between other repeat regions.

RESULTS

BAC Library Construction, and Selection and Sequencing BACs

A BAC library of AP85-441 (*S. spontaneum*, 2n = 4X = 32) and a BAC library of LA Purple (*S. officinarum*, 2n = 8X = 80) were constructed using *Hin*d III partially digested high-molecular-weight DNA. The BAC library of AP85-441 consists of 38,400 clones and the BAC library of LA Purple consists of 76,800 clones. We randomly picked 120 clones from each library to estimate the average insert size. The average insert size of the BAC library of AP85–441 was estimated at 110 kb and the one of the BAC library of LA Purple was estimated at 120 kb. Since the genome sizes of AP85–441 and LA Purple are 3.36 Gb/2C and 7.66 Gb/2C (Zhang et al., 2012), the BAC libraries of AP85-441 and LA Purple represent approximately 1.26 and 1.20 genome equivalents, respectively.

We used the probes designed for the genes on sucrose, lignin, and cellulose biosynthesis pathways to screen the two *Saccharum* BAC libraries and selected 53 LA Purple BACs (named with So) and 44 AP85–441 BACs (named with Ss) for sequencing. The total length of the assembled sequence for the 97 BACs is 10,858,850 bp, 5,847,280 bp for the 53 So BACs and 5,011,570 bp

for the 43 Ss BACs. These sequences represent approximately 0.08% of the LA purple genome and 0.15% of the AP85–441 genome based on an estimated genome size of 7.66 Gb for LA Purple and 3.36 Gb for AP85–441 (Zhang et al., 2012).

Among the 97 BACs, 79 BACs (41 So BACs and 38 Ss BACs) could be completed by primer walking, and each was assembled into a single contig. Seven BACs (5 So BACs and 2 Ss BACs) were each assembled into two ordered and oriented contigs. Three BACs (2 So BACs and 1 Ss BACs) were each assembled into three ordered but not oriented contigs. The rest 8 BACs (5 So BACs and 3 Ss BACs) were assembled into 7–21 contigs, of which the internal contigs couldn't be ordered and oriented. Sequence assembly statistics of the 97 BACs was summarized in **Supplementary Table S2**. The assembled BACs have been deposited in GenBank and the GenBank accession numbers are MH182499-MH182581 and KU685404-KU685417.

Gene Prediction and Annotation

We used MAKER to annotate the *Saccharum* BACs and obtained 778 gene models that had an Annotation Edit distance (AED) score < 1.00 and/or had a PFAM domain. The AED score measures the congruence between an annotation with its supporting evidence, and ranges from 0 to 1, where value 0 indicates perfect match of annotation to the evidence and value 1 indicates no evidence support of annotation. We filtered 29 gene models that had TE-related PFAM domains and AED value of 1.00. The remaining 749 genes models (401 from Ss BACs and 348 from So BACs) had AED score < 1.00 and/or had a non-TE related PFAM domain. The Ss BACs have a relatively higher gene density (approximately 80 genes per Mb) compared to the So BACs (63 genes per Mb), which is consistent with the lower repeat content in Ss BACs than in So BACs (See details in "Repeat content in selected *Saccharum* BACs" and **Table 1**). The functional annotation of gene models was based on sequence similarity search in the UniRef90 database (**Supplementary Table S3**).

Approximately 86% of the gene models in Ss BACs and 89% of the gene models in So BACs had an AED ≤ 0.5 (**Supplementary Figure S1**). Although six gene models were annotated as TE-related genes, we did not filter them because they could be bona fide expressed TEs as evidenced by their AED scores < 1.00. Thirty-two gene models may be pseudogenes because they had an AED score of 1.00 but contained non-TE related PFAM domains. Twenty-eight gene models with AED < 1.00 might be caused by artifacts or spurious protein alignments as they do not contain a PFAM domain and had an eAED score of 1.00.

Repeat Content in Selected *Saccharum* BACs

We compiled a custom repeat database for sugarcane and used RepeatMasker to estimate the repeat content in selected *Saccharum* BACs using the sugarcane repeat library. The So BACs and Ss BACs contain 53 and 41% repetitive sequences, respectively (**Table 1**). This repeat content may be underestimated because some bona fide repeats may escape detection due to their low copy number in the examined BACs and the repeat consensus sequences may not capture the full range of the repeat sequence variation. Like in other plants, LTR retrotransposons are the most abundant repeat in *Saccharum*

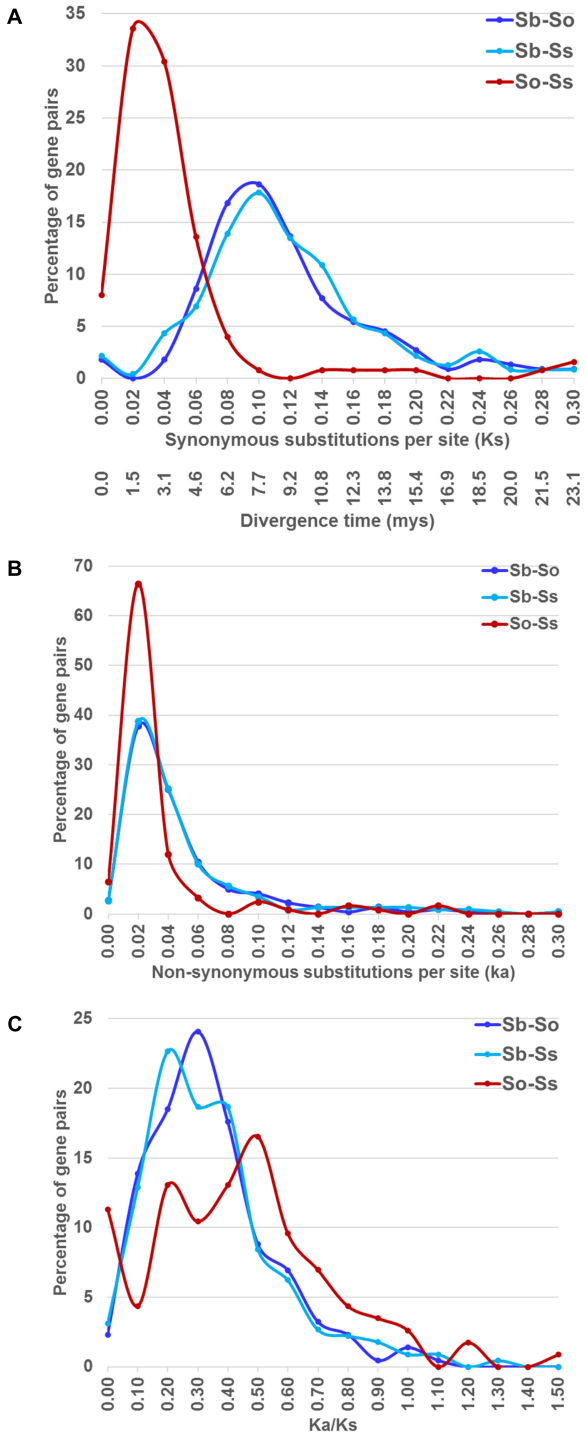

FIGURE 4 | The frequency distribution of the *Ks* and the corresponding divergence times **(A)**, *Ka* **(B)**, and *Ka/Ks* values **(C)** between sorghum and *Saccharum*, and between *S. officinarum* and *S. spontaneum*.

TABLE 2 | Number of syntenic gene pairs used for calculation of *Ks*, *Ka*, and *Ka/Ks* ratios.

	Sb-So	Sb-Ss	So-Ss
Total gene pairs	**220**	**230**	**125**
Pairs with *Ks* < 0.5	208 (94.55%)	219 (95.22%)	122 (97.60%)
Pairs with *Ka* < 0.5	209 (95.00%)	223 (96.96%)	120 (96.00%)
Pairs with *Ka/Ks* < 1.00*	215 (97.73%)	221 (96.09%)	107 (85.60%)

**The Ka/Ks value for 4 Sb–So (1.82%), 5 Sb–Ss (2.17%), and 10 So–Ss (7.87%) gene pairs could not be determined because the Ks values of these comparison were 0.*

BACs, accounting for 45% of the So BAC sequences and 33% of the Ss BAC sequences. The maximus lineage of the Ty1/Copia type and the Del lineage of the Ty3/Gypsy type elements form the largest fraction of LTR retrotransposon in both So and Ss BACs. In general, So BACs contain a higher total interspersed repeat content and total LTR retrotransposon content than Ss BACs. For the major LTR retrotransposons, a much higher percentage of Max lineage (Copia), Del lineage (Gypsy), and unclassified Gypsy LTR retrotransposons was observed in So BACs than in Ss BACs. Some of the unclassified Gypsy elements are possibly LARD elements that are related to Del.

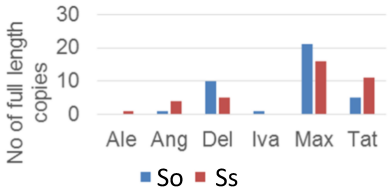

FIGURE 5 | Full-length LTR retrotransposon copies in So and Ss BACs.

Identification of Syntenic Regions Between *Saccharum* and Sorghum and Between *S. officinarum* and *S. spontaneum*

We used SynMap to identify syntenic regions between *Saccharum* and sorghum genomes. Fifty-seven syntenic blocks were identified by mapping 87 *Saccharum* BACs (45 So and 42 Ss BACs) against sorghum genome based on synteny of 205 So and 227 Ss gene models to the sorghum gene models (**Supplementary Table S3**). The syntenic regions for seven *Saccharum* BACs (6 So BACs and 1 Ss BAC) could not be identified by SynMap due to lack of a minimum of two genes syntenic to sorghum genes. We individually BLASTed these 7 *Saccharum* BACs into sorghum genome and identified seven syntenic blocks of which three have been identified by other *Saccharum* BACs using SynMap. The map location of the 94 *Saccharum* BACs on sorghum chromosomes are summarized in **Figure 1** and **Supplementary Table S3**. Based on the map location in sorghum genome, we grouped the 94 *Saccharum* BACs into 61 homology groups. We further grouped the 61 homology groups into 8 types based on the number of So and Ss BACs mapped to a sorghum syntenic region. The eight types of homology groups were named Sb-2So-2Ss, Sb-3So-1Ss, Sb-2So-0Ss, Sb-2So-1Ss, Sb-1So-2Ss, Sb-0So-1Ss, Sb-1So-1Ss, Sb-1So-0Ss. The detailed information of the 61 homology groups can be found in **Supplementary Table S4**.

A schematic representation of a syntenic region between sorghum, *S. officinarum* (BACs So104I06 and So146O02), and *S. spontaneum* (BACs Ss03A17 and Ss32F07) is shown in **Figure 2**. The schematic for additional homologous groups is shown in **Supplementary Figure S2**. A high degree of collinearity in genic regions was observed between *Saccharum* and sorghum and between *S. officinarum* and *S. spontaneum*. The collinearity was interrupted by interspersed repeats (**Figure 2**).

We used the mRNA coordinates of the syntenic genes to delineate and assess the pairwise difference in the length of the syntenic regions and the syntenic genes from sorghum and *Saccharum*. Of the 51 syntenic regions identified between *S. officinarum* and sorghum genomes, 29 showed expansion in *S. officinarum* and 22 showed expansion in sorghum (**Figure 3A**). Of the 44 syntenic regions identified between *S. spontaneum* and sorghum gnomes, 33 showed expansion in *S. spontaneum* and 11 showed expansion in sorghum (**Figure 3A**). And, of the 31 syntenic regions identified between *S. officinarum* and *S. spontaneum*, 17 had expanded in *S. officinarum* and 14 had expanded in *S. spontaneum*. Most expanded regions had up to 2-fold expansion, although there were few outliers (>3-fold expansion) that might be caused by genome rearrangements, genome mis-assembly and/or high repeat insertions. Including the outliers, the total length of the syntenic regions in sorghum was 1.1-fold of *S. officinarum* and 0.96-fold of *S. spontaneum*. After excluding the outliers (with >3-fold expansion), the total length of syntenic regions in sorghum was 0.92-fold of *S. officinarum* and 0.77-fold of *S. spontaneum*. Overall, *S. spontaneum* showed expansion relative to *S. officinarum*, and both *S. officinarum* and *S. spontaneum* showed expansion relative to sorghum.

We also compared the expansion within the annotated genes and found that the expansion in genic regions was at a much smaller scale (**Figure 3B**). Approximately half or more than half (46–65%) of gene pairs showed < 1.3-fold expansion. Approximately 20% of *S. officinarum* and *S. spontaneum* genes showed < 1.3-fold expansion relative to sorghum genes, approximately 27% of sorghum genes showed < 1.3-fold expansion relative to *S. officinarum* and *S. spontaneum* genes, and approximately 35% of *S. officinarum* genes and 30% of *S. spontaneum* genes had <1.3-fold expansion relative to *S. spontaneum* and *S. officinarum* genes, respectively. Our result indicated that the expansion of syntenic regions in *Saccharum* was largely caused by the expansion in the intergenic regions.

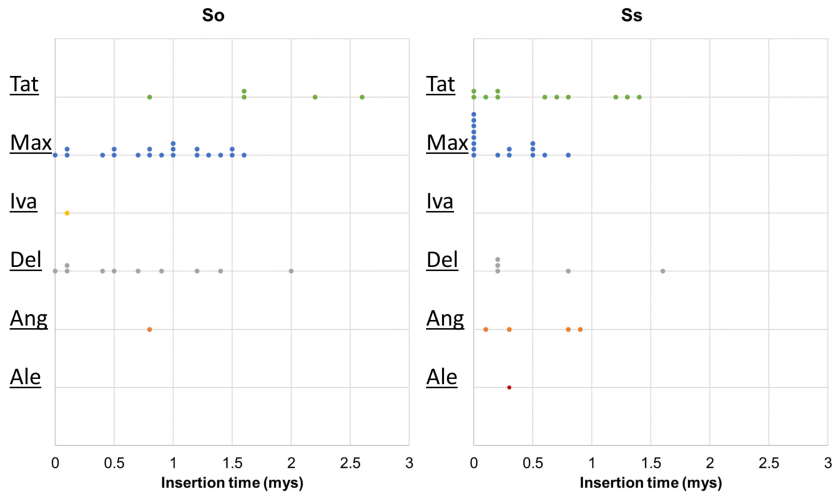

FIGURE 6 | Insertion time of LTR retrotransposon families. The insertion time of LTR retrotransposon families in So (left graph) and Ss (right graph) BACs are shown. The *X* axis represents the insertion time (mys). Each dot in the graph represents insertion time of one element and these are stacked when more than one element has the same insertion time for easy visualization of copy number. The insertion time was calculated based on substitution rate of 1.3×10^{-8} (Ma and Bennetzen, 2004).

Evolutionary Divergence Between Syntenic Gene Pairs

We estimated the *Ks* and *Ka* values of syntenic gene pairs between sorghum and *S. officinarum*, sorghum and *S. spontaneum*, and *S. officinarum* and *S. spontaneum*. The frequency distribution of the *Ks*, *Ka*, and *Ka*/*Ks* for the three comparisons is shown in **Figure 4**. The distribution of the *Ks* and *Ka* values of sorghum/*S. officinarum* and sorghum/*S. spontaneum* showed similar patterns. The peak *Ks* value for syntenic genes between sorghum and *S. officinarum* and between sorghum and *S. spontaneum* was 0.10 and the estimated divergence time was 7.7 mys. The peak *Ks* value of syntenic gene pairs between *S. officinarum* and *S. spontaneum* was 0.02 and the estimated divergence time was 1.5 mys. The peak *Ka* value for syntenic gene pairs was 0.2 for sorghum/*S. officinarum* and sorghum/*S. spontaneum*, and 0.1 for *S. officinarum*/*S. spontaneum*. The *Ka*/*Ks* values of most gene pairs (86–98%) was less than 1.00 suggesting that most syntenic gene pairs are under purifying selection (**Table 2**).

Insertion Time of LTR Retrotransposon Lineages in *S. officinarum* and *S. spontaneum*

Retrotransposon activation can be triggered by many factors including genome duplication. Therefore, it would be interesting to see the impact of genome duplication on LTR retrotransposons in *Saccharum* genomes. We extracted the full-length LTR retrotransposon copies from the So and Ss BACs and estimated their insertion times. The number of full-length LTR retrotransposon copies extracted from So (38 copies) and Ss (37 copies) BACs were similar (**Figure 5**). However, there were more Del and Max lineage members in So BACs than in Ss BACs. Overall, the full-length LTR retrotransposons in Ss BACs are younger than in So BACs (**Figure 6**). In Ss BACs, 67 and 89% of the full-length LTR retrotransposons are younger than 0.5 and 1 million years, respectively. In So BACs, 32 and 60% of the full-length elements are younger than 0.5 and 1 million years, respectively (**Figure 6**). None of the full-length LTR retrotransposons in Ss BACs are older than 2 mys, which is the estimated time when *S. officinarum* and *S. spontaneum* diverged. Interesting, none of the intact LTR retrotransposons were shared between *S. officinarum* and *S. spontaneum*.

Since *S. officinarum* and *S. spontaneum* diverged from a common ancestor recently, we would expect that remnants of some LTR retrotransposon fragments predating the divergence of *S. officinarum* and *S. spontaneum* have been retained and can be identified in the two genomes. TE insertions into the genome or within other TEs form unique junctions at their insertion sites, which can be used as markers even though the original copy has mostly degenerated (Luce et al., 2006). We identified signatures of shared LTR retrotransposon insertions between paired homologous BACs. A total of 18 LTR junction markers were identified in paired homologous BACs between *S. officinarum* and *S. spontaneum*, 11 were identified in paired homologous BACs within *S. officinarum*, and 4 were identified in paired homologous BACs within *S. spontaneum* (**Table 3**). It was estimated that *S. officinarum* and *S. spontaneum* diverged from a common ancestor approximately 1.5 – 2 mys (Jannoo et al., 2007). Interestingly, the insertion times of all the LTR junction markers shared by homologous BACs within *S. officinarum* and within *S. spontaneum* were estimated at $\leq$2 mys, while the insertion times of all except three LTR junction markers shared between *S. officinarum* and *S. spontaneum* were estimated $>$ 1.5 mys (**Figure 7**).

The LTR junction marker with the lowest divergence between *S. officinarum* and *S. spontaneum* was present in three So BACs (So104O01, So33C13, and So75F14) and one Ss (Ss41F02)

TABLE 3 | LTR junction marker identified in paired homologous BACs in *Saccharum* BACs.

Marker name	BAC1 ID	BAC1 coordinate		BAC2 ID	BAC2 coordinate		Marker type	Aligned length (%)	Identity (%)
So/Ss.1	So70L01	105063	105165	Ss33E24	101220	101315	End_Del	95.556	45
So/Ss.2	So01G09	48178	48075	Ss04J15	99569	99672	End_Max	98.333	60
So/Ss.3	So01G09	48178	48075	Ss41M03	9167	9065	End_Max	98.333	60
So/Ss.4	So104O01	49103	49198	Ss41F02	20444	20539	End_Max	100	60
So/Ss.5	So141L21	69996	70095	Ss33C03	63566	63664	End_Max	100	60
So/Ss.6	So141L21	92850	92947	Ss33C03	80868	80964	End_Max	98.333	60
So/Ss.7	So192M06	21741	21644	Ss34F19	35289	35386	End_Max	100	60
So/Ss.8	So33C13	29101	29006	Ss41F02	20444	20539	End_Max	100	60
So/Ss.9	So34B02	58358	58265	Ss84H16	86435	86527	End_Max	93.333	60
So/Ss.10	So75F14	25627	25722	Ss41F02	20444	20539	End_Max	100	60
So/Ss.11	So34B02	99058	99158	Ss84H16	78551	78450	Start_Ang	96.667	60
So/Ss.12	So01G09	50075	49976	Ss41M03	10993	10895	Start_Max	98.333	60
So/Ss.13	So01G09	50075	49976	Ss04J15	97667	97766	Start_Max	96.667	60
So/Ss.14	So141L21	89269	89368	Ss33C03	77272	77364	Start_Max	96.667	60
So/Ss.15	So141L21	68024	68123	Ss33C03	52616	52715	Start_Max	100	60
So/Ss.16	So155N20	60505	60592	Ss80F19	4670	4583	Start_Max	98.333	60
So/Ss.17	So34B02	57372	57471	Ss84H16	87420	87334	Start_Max	96.667	60
So/Ss.18	So86E01	40124	40222	Ss14E05	38389	38474	Start_Max	91.667	60
Ss/Ss.1	Ss04J15	99569	99672	Ss41M03	9167	9065	End_Max	96.667	60
Ss/Ss.2	Ss32E01	33882	33785	Ss69K24	2745	2842	Start_Ale	100	60
Ss/Ss.3	Ss04J15	97667	97766	Ss41M03	10993	10895	Start_Max	98.333	60
Ss/Ss.4	Ss32E01	7753	7654	Ss69K24	28879	28978	Start_Max	100	60
So/So.1	So04K09	13971	14067	So93O11	100606	100510	End_Del	98.276	58
So/So.2	So04K09	45961	46057	So93O11	92738	92643	End_Max	98.333	60
So/So.3	So04K09	45158	45257	So93O11	93534	93435	End_Max	100	60
So/So.4	So104O01	49103	49198	So75F14	25627	25722	End_Max	100	60
So/So.5	So104O01	49103	49198	So33C13	29101	29006	End_Max	100	60
So/So.6	So33C13	29101	29006	So75F14	25627	25722	End_Max	100	60
So/So.7	So104I06	97112	97211	So146O02	3957	3858	End_Tat	98.333	60
So/So.8	So171B07	22361	22461	So33D14	58463	58362	Start_Del	96.667	60
So/So.9	So04K09	46549	46461	So93O11	92160	92248	Start_Max	98.333	60
So/So.10	So04K09	43263	43361	So93O11	95353	95254	Start_Max	98.333	60
So/So.11	So104I06	79770	79868	So146O02	16385	16286	Start_Tat	98.333	60

BAC. A 7 kb-long multiple alignment was generated from the homologous region containing the LTR junction marker from the four BACs and used to estimate the divergence time of the intergenic region. The *K* values based on the homologous intergenic region showed that Ss41F02 diverged from the common ancestor of So104O01, So33C13, and So75F14 (K = 0.035–0.039) first, followed by the divergence of So33C13 from the common ancestor of So75F14 and So104O01 (K = 0.023 and 0.025), and So104O01 and So75F14 diverged the most recently (K = 0.012). The same pattern of divergence was observed using the divergence (Ks) of a syntenic gene shared by all four BACs (**Figure 8**).

DISCUSSION

Sugarcane (*Saccharum*) is closely related to sorghum (*Sorghum bicolor*). The two progenitors of modern sugarcane, *S. officinarum* and *S. spontaneum*, are octoploids, which have experienced two rounds of whole genome duplications since the divergence of *Saccharum* and sorghum. The divergence time between sorghum and sugarcane has been variously estimated at 8–9 mys based on Adh1 gene (Jannoo et al., 2007), 7.7 mys based on 67 pairs of orthologous genes (Wang et al., 2010), 5.0–7.4 mys based on three homologous regions (Vilela et al., 2017), and 5.4 mys by Kim et al. (2014). Similarly, the divergence time of *S. officinarum* and *S. spontaneum* was also variably estimated at 1.5–2 mys based on Adh1 gene (Jannoo et al., 2007), and 2.5–2.8 mys based on TOR haplotypes (Vilela et al., 2017). We estimated the divergence time of sugarcane and sorghum at 7.7 mys (Ks = 0.10) and the divergence time of *S. officinarum* and *S. spontaneum* at 1.5 mys (Ks = 0.02) based on synonymous distance between syntenic gene pairs from *S. officinarum*, *S. spontaneum* and sorghum genomes. Our divergence time estimates overlap with those reported in previous studies and are expected to be more accurate because we used the mutation rate of a much larger number of genes from the two sugarcane progenitors.

The evolutionary history of polyploidization events in the genus *Saccharum* is still debated. Kim et al. (2014) proposed that allopolyploidy occurred in the common ancestor of *Saccharum*

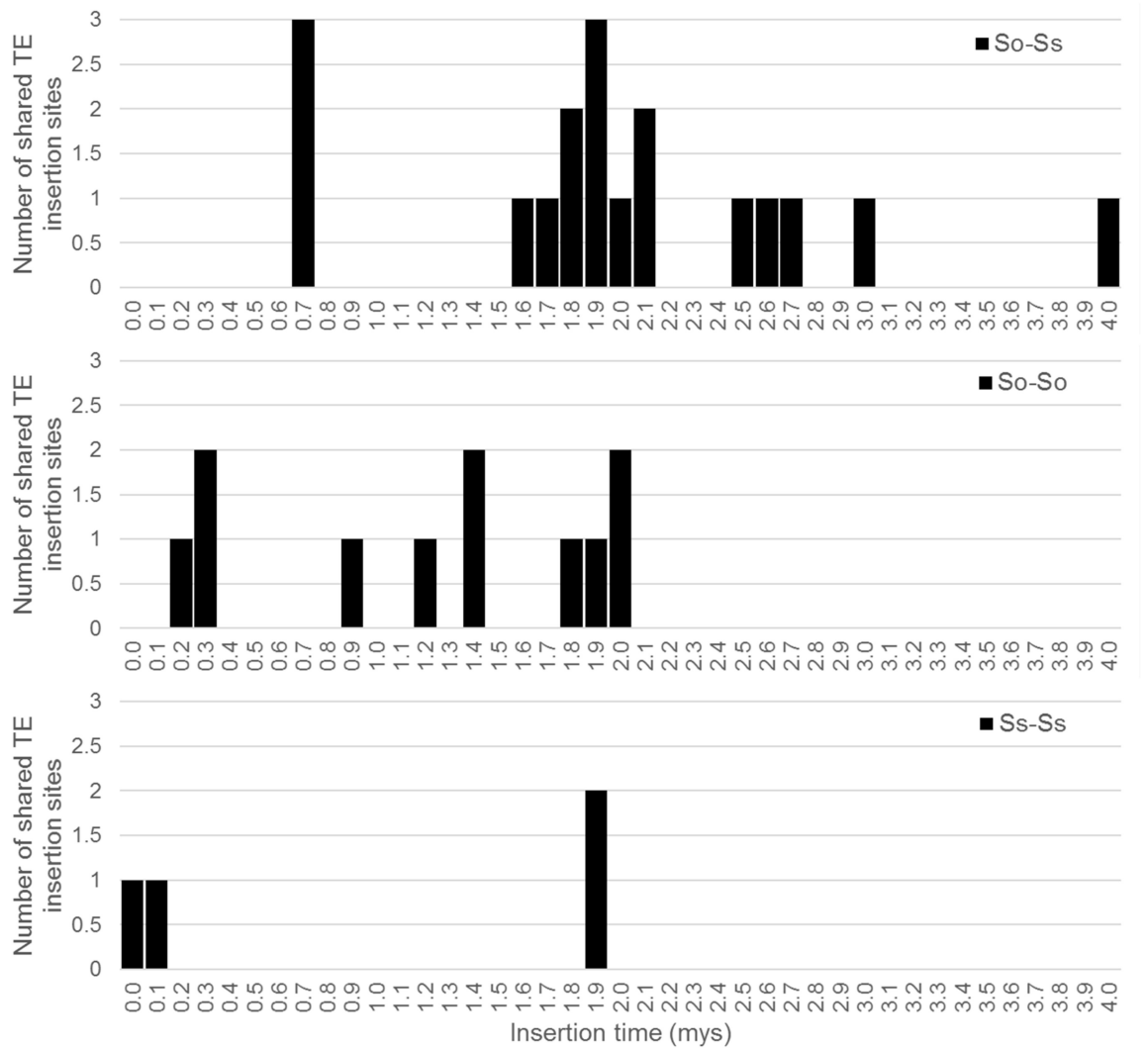

FIGURE 7 | Insertion times of LTR junction markers shared between *Saccharum* BACs. The *X* axis represents the insertion time (mys) and the *Y* axis represents the number of shared LTR junction markers. The insertion time was calculated based on substitution rate of 1.3×10^{-8} per site per year (Ma and Bennetzen, 2004).

and *Miscanthus*, followed by *Saccharum*-specific autopolyploidy based on the distribution of *Ks* value peaks between *Saccharum* and *Miscanthus* paralogs. The authors used sorghum exons to identify paralogous *Miscanthus* exons, which were subsequently used to identify sugarcane paralogs from NCBI EST database. The authors used 2368 pairs of *Miscanthus* exons (equivalent to ∼391 genes, assuming 6.05 exons per transcript estimated for sorghum) to identify sugarcane paralogs from EST database. However, it is not clear whether the sugarcane paralogs were from *S. officinarum* only, as most ESTs in GenBank are from the sugarcane hybrid R570 which contains about 20% of the genome from *S. spontaneum*. Furthermore, a different research group reported that *S. officinarum* experienced two rounds of autopolyploidization and *S. spontaneum* experienced multiple polyploidization events independently after the two species separated from each other based on the distribution of shared TEs at the TOR and LFY haplotypes derived from *S. officinarum* and *S. spontaneum* genomes in the sugarcane hybrid R570 (Vilela et al., 2017). The authors found that most TE insertions occurred after the estimated divergence of *S. officinarum* and *S. spontaneum* at 2.5 to 3.5 mys and some of these insertions were restricted to *S. officinarum* haplotypes (Vilela et al., 2017). In this study, the authors did not find evidence of allopolyploidy shared between *Saccharum* and *Miscanthus* based on *Ks* values and shared TE insertions.

If *Saccharum* lineage originated from an allopolyploid ancestor followed by *Saccharum*-specific autopolyploidy, the distribution of *Ks* values of *S. officinarum* and *S. spontaneum* gene pairs should form two peaks, the older peak representing the divergence between the two distinct sub-genomes of the allopolyploid ancestor and the younger peak representing the divergence between the genes derived from the two sub-genomes via autopolyploidization. In our study, we detected a single sharp *Ks* peak at 0.02, which represents the divergence of

Evolutionary divergence (Ks) of the homeologous gene

	So75F14_010-RA	So104O01_030-RA	So33C13_030-RA	Ss41F02_020-RA
So75F14_010-RA		0.009	0.020	0.045
So104O01_030-RA	0.009		0.024	0.047
So33C13_030-RA	0.020	0.024		0.040
Ss41F02_020-RA	0.045	0.047	0.040	

Evolutionary divergence time of the homeologous gene

	So75F14_010-RA	So104O01_030-RA	So33C13_030-RA	Ss41F02_020-RA
So75F14_010-RA		0.692	1.538	3.462
So104O01_030-RA	0.692		1.846	3.615
So33C13_030-RA	1.538	1.846		3.077
Ss41F02_020-RA	3.462	3.615	3.077	

Evolutionary divergence (K) of the homeologous LTR

	So75F14_18858-30671	So104O01_42341-54147	So33C13_24055-35863	Ss41F02_13443-25488
So75F14_18858-30671		0.012	0.023	0.035
So104O01_42341-54147	0.012		0.025	0.036
So33C13_24055-35863	0.023	0.025		0.039
Ss41F02_13443-25488	0.035	0.036	0.039	

Evolutionary divergence time of the homeologous LTR

	So75F14_18858-30671	So104O01_42341-54147	So33C13_24055-35863	Ss41F02_13443-25488
So75F14_18858-30671		0.472	0.900	1.355
So104O01_42341-54147	0.472		0.946	1.390
So33C13_24055-35863	0.900	0.946		1.514
Ss41F02_13443-25488	1.355	1.390	1.514	

FIGURE 8 | Pairwise evolutionary distance for a homeologous gene and a LTR. The homeologous gene and LTR sequence share a higher similarity among the three So BACs than to the Ss BAC.

S. officinarum and *S spontaneum* at 1.5 mys. Our result does not support the hypothesis of allopolyploidy occurred in the ancestor of *Saccharum* and *Miscanthus* followed by *Saccharum*-specific autopolyploidy.

Transposable elements form a large fraction of plant genomes. Although transposable element activity is tightly controlled in plant genomes by silencing or eliminating the TE copies, retrotransposition of TEs can be induced by stress (Wessler, 1996; Ito et al., 2016), tissue culture (Hirochika et al., 1996; Rhee et al., 2010), or events such as hybridization and polyploidy (Kashkush et al., 2002; Vicient and Casacuberta, 2017). Transposable element activation following polyploidy has been reported in numerous studies. Periodic bursts of centromeric LTR retrotransposon activity occurred after allopolyploidy through repeated formation of recombinants in maize genome (Sharma et al., 2008). Similarly, specific LTR retrotransposon families showed proliferation following autopolyploidy in the Buckler Mustard species complex (Bardil et al., 2015) and allopolyploidy in several other plant systems (Parisod et al., 2010; Senerchia et al., 2014). With the passage of time, TE insertions degenerate due to mutations, nested insertions, and deletions, making it difficult to identify shared insertions in diverged genomes. The half-life of LTR retrotransposons is shorter in smaller genomes such as *Arabidopsis* and rice and longer in large genomes such as wheat (Wicker and Keller, 2007). The half-life of LTR retrotransposons in rice, one of the smallest cereal genomes, was estimated at 4–6 my (Ma and Bennetzen, 2004; Zhang and Gao, 2017), which is longer than the estimated time of allopolyploidy in sugarcane at 3.8–4.6 mys (Kim et al., 2014). Surprisingly, most full-length LTR retrotransposon copies have inserted recently in both *S. officinarum* and *S. spontaneum*, long after the time of allopolyploidy (3.8 mys) proposed by Kim et al. (2014). In fact, of the 38 full-length retrotransposon elements identified in *S. officinarum* and 37 elements in *S. spontaneum*, none of them were older than 2.6 my and most had inserted in *S. officinarum* and *S. spontaneum* within the recent 1.6 and 0.9 my, respectively. Although a few full-length LTR retrotransposon insertions were shared by homologous chromosomes within *S. officinarum* and within *S. spontaneum*, no full-length elements were shared between *S. officinarum* and *S. spontaneum*. If retrotransposition was activated following allopolyploidy, a large number of young TEs should be identified in both *S. officinarum* and *S. spontaneum* genomes. A dearth of TE insertions shared by *S. officinarum* and *S. spontaneum* supports the latter hypothesis that two or more autopolyploidization events occurred independently in *S. officinarum* and *S. spontaneum* after their divergence.

Contrary to earlier expectations, however, it is possible that retrotransposition was not activated following allopolyploidy in *Saccharum* or that the LTR insertions were purged from *Saccharum* genome rather quickly.

Although no shared full-length LTR retrotransposons were identified in *S. officinarum* and *S. spontaneum*, several remnants of shared TEs were identified based on unique TE junctions in *S. officinarum* and *S. spontaneum*. In general, the estimated number of nucleotide substitutions per site (*K*) between *S. officinarum* and *S. spontaneum* were much higher than those between homologous regions within *S. officinarum* and within *S. spontaneum*. Divergence time estimated using an intergenic region harboring a TE-junction shared by 3 So and 1 Ss BACs revealed that the *S. spontaneum* intergenic region was distant to those from the 3 homologous regions in *S. officinarum*. In addition, the same pattern of divergence was observed using the divergence (*Ks*) of a syntenic gene shared by all four BACs. Our result supports the latter hypothesis that *S. officinarum* experienced independent autopolyploidization events following its divergence from *S. spontaneum* (Vilela et al., 2017). However, we cannot exclude the possibility that high recombination and gene conversion may have homogenized the regions we examined from *S. officinarum* and *S. spontaneum*. Therefore, close examination of shared TEs at several other locations is warranted.

In summary, *S. officinarum* and *S. spontaneum* share a high degree of collinearity in genic regions. We did not find evidence of an early allopolyploidy in *Miscanthus–Saccharum* ancestor as proposed by Kim et al. (2014). The presence of many young LTR TEs, the absence of TEs closer to the proposed time of allopolyploidy, and high similarity of intergenic regions and a syntenic gene in at least 3 So BACs relative to the Ss BAC lend strong support to the hypothesis that *S. officinarum* and *S. spontaneum* experienced at least two rounds of independent polyploidizations in each lineage after their divergence from each other roughly 2 mys. The *S. officinarum* and *S. spontaneum* BAC libraries are a valuable resource for genomic studies of *Saccharum* and provide the foundation for identification of *S. spontaneum* and *S. officinarum* fractions in modern sugarcane genome. These BAC libraries can also be used for identification and characterization of targeted gene families, and for comparative and evolutionary genomics studies in sugarcane.

DATA AVAILABILITY STATEMENTS

The assembled BACs have been deposited in GenBank under the accession numbers MH182499-MH182581 and KU685404-KU685417.

AUTHOR CONTRIBUTIONS

QY and RM designed the experiments, coordinated, and organized the all research activities. AS, JS, QL, RS, NR, KW, and JZ conducted the experiment and data analysis. AS and QY wrote the manuscript. All the authors read and approved the final manuscript.

FUNDING

This project was funded by the United States Department of Energy Office of Science and Office of Biological and Environmental Research (BER) grant no. DESC0010686 to RM and QY, Texas A&M AgriLife Research Bioenergy Initiatives to QY, Texas A&M AgriLife Research Genomics Seed Grant Program to QY, the United States Department of Agriculture National Institute of Food and Agriculture Hatch Project TEX0-1-9374 to QY, and the National Natural Science Foundation of China grant no. 31628013 to QY and KW.

ACKNOWLEDGMENTS

The open access publishing fees for this article have been covered by the Texas A&M University Open Access to Knowledge Fund (OAKFund), supported by the University Libraries and the Office of the Vice President for Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01414/full#supplementary-material

REFERENCES

Adams, K. L., and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8, 135–141. doi: 10.1016/j.pbi.2005.01.001

Bardil, A., Tayalé, A., and Parisod, C. (2015). Evolutionary dynamics of retrotransposons following autopolyploidy in the Buckler Mustard species complex. *Plant J.* 82, 621–631. doi: 10.1111/tpj.12837

Botha, F. C., and Moore, P. H. (2014). "Biomass and bioenergy," in *Sugarcane: Physiology, Biochemistry and Functional Biology*, eds P. H. Moore and F. Botha (Oxford: Wiley-Blackwell), 521–540.

Bremer, G. (1961). Problems in breeding and cytology of sugar cane. *Euphytica* 10, 59–78. doi: 10.1007/BF00037206

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., et al. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18, 188–196. doi: 10.1101/gr.6743907

D'Hont, A., Grivet, L., Feldmann, P., Glaszmann, J. C., Rao, S., and Berding, N. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Mol. Gen. Genet.* 250, 405–413. doi: 10.1007/BF02174028

D'Hont, A., Souza, G. M., Menossi, M., Vincentz, M., Van-Sluys, M.-A., Glaszmann, J. C., et al. (2008). "Sugarcane: a major source of sweetness, alcohol, and bio-energy," in *Genomics of Tropical Crop Plants Plant Genetics and Genomics: Crops and Models*, Vol. 1, eds P. H. Moore and R. Ming (New York, NY: Springer), 483–513.

Domingues, D. S., Cruz, G. M., Metcalfe, C. J., Nogueira, F. T., Vicentini, R., de, S., et al. (2012). Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. *BMC Genomics* 13:137. doi: 10.1186/1471-2164-13-137

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PLoS One* 6:e16526. doi: 10.1371/journal.pone.0016526

Gaut, B. S. (2002). Evolutionary dynamics of grass genomes. *New Phytol.* 154, 15–28. doi: 10.1046/j.1469-8137.2002.00352.x

Gaut, B. S., Morton, B. R., McCaig, B. C., and Clegg, M. T. (1996). Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc. Natl. Acad. Sci. U.S.A.* 93, 10274–10279. doi: 10.1073/pnas.93.19.10274

Gómez-Merino, F. C., Sentíes-Herrera, H. E., and Trejo Téllez, L. I. (2014). "Sugarcane as a novel biofactory: potencialities and challenges," in *Biosystems Engineering: Biofactories for Food Production in the Century XXI*, eds R. Guevara-Gonzalez and I. Torres-Pacheco (Cham: Springer).

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Haas, B. J., Delcher, A. L., Wortman, J. R., and Salzberg, S. L. (2004). DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* 20, 3643–3646. doi: 10.1093/bioinformatics/bth397

Han, Y., and Wessler, S. R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 38, e199. doi: 10.1093/nar/gkq862

Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H., and Kanda, M. (1996). Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl. Acad. Sci. U.S.A.* 93, 7783–7788. doi: 10.1073/pnas.93.15.7783

Irvine, J. E. (1999). Saccharum species as horticultural classes. *Theor. Appl. Genet.* 98, 186–194. doi: 10.1007/s001220051057

Ito, H., Kim, J.-M., Matsunaga, W., Saze, H., Matsui, A., Endo, T. A., et al. (2016). A stress-activated transposon in *Arabidopsis* induces transgenerational abscisic acid insensitivity. *Sci. Rep.* 6:23181. doi: 10.1038/srep23181

Jannoo, N., Grivet, L., Chantret, N., Garsmeur, O., Glaszmann, J. C., Arruda, P., et al. (2007). Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. *Plant J.* 50, 574–585. doi: 10.1111/j.1365-313X.2007.03082.x

Jiao, Y., and Paterson, A. H. (2014). Polyploidy-associated genome modifications during land plant evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130355. doi: 10.1098/rstb.2013.0355

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi: 10.1159/000084979

Kashkush, K., Feldman, M., and Levy, A. A. (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 160, 1651–1659.

Kim, C., Wang, X., Lee, T.-H., Jakob, K., Lee, G.-J., and Paterson, A. H. (2014). Comparative analysis of *Miscanthus* and *Saccharum* reveals a shared whole-genome duplication but different evolutionary fates. *Plant Cell* 26, 2420–2429. doi: 10.1105/tpc.114.125583

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5:59. doi: 10.1186/1471-2105-5-59

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Langham, R. J., Walsh, J., Dunn, M., Ko, C., Goff, S. A., and Freeling, M. (2004). Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* 166, 935–945.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., et al. (2011). The gypsy database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* 39, D70–D74. doi: 10.1093/nar/gkq1061

Luce, A. C., Sharma, A., Mollere, O. S. B., Wolfgruber, T. K., Nagaki, K., Jiang, J., et al. (2006). Precise centromere mapping using a combination of repeat junction markers and chromatin immunoprecipitation-polymerase chain reaction. *Genetics* 174, 1057–1061. doi: 10.1534/genetics.106.060467

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J. Cell Mol. Biol.* 53, 661–673. doi: 10.1111/j.1365-313X.2007.03326.x

Ma, J., and Bennetzen, J. L. (2004). Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12404–12410. doi: 10.1073/pnas.0403715101

Ming, R., Moore, P. H., Zee, F., Abbey, C. A., Ma, H., and Paterson, A. H. (2001). Construction and characterization of a papaya BAC library as a foundation for molecular dissection of a tree-fruit genome. *Theor. Appl. Genet.* 102, 892–899. doi: 10.1007/s001220000448

Nussbaumer, T., Martis, M. M., Roessner, S. K., Pfeifer, M., Bader, K. C., Sharma, S., et al. (2013). MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res.* 41, D1144–D1151. doi: 10.1093/nar/gks1153

Otto, S. P., and Whitton, J. (2000). Polyploid incidence and evolution. *Annu. Rev. Genet.* 34, 401–437. doi: 10.1146/annurev.genet.34.1.401

Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., et al. (2010). Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytol.* 186, 37–45. doi: 10.1111/j.1469-8137.2009.03096.x

Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E. J. P. (2011). MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons. *PLoS One* 6:e22594. doi: 10.1371/journal.pone.0022594

Rhee, Y., Sekhon, R. S., Chopra, S., and Kaeppler, S. (2010). Tissue culture-induced novel epialleles of a Myb transcription factor encoded by pericarp color1 in maize. *Genetics* 186, 843–855. doi: 10.1534/genetics.110.117929

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584 doi: 10.7717/peerj.2584

Sambrook, J., Maniatis, T., Fritsch, E. F., and Laboratory, C. S. H. (1987). *Molecular Cloning: A Laboratory Manual*, 2nd Edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Senerchia, N., Felber, F., and Parisod, C. (2014). Contrasting evolutionary trajectories of multiple retrotransposons following independent allopolyploidy in wild wheats. *New Phytol.* 202, 975–985. doi: 10.1111/nph.12731

Sharma, A., Schneider, K. L., and Presting, G. G. (2008). Sustained retrotransposition is mediated by nucleotide deletions and interelement recombinations. *Proc. Natl. Acad. Sci. U.S.A.* 105, 15470–15474. doi: 10.1073/pnas.0805694105

Smit, A. F., and Hubley, R. (2008). *RepeatModeler Open-1.0.* Available at: http://www.repeatmasker.org/

Smit, A. F., Hubley, R., and Green, P. (1996). *RepeatMasker Open-3.0.* Available at: http://www.repeatmasker.org

Soltis, D. E., Albert, V. A., Leebens-Mack, J., Bell, C. D., Paterson, A. H., Zheng, C., et al. (2009). Polyploidy and angiosperm diversification. *Am. J. Bot.* 96, 336–348. doi: 10.3732/ajb.0800079

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi: 10.1093/nar/gkl200

Stebbins, G. L. Jr. (1947). Types of polyploids: their classification and significance. *Adv. Genet.* 1, 403–429. doi: 10.1016/S0065-2660(08)60490-3

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23, 1282–1288. doi: 10.1093/bioinformatics/btm098

Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18, 411–424. doi: 10.1038/nrg.2017.26

Vicient, C. M., and Casacuberta, J. M. (2017). Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* 120, 195–207. doi: 10.1093/aob/mcx078

Vilela, M., de, M., Del Bem, L. E., Van Sluys, M.-A., de Setta, N., Kitajima, J. P., et al. (2017). Analysis of three sugarcane homo/homeologous regions suggests independent polyploidization events of *Saccharum officinarum* and *Saccharum spontaneum*. *Genome Biol. Evol.* 9, 266–278. doi: 10.1093/gbe/evw293

Wang, J., Roe, B., Macmil, S., Yu, Q., Murray, J. E., Tang, H., et al. (2010). Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics* 11:261. doi: 10.1186/1471-2164-11-261

Wessler, S. R. (1996). Turned on by stress. Plant retrotransposons. *Curr. Biol.* 6, 959–961.

Wicker, T., and Keller, B. (2007). Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Res.* 17, 1072–1081. doi: 10.1101/gr.6214107

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yu, Q., Guyot, R., de Kochko, A., Byers, A., Navajas-Pérez, R., Langston, B. J., et al. (2011). Micro-collinearity and genome evolution in the vicinity of an ethylene receptor gene of cultivated diploid and allotetraploid coffee species (Coffea). *Plant J. Cell Mol. Biol.* 67, 305–317. doi: 10.1111/j.1365-313X.2011.04590.x

Zhang, J., Nagai, C., Yu, Q., Pan, Y.-B., Ayala-Silva, T., Schnell, R. J., et al. (2012). Genome size variation in three *Saccharum* species. *Euphytica* 185, 511–519. doi: 10.1007/s10681-012-0664-6

Zhang, J., Zhou, M., Walsh, J., Zhu, L., Chen, Y., and Ming, R. (2013). "Sugarcane genetics and genomics," in *Sugarcane: Physiology, Biochemistry, and Functional Biology*, eds P. H. Moore and F. C. Botha (Hoboken, NY: John Wiley & Sons Ltd), 623–643. doi: 10.1002/9781118771280.ch23

Zhang, Q.-J., and Gao, L.-Z. (2017). Rapid and recent evolution of LTR retrotransposons drives rice genome evolution during the speciation of AA-genome *Oryza* species. *G3* 7, 1875–1885. doi: 10.1534/g3.116.037572

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.2018.01250

Genome-Wide Association Study in Pseudo-F_2 Populations of Switchgrass Identifies Genetic Loci Affecting Heading and Anthesis Dates

*Megan Taylor[1], Carl-Erik Tornqvist[2†], Xiongwei Zhao[1,3], Paul Grabowski[4†], Rebecca Doerge[1,5], Jianxin Ma[1], Jeffrey Volenec[1], Joseph Evans[6†], Guillaume P. Ramstein[2†], Millicent D. Sanciangco[6], C. Robin Buell[6], Michael D. Casler[2,4] and Yiwei Jiang[1]**

[1] Department of Agronomy, Purdue University, West Lafayette, IN, United States, [2] U.S. Department of Energy, Great Lakes Bioenergy Research Center and Department of Agronomy, University of Wisconsin-Madison, Madison, WI, United States, [3] Maize Research Institute, Sichuan Agricultural University, Chengdu, China, [4] U.S. Dairy Forage Research Center, United States Department of Agriculture-Agricultural Research Service, Madison, WI, United States, [5] Department of Biology and Department of Statistics, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA, United States, [6] U.S. Department of Energy, Great Lakes Bioenergy Research Center and Department of Plant Biology, Michigan State University, East Lansing, MI, United States

Switchgrass (*Panicum virgatum*) is a native prairie grass and valuable bio-energy crop. The physiological change from juvenile to reproductive adult can draw important resources away from growth into producing reproductive structures, thereby limiting the growth potential of early flowering plants. Delaying the flowering of switchgrass is one approach by which to increase total biomass. The objective of this research was to identify genetic variants and candidate genes for controlling heading and anthesis in segregating switchgrass populations. Four pseudo-F_2 populations (two pairs of reciprocal crosses) were developed from lowland (late flowering) and upland (early flowering) ecotypes, and heading and anthesis dates of these populations were collected in Lafayette, IN and DeKalb, IL in 2015 and 2016. Across 2 years, there was a 34- and 73-day difference in heading and a 52- and 75-day difference in anthesis at the Lafayette and DeKalb locations, respectively. A total of 37,901 single nucleotide polymorphisms obtained by exome capture sequencing of the populations were used in a genome-wide association study (GWAS) that identified five significant signals at three loci for heading and two loci for anthesis. Among them, a homolog of FLOWERING LOCUS T on chromosome 5b associated with heading date was identified at the Lafayette location across 2 years. A homolog of ARABIDOPSIS PSEUDO-RESPONSE REGULATOR 5, a light modulator in the circadian clock associated with heading date was detected on chromosome 8a across locations and years. These results demonstrate that genetic variants related to floral development could lend themselves to a long-term goal of developing late flowering varieties of switchgrass with high biomass yield.

Keywords: GWAS, candidate gene, heading, flowering time, *Panicum virgatum*

OPEN ACCESS

Edited by:
Genlou Sun,
Saint Mary's University, Canada

Reviewed by:
Xuewen Wang,
University of Georgia, United States
Zan Wang,
Institute of Animal Science (CAAS), China

***Correspondence:**
Yiwei Jiang
yjiang@purdue.edu

†Present address:
Carl-Erik Tornqvist, DNASTAR, Inc., Madison, WI, United States
Paul Grabowski, Department of Integrative Biology, The University of Texas at Austin, Austin, TX, United States
Joseph Evans, Corteva Agriscience™ Agriculture Division of DowDuPont™, Johnston, IA, United States
Guillaume P. Ramstein, Institute of Biotechnology, Cornell University, Ithaca, NY, United States

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *24 December 2017*
Accepted: *06 August 2018*
Published: *13 September 2018*

Citation:
Taylor M, Tornqvist C-E, Zhao X, Grabowski P, Doerge R, Ma J, Volenec J, Evans J, Ramstein GP, Sanciangco MD, Buell CR, Casler MD and Jiang Y (2018) Genome-Wide Association Study in Pseudo-F_2 Populations of Switchgrass Identifies Genetic Loci Affecting Heading and Anthesis Dates. Front. Plant Sci. 9:1250. doi: 10.3389/fpls.2018.01250

INTRODUCTION

Switchgrass is a C4 perennial bioenergy and forage grass. Switchgrass has been chosen as a herbaceous species for biofuel feedstock development due to its adaptation across climates, high biomass yields, tolerance to marginal conditions, and low input requirements. Switchgrass consists of upland and lowland ecotypes. Upland types are commonly tetraploid (2n = 4x = 36), but can be octoploid (2n = 8x = 72) or hexaploid (2n = 6x = 54 chromosomes), whereas lowland ecotypes are typically tetraploids (Narasimhamoorthy et al., 2008). However, switchgrass displays disomic inheritance at the tetraploid ploidy level (Casler, 2012). Upland ecotypes are adapted to northern climates with earlier flowering times and producing low biomass, while lowland ecotypes are adapted to southern climates with later flowering times and production of high biomass (Casler, 2012). Northern accessions of switchgrass reach peak biomass at flowering time, about 6–8 weeks before killing frost. Delaying flowering by 4–5 weeks can increase biomass yield by 30 to 50% (Casler, 2012). Theoretically, a delay in flowering time could be achieved by the use of either upland or lowland ecotypes. For upland ecotypes, this would involve intensive selection for late flowering within adapted germplasm. For lowland ecotypes, this would involve intensive selection for cold tolerance and adaptability within populations from southern latitudes that are already 4–6 weeks later in flowering compared to northern populations (Casler and Boe, 2003; Casler et al., 2004). Switchgrass requires short days to flower. When short-day grasses are grown in long day conditions, tillers remain vegetative for a longer period of time, resulting in more phytomers and delayed and flowering (Van Esbroeck et al., 2003).

Plants possess an internal biological clock, the circadian clock, which responds to day length and sends signals altering the plant for upcoming seasonal changes (Nuñez and Yamada, 2017). The transition from vegetative to reproductive phase causes a variety of signals and pathways to be activated in plant species. Several pathways regulate flowering time, including photoperiod, circadian clock, vernalization, autonomous, hormone and the aging (Khan et al., 2014; Blumel et al., 2015; Tornqvist et al., 2017). Photoperiod and circadian clock pathways are conserved in most plant species and work jointly using diurnal rhythms of the circadian clock gene expression to induce expression of downstream genes dependent on the light cycle (Song et al., 2010; Johansson and Staiger, 2015). Flowering in *Arabidopsis thaliana* relies on the day-length-dependent induction of FLOWERING LOCUS T (FT), encoding Florigen, a well-characterized protein which is synthesized in the leaf tissue and moves to the shoot apex to initiate floral development (Song et al., 2013). The expression of FT is primarily regulated by the transcriptional activator CONSTANS (CO), whose activity is tightly controlled by circadian clock and light (Song et al., 2013). Upon arriving in the meristem, FT binds to FLOWERING LOCUS D (FD), a bZIP transcription factor to form the FT-FD complex, regulating meristem identity genes (Wickland and Hanzawa, 2015). APETALA (AP1) and FRUITFULL (FUL) are meristem identity genes that are directly activated by the FT-FD complex and signal the primordia to begin making reproductive, non-vegetative, structures in *A. thaliana* (Blumel et al., 2015). In switchgrass, *PvFT1, PvAPL1-3*, and *PvSL1,2* have been identified as critical regulatory factors controlling floral initiation and development of floral organs (Niu et al., 2016). Overexpression of *PvFT1* was found to induce extremely early flowering in switchgrass (Niu et al., 2016).

Vernalization is required for flowering in some plant species (Kim et al., 2009), but the relevant mechanisms are not conserved. In *A. thaliana*, the vernalization pathway relies on the interaction of FLOWERING LOCUS C (FLC) with other regulators to control flowering development (Khan et al., 2014). FLC prevents flowering by preventing the transcriptional activation of SUPPRESSOR OF OVEREXPRESSION OF CO1 (SOC1) and FT by interfering with their chromatin (Helliwell et al., 2006). This affects the ability of the photoperiod and circadian clock pathway to activate floral integrators. VERNALIZATION1 (VRN1), VERNALIZATION2 (VRN2), and VERNALIZATION3 (VRN3) are three important genes that control the vernalization pathway in winter wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*) varieties (Kim et al., 2009). VRN1 encodes a MADS-Box transcription factor that is similar to AP1 and FUL meristem identity genes (Trevaskis et al., 2003). VRN2 acts as a floral repressor by blocking VRN3, the cereal homolog to FT (Yan et al., 2006). VRN1 also has a dual role in the downstream promotion of flowering using VRN3 and in cold-induced upstream repression of VRN2 (Kim et al., 2009). The dual role of VRN1 in the cereal vernalization pathway creates a positive flowering feedback loop that is not found in *A. thaliana*. In addition, the repression of VRN1 in *Brachypodium distachyon* using REPRESSOR OF VERNALIZATION1 (RVR1) is required for vernalization (Woods et al., 2017), indicating that a variety of pathways control vernalization within grasses and cereals. Several grass species do not require vernalization, including switchgrass. This may be related to flowering time being in the summer and autumn, but not in spring (Michaels and Amasino, 2000).

Homologs of FT, FTLIKE9/10 (FTL9/10), and AGAMOUS-LIKE 16 (AGL16) associated with heading date have been identified through GWAS in natural populations of switchgrass originating primarily from northern latitudes of its range (Grabowski et al., 2017). In addition, quantitative trait loci (QTL) for heading and anthesis dates have been detected in a pseudo-F_2 switchgrass population of 342 genotypes (Tornqvist et al., 2018). While phenotypic variation of flowering time in switchgrass is largely driven by the latitude of genotype origin (McMillan, 1965; Casler and Boe, 2003), genetic mechanisms underlying flowering time are not yet well understood in this species. We have developed four tetraploid switchgrass mapping populations by creating two reciprocal pseudo-F_2 crosses derived from an upland, early flowering and a lowland late flowering ecotype. Using phenotypic data for heading and anthesis dates and genotypic data based on exome capture sequencing, we conducted GWAS to identify genetic loci and candidate genes affecting heading and anthesis dates across two geographical locations. The results will provide insights into genetic mechanisms of flowering time and could assist in developing late flowering varieties of switchgrass with high biomass yield.

MATERIALS AND METHODS

Plant Materials and Planting Design

The four crosses represent second-generation crosses, originating from plant B901 of the 'Ellsworth' lowland population and plant S041 from the 'Summer' upland population. The initial cross was made in 2012 at the U.S. Dairy Forage Research Center in Madison, WI. F_1 seeds were harvested from the B901 parent as the female and germinated in 2013. Four random F_1 plants were selected and designated as numbers BS1, BS3, BS4, and BS7. The following four pseudo-F_2 crosses were made in 2013: BS1 × BS7 (318 genotypes), BS7 × BS1 (98 genotypes), BS3 × BS4 (114 genotypes), and BS4 × BS3 (58 genotypes). This created a total of 588 tetraploid genotypes. The initial cross was made between upland and lowland ecotypes to generate as much allelic variation as possible in flowering time, based on phenotypic differentiation between the two ecotypes. The second set of crosses was made to generate segregation at all relevant loci that might have been homozygous in the two original parents, which would then have been non-segregating heterozygotes in the F1 individuals.

Newly germinated seedlings were transplanted to containers (2.5 cm diameter) and grown under natural and supplemental light in a greenhouse. The tillers were split multiple times so that each F_2 genotype contained four tillers. Parents and grandparents were similarly propagated to serve as controls. After individual tillers began producing new tillers, they were separated into two groups: one or two clones of each F_2 genotype assigned to each of two locations, DeKalb, IL (41.77 N) and Lafayette, IN (40.43 N). Seedlings were transplanted in July 2014 and arranged in an augmented experimental design with 10 blocks at each location, similar to the design that was previously described (Casler et al., 2015). A total of 588 F_2 genotypes with one or two clonal replicates each were randomly assigned across the 10 blocks with all parents and grandparents included in each block as controls. The spacing between adjacent plants was 0.9 m. Plants were fertilized with 100 kg N ha^{-1} in early spring for 2015 and 2016. Weed control was maintained by applying a pre-emergent herbicide. Spot treatments and weeding were employed throughout the growing season to ensure weed pressure was minimal.

Phenotyping

Heading and anthesis dates were recorded for 588 switchgrass genotypes for both locations in 2015 and 2016. Heading date was determined when 50% of the tillers had emerged panicles, while anthesis was defined by the presence of one floret flowering for 50% of the tillers. Analysis of variance was completed in PROC MIXED with blocks as the only random effect (SAS Institute, Version 9.1, Cary, NC, United States). The augmented design was implemented by using the control genotypes within each block as adjustment factors for block-to-block variation, treating them as random covariates within the mixed model analysis. A completely random effects model was used for estimating broad-sense heritability. Heritability (H^2) was estimated as follows: $H^2 = \sigma^2_g/(\sigma^2_g + \sigma^2_{ge}/l + \sigma^2_e/rl)$, where σ^2_g is the variance component for genotype, σ^2_{ge} for genotype-by-environment, σ^2_e for error, r number of replications, and l is the number of environments (Sukumaran et al., 2016). Least squares means for both heading and anthesis dates were generated for each location separately as well as the combined location and year.

Genotyping

The four pseudo-F_2 populations were genotyped using exome capture sequencing as described previously in Evans et al. (2015) and adapted in Tornqvist et al. (2018) to produce 101-nucleotide paired-end reads, with an average of 12 M reads per sample (**Supplementary Table S1**). Reads were initially examined for quality using FastQC 0.11.5[1] and trimmed using CutAdapt v1.9.1 (Martin, 2011). Reads were aligned to the *Panicum virgatum* genome (v.1.1 hardmasked)[2] using Bowtie v0.12.7 (Langmead et al., 2009) allowing only uniquely mapping reads with a single mismatch in the seed region, and a minimum read length of 35 nucleotides. Reads were sorted and indexed with Samtools v0.1.19 (Li et al., 2009). Pileup files were generated using Samtools (v0.1.19) mpileup with BAQ disabled and map quality adjustment disabled. An initial set of single nucleotide polymorphisms (SNPs) were called using read data only at SNP loci identified previously in a northern switchgrass diversity panel of 537 individuals (Evans et al., 2015) and filtered to remove any alleles not present in the original dataset.

All raw SNPs were filtered using a custom script (**Supplementary R scripts Data.processing**) in the R programming language (R Core Team, 2014), with a call rate > 0.8 and sequencing depth > 1.6 (set from the expected depth for call rate = 0.8, according to a Poisson distribution). SNP markers were filtered for MAF > 0.05, based on simple genotype calls (directly based on co-occurrence of polymorphic reads), and genotype probabilities (as estimated from the expectation-maximization algorithm of Martin et al., 2010). Markers were retained if they satisfied the aforementioned filtering criteria in all four segregating families. After marker filtering, markers were imputed with the expectation-maximization algorithm of Poland et al. (2012) setting imputed values < 0 and > 2 to 0 and 2, respectively (**Supplementary R scripts Imputation**). After imputation, we performed a chi-squared test to test for Hardy-Weinberg equilibrium (HWE) and discarded markers for which the p-value was $<10^{-5}$ in any of the four populations.

Genome-Wide Association Analysis

Principal component analysis (PCA) and genomic kinship (K) were calculated using TASSEL 5.0 software, with centered IBS method for K (Bradbury et al., 2007). Quantile–quantile (Q-Q) plots for model comparisons of simple linear (S), PCA, genomic kinship (K), and PCA + K across traits were generated using 'qqman' package in R (R Core Team, 2014), and the best fit model was selected for association analysis of each trait. Associations between SNPs and heading or anthesis dates were analyzed using the mixed linear model (MLM) in TASSEL 5.0 software (Bradbury et al., 2007) with the following data set: (1) across

[1] www.bioinformatics.babraham.ac.uk

[2] phytozome.jgi.doe.gov

2 years in each location; (2) across two locations and 2 years; and (3) across two locations in each year. Associations were considered to be significant only at a *P*-value lower than 0.05/N, where N was 37,901 SNPs.

Candidate Gene Identification

Using the *P. virgatum* genome assembly v.1.1 (DOE-JGI)[3], candidate genes containing SNPs or adjacent to SNPs extended to 10-, 20-, 30-, and 50-kb region were identified. For genes on unanchored contigs in the *P. virgatum* genome, we predicted the genomic location based on homology to the *Setaria italica* (Bennetzen et al., 2012) and *Sorghum bicolor* (Paterson et al., 2009) genomes using the PHYTOMINE tool in PHYTOZOME (Goodstein et al., 2012). Sequences were acquired from the National Center for Biotechnology Information online nucleotide database. The BLAST and Phytomine feature were also employed to gain further information regarding sequence similarity and putative function of genes identified. The protein sequences of the published *Arabidopsis thaliana* FT (Accession AB027505) was used to search for FT genes orthologs in switchgrass (v1.1) using BLASTP program with an *E*-value of 1E-5.

Gene Expression by Quantitative Real-Time RT-PCR

Based on heading and anthesis dates, two genotypes of early flowering (3 and F_2 individual 7071) and two genotypes of late flowering (1 and F_2 individual 7055) were selected for examining gene expression profile using qRT-PCR. Phytomer tissues were collected on May 8th of 2017, representing V2-V3 stages of vegetative growth. The phytomer was defined as a node, the leaf at the node, a lateral bud, and an internode (Buck-Sorlin, 2013). Each sample consisted of three tillers per plant and had three replicates for each genotype. Briefly, total RNA was isolated using a Direct-zol^{TM} RNA MiniPrep Kit (Zymo Research Corp., Irvine, CA, United States) and then reverse transcription was performed with an iScript ^{TM} cDNA Synthesis Kit (Bio-Rad, Hercules, CA, United States). A volume of 10 μL mixture was used for all qPCRs reaction containing 1 μL of cDNA, the relevant primers, and iTaq Universal SYBR® Green (Bio-Rad, Hercules, CA, United States) in Mx3000P qPCR system (Agilent Technologies, Santa Clara, CA, United States), with reaction for 10 min at 95°C followed by 40 amplification cycles of 10 s at 95°C, 30 s at 55°C, and 30 s at 72°C. Primer sequences for target genes and for switchgrass housekeeping gene of elongation factor 1-alpha (*eEF-1α*) (Gimeno et al., 2014) were listed in **Supplementary Table S2**. The method of $2^{-\Delta\Delta C_T}$ (Livak and Schmittgen, 2001) was used to calculate the relative expression level among early and late flowering genotypes. The analysis included three biological replicates and three technical replicates for each sampling time.

Data Availability

Raw reads for pseudo-F_2 populations have been deposited in the National Center for Biotechnology Information Sequence Read Archive under BioProject ID (PRJNA450338). The initial SNP calls based on the positions identified in Evans et al. (2015) are available on the Dryad Digital Repository under doi (to be released upon publication). The final filtered SNP matrices used in the analyses were shown in **Supplementary Table S3**.

[3]http://phytozome.jgi.doe.gov/

RESULTS

Phenotypic Variation

Heading and anthesis dates were recorded in 2015 and 2016 at Lafayette and DeKalb locations. Across 2 years, heading ranged from 177 to 211 days of the year and anthesis ranged from 193 to 245 days of the year at Lafayette, while heading ranged from 179 to 252 and anthesis varied from 186 to 261 days of the year at DeKalb (**Table 1**). Overall, there was a 34- and 73-day difference in heading and a 52- and 75-day difference in anthesis at the Lafayette and DeKalb locations, respectively. There were significant variation with respect to genotype, location, year, genotype × year, genotype × location, and genotype × location × year (**Table 2**). Broad-sense heritability was high enough for heading date (0.73) and anthesis date (0.74). Thus, GWAS analyses were conducted separately for each location and year and combined across locations and/or years only when the results were homogeneous. The trends in the relationship between heading or anthesis date and accumulated growing degree (GDD) were very similar in year 2015 and 2016 (**Supplementary Figure S1**). There was also a strong linear correlation ($r > 0.99$) between heading or anthesis date with GDD across years, genotypes, and locations (**Supplementary Figure S1**). Thus, day of year was chosen for calculating heading and anthesis dates and subsequently used for GWAS analysis.

Genotyping and Principal Component Analysis

After filtering raw SNPs and fitting segregation patterns, a total of 37,901 SNPs was generated across all populations. PCA across 588 genotypes showed differentiation among the four sibling populations (**Supplementary Figure S2**). Two distinct groups were formed in the first principal component (PC1) separating sibling populations based on reciprocal crosses (BS1 × BS7 and BS7 × BS1; BS3 × BS4, and BS4 × BS3) (**Supplementary Figure S2**).

GWAS for Heading and Anthesis

Quantile–quantile plots verified the adequate model for controlling false positives for GWAS of heading and anthesis

TABLE 1 | Range and mean values for heading and anthesis dates in Lafayette, IN and DeKalb, IL across 2015 and 2016 years.

Location	Trait	Range (day of the year)	Mean (day of the year)
Lafayette	Heading	177–211	187
	Anthesis	193–245	218
DeKalb	Heading	179–252	203
	Anthesis	186–261	233

TABLE 2 | Mixed model analysis of variance for fixed effects for heading and anthesis dates in Lafayette, IN and DeKalb, IL across two years of 2015 and 2016.

		df	Type III SS	*F*-value	Significance
Heading	Year (Y)	1	31521	609.74	***
	Location (L)	1	103081	1993.99	***
	Y × L	1	38258	740.08	***
	Genotype (G)	587	96817	3.19	***
	G × Y	566	39223	1.34	**
	G × L	352	24411	1.34	***
	G × L × Y	318	26457	1.61	***
Anthesis	Year (Y)	1	34633	797.53	***
	Location (L)	1	115341	2656.02	***
	Y × L	1	57146	1315.94	***
	Genotype (G)	583	112532	4.44	***
	G × Y	544	50519	2.14	***
	G × L	346	31693	2.11	***
	G × L × Y	257	21834	1.96	***

*Significant at 0.05 significance level. **Significant at 0.01 significance level. ***Significant at 0.001 significance level.

dates (**Supplementary Figure S3**). Comparisons of observed and expected -$\log_{10}$ (*P*) showed that PCA plus K model was most suitable for analyzing SNP-trait associations. GWAS was performed in three ways to test for associations. First, GWAS was performed for heading and anthesis dates across 2 years for the Lafayette and DeKalb locations separately. The Lafayette location had a significant SNP identified on chromosome 5b for heading (**Figure 1**), but no significant SNPs were identified for DeKalb location for both traits. Across both locations and both years, one significant SNP for heading date was identified on chromosome 8a (**Figure 1**). GWAS was also completed for each year at both locations. Data from 2015 did not yield any significant SNPs across both locations, but year 2016 data contributed to three significant SNPs (**Figure 1**). The year 2016 combined for both locations had a significant SNP for heading on chromosome 2b, while one significant SNP for anthesis date on chromosome 9a and one SNP for anthesis date in the unanchored region 14 were also identified (**Figure 1**).

GWAS and Candidate Genes for Heading and Anthesis

Heading in Lafayette Across Two Years

The SNP on chromosome 5b for heading was located within the gene *Pavir.Eb00235*, which encodes a brassinosteroid signaling regulator that regulates transcription (**Table 3**). Genotypes with homozygous C:C alleles at SNP position rs1088884 had significantly later heading date than those carrying heterozygous T:C and homozygous T:T for early heading (**Figure 2**). Several genes of interest were located within 30 kb (**Table 4**), including a homolog of FLOWERING LOCUS T (FT), homologs of ARABIDOPSIS NOD26-LIKE INTRINSIC PROTEIN (AtNLM1;2), and a homolog of LIGHT REGULATED ZINC FINGER PROTEIN (LZF1).

Heading at Lafayette and DeKalb Across Two Years

The SNP on chromosome 8a for heading at Lafayette and DeKalb across both years was located within *Pavir.Ha01813*, an exo70 family protein subunit, which functions in the production of an octameric protein implicated in tethering secretory vesicles to the plasma membrane (**Table 3**). Genotypes with homozygous A:A at SNP position rs628677 had the same heading dates compared to homozygous T:T (**Figure 2**). Several genes of interest were identified within 20 kb (**Table 4**). Notable homologs included ARABIDOPSIS PSEUDO-RESPONSE REGULATOR 5 (APPR5/PRR5) and UDP-Glycosyltransferase.

Heading at Lafayette and DeKalb in 2016

Identification of candidate genes for heading at both locations for 2016 was expanded to 50 kb due to the number of uncharacterized genes surrounding significant SNPs. The significant SNP on chromosome 2b for heading was located within *Pavir.Bb00124*, a gene with unknown function. Genotypes with homozygous alleles T:T at SNP position rs888297 had significantly later heading than those carrying heterozygous T:C and homozygous C:C for early heading (**Figure 2**). There were four genes of interest found within the 50 kb region identified by GWAS, including a homolog of PROLINE TRANSPORTER 1 (PRO1), a homolog of serine/threonine-protein kinase WNK, and a homolog of SULFOQUINOVOSYLDIACYL GLYCEROL 2 (SQDG2) (**Supplementary Table S4**).

Anthesis at Lafayette and DeKalb in 2016

The SNP on chromosome 9a associated with anthesis at both locations was located within the gene *Pavir.Ia02791* that encodes ALANYL-tRNA SYNTHETASE (ALATS) (**Table 3**). At SNP position rs712216, genotypes with homozygous alleles T:T and heterozygous T:A had substantially later anthesis dates than the one genotype carrying homozygous alleles A:A (**Figure 2**). A gene encoding a homolog of a hAT transposon superfamily protein was identified within 10 kb, and a homolog of ribonuclease H-like superfamily protein and a homolog of a RNA binding family protein were identified within the 50 kb region (**Table 4** and **Supplementary Table S4**). These genes primarily were related to nucleotide binding or protein dimerization, which could interact with developmental processes that control flowering, but currently, the function of these proteins in relation to floral development is not well understood. The SNP position rs2175421 in an undefined region was also associated with anthesis at both locations. *Pavir.J40827* encoding ADP-ribosylation factor-like factor was identified in this region (**Supplementary Table S4**). Genotypes with homozygous alleles A:A and heterozygous A:C had later anthesis dates than those carrying homozygous alleles C:C (**Figure 2**).

Gene Expression in Genotypes Contrasting With Flowering Time

The expression of four candidate genes *BZR1*, *PRR5*, *UDPG*, and *WNK* was analyzed in two early flowering genotypes (3 and 7071) and two late flowering genotypes (1 and 7075) (**Figure 3**). Relative to the early flowering genotype 3 and 7071, expression levels of *BZR1* were significantly higher in the two flowering genotypes

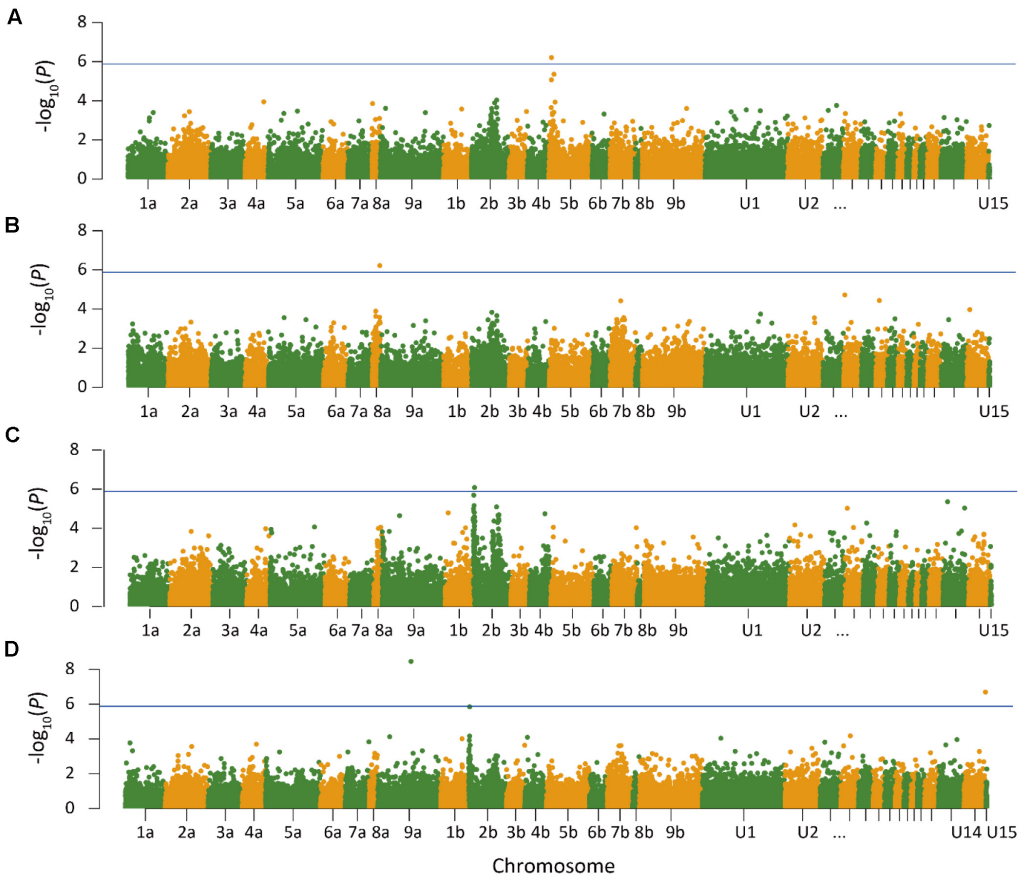

FIGURE 1 | Manhattan plot of genome wide association analysis of switchgrass (*Panicum virgatum*) for Lafayette, IN across year 2015 and 2016 **(A)**, heading for Lafayette, IN and DeKalb, IL across year 2015 and 2016 **(B)**, heading for Lafayette, IN and DeKalb, IL in year 2016 **(C)**, and anthesis for Lafayette, IN and DeKalb, IL in year 2016 **(D)**. Lines indicate threshold for significant single nucleotide polymorphism. The *x*-axis indicates the physical position of the SNPs on the 18 switchgrass chromosomes; positional information in U1–U15 refer to unanchored scaffolds and contigs.

TABLE 3 | Significant SNPs for heading and anthesis dates identified by GWAS for Lafayette and Lafayette/DeKalb location.

SNP	Chr.	Position	Trait	Sample set	Year	No. of genotypes	*P*-values	Nearest gene ID
rs1088884	5b	3772986	Heading	Lafayette	2015/2016	538	6.15E-07	*Pavir.Eb00235*
rs628677	8a	51715776	Heading	Lafayette/DeKalb	2015/2016	586	6.03E-07	*Pavir. Ha01813*
rs888297	2b	1737687	Heading	Lafayette/DeKalb	2016	570	8.03E-07	*Pavir. Bb00124*
rs712216	9a	55104939	Anthesis	Lafayette/DeKalb	2016	563	3.50E-09	*Pavir.Ia02791*
rs2175421	Undefined 14	49102800	Anthesis	Lafayette/DeKalb	2016	563	6.33E-08	*Pavir.J40827*

(**Figure 3**). The higher expression of *PRR5* was also observed in the late flowering genotype 1, compared to other three genotypes (**Figure 3**). The early flowering genotype 7071 and late genotype 1 had higher expression of *UDPG*, while 7071 showed higher expression of *WNK* (**Figure 3**).

DISCUSSION

Significant interactions for heading and anthesis dates involving genotypes, locations, and years may have contributed to variability in GWAS results. Five significant SNPs at multiple loci on chromosome 5b, 8a, 2b, 9a, and undefined region of 14 were associated with either heading or anthesis date. These signals were detected in Lafayette location across 2 years, Lafayette and DeKalb locations across 2 years, and two locations in year 2016, but no signals were detected in year 2015. The higher average temperatures, GDD and precipitation from April to September across two locations observed in 2016 than in 2015 may have contributed to variations of growth and flowering time between the 2 years. Moreover, 2016 was the second year after planting and the grass plants were more established than year 2015, which may play a role in phenotypic variation. Comparing two locations, one signal was found in Lafayette across 2 years,

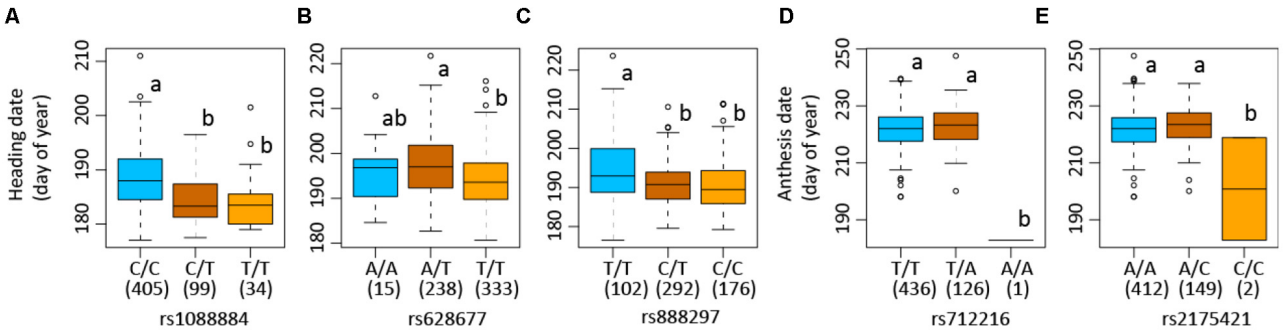

FIGURE 2 | Allelic variations of single nucleotide polymorphisms (SNPs) associated with heading and anthesis dates in switchgrass (*Panicum virgatum*). Boxplots show heading or anthesis date in days of the year. Numbers of each genotype are indicated in parenthesis in box plots. Numbers following "rs" are SNP position. Plots show samples used in genome-wide association studies in which association is detected: Lafayette, IN across year 2015 and 2016 **(A)**, heading for Lafayette, IN and DeKalb, IL across year 2015 and 2016 **(B)**, heading for Lafayette, IN and DeKalb, IL in year 2016 **(C)**, and anthesis for Lafayette, IN and DeKalb, IL in year 2016 **(D)**. Columns with the same letter were not significantly different at $P < 0.05$.

TABLE 4 | Candidate gene associations for heading and anthesis dates identified by GWAS for Lafayette and DeKalb for 2015 and 2016.

Distance (kb)	Switchgrass gene	Chr	*A. thaliana* homolog	*O. sativa* homolog	Function	Trait	Location	Year
Inside	*Pavir.Eb00235*	5B	AT1G78700.1	Os01g10610.1	BES1/BZR1 homolog 4	Heading	Lafayette	2015/2016
	Pavir.Ha01813	8A	AT5G52340.1	Os11g05880.1	Exocyst subunit exo70 family protein A2	Heading	Lafayette and DeKalb	2015/2016
10	*Pavir.Eb00236*	5B	AT4G18910.1	Os05g11560.1	NOD26-like intrinsic protein 1;2	Heading	Lafayette	2015/2016
	Pavir.Eb00237	5B	AT1G65480.1	Os01g10590.1	Flowering Locus T	Heading	Lafayette	2015/2016
	Pavir.Ha01814	8A	AT2G34320.1	Os03g30510.1	Polynucleotidyl transferase	Heading	Lafayette and DeKalb	2015/2016
	Pavir.Ha01812	8A	AT5G22870.1	Os11g05870.1	Late embryogenesis abundant	Heading	Lafayette and DeKalb	2015/2016
	Pavir.Ia02791	9A	AT1G50200.1	Os10g10244.1	Alanyl-tRNA synthetase	Anthesis	Lafayette and DeKalb	2016
	Pavir.Ia02790	9A	AT5G31412.1	Os01g52430.1	HAT transposon superfamily protein	Anthesis	Lafayette and DeKalb	2016
20	*Pavir.Bb00125*	2B	AT5G01220.1	Os07g01030.1	Sulfoquinovosyldiacylglycerol 2	Heading	Lafayette and DeKalb	2016
	Pavir.Bb00123	2B	AT2G39890.1	Os07g01090.1	Proline transporter 1	Heading	Lafayette and DeKalb	2016
	Pavir.Ha01815	8A	AT5G24470.1	Os11g05930.1	Pseudo-response regulator 5 (PRR5)	Heading	Lafayette and DeKalb	2015/2016
	Pavir.Ha01810	8A	AT3G11660.1	Os11g05860.1	NDR1/HIN1-like 1	Heading	Lafayette and DeKalb	2015/2016
	Pavir.Ha01817	8A	AT3G44190.1	Os11g05970.1	FAD/NAD(P)-binding oxidoreductase	Heading	Lafayette and DeKalb	2015/2016
30	*Pavir.Eb00232*	5B	AT1G53380.2	Os01g10680.2	Plant protein of unknown function	Heading	Lafayette	2015/2016
	Pavir.Eb00238	5B	AT1G78600.2	Os01g10580.1	Light-regulated zinc finger protein 1	Heading	Lafayette	2015/2016
	Pavir.Ha01819	8A	AT1G06140.1	Os11g05980.1	Pentatricopeptide repeat (PPR)	Heading	Lafayette and DeKalb	2015/2016
	Pavir.Ha01820	8A	AT3G11670.1	Os11g05990.1	UDP-Glycosyltransferase	Heading	Lafayette and DeKalb	2015/2016
	Pavir.Ha01808	8A	AT5G44720.2	Os09g38772.1	Molybdenum cofactor sulfurase	Heading	Lafayette and DeKalb	2015/2016
	Pavir.Bb00120	2B	AT3G09220.1	Os07g01110.1	Laccase 7	Heading	Lafayette and DeKalb	2016

but not in DeKalb. Similarly, the higher average temperatures and GDD from April to September in Lafayette than DeKalb may cause variation of growth and possibly lead to delayed flowering time at DeKalb. There were fewer signals detected for anthesis than heading in this study. Some genotypes with heading emergence but did not flower at the end of growing season. This may have influenced signal identifications related to anthesis.

It is worth mentioning that QTL mapping for flowering time has been performed using one of the four pseudo-F_2 populations (BS1 × BS7) and same phenotypic data for this particular population in this study. Interestingly, 9 QTLs have been detected including one for heading and one for anthesis on chromosome 2b in Dekalb, 2016 and one for heading on chromosome 8a across Lafayette and DeKalb locations and 2 years of 2015 and 2016 (Tornqvist et al., 2018). Although these QTLs on chromosome 2b and 8a were not the same one identified in this study, all signals detected through GWAS and QTL mapping could be important targets for elucidating genetic control of flowering time in switchgrass. In addition, QTL mapping identified one signal on chromosome 2a separately for each location and year and combined across locations and/or years, but was not found in this study. Within this QTL region on chromosome 2a, homologs of flowering time genes were identified such as PSEUDO RESPONSE REGULATOR 5 (PRR5), SUPPRESSOR OF FRIGIDA 4, and APETALA 1, which are involved in the circadian clock, vernalization, and floral meristem identity, respectively (Tornqvist et al., 2018).

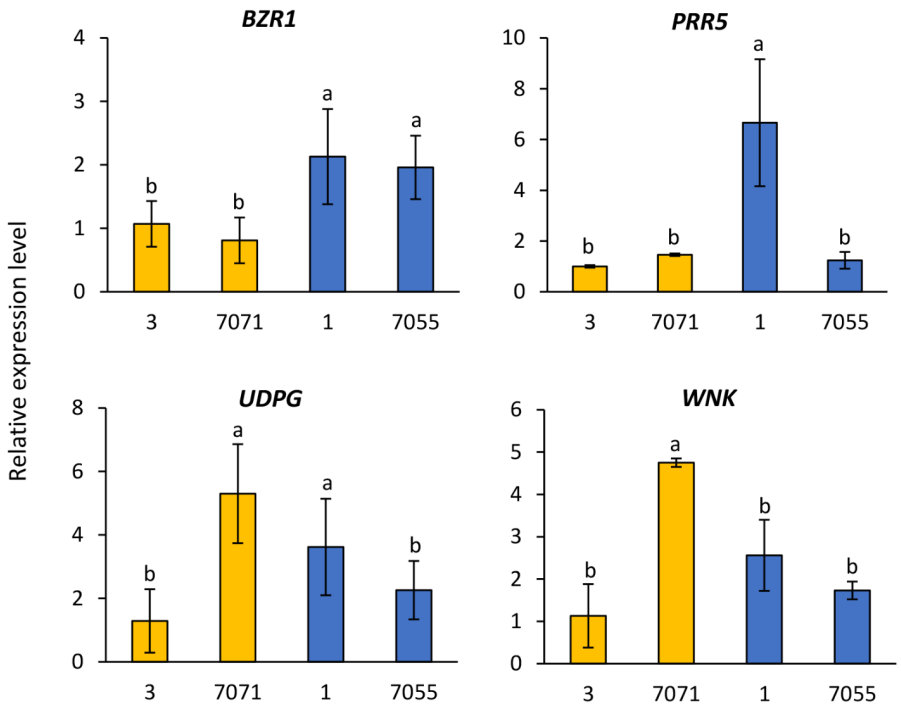

FIGURE 3 | Relative gene expression level in the early flowering genotypes (3 and 7071) and late flowering genotypes (1 and 7055). The expression data were normalized relative to early genotype (3). Columns with the same letter were not significantly different at $P < 0.05$. BES1/BZR1, Brassinosteroid signaling positive regulator; PRR5, PSEUDO-RESPONSE REGULATOR 5; UDPG, UDP-Glycosyltransferase; WNK, Serine/threonine-protein kinase WNK.

Genetic control of flowering has also been reported in Setaria, a panicoid grass closely related to switchgrass (Mauro-Herrera et al., 2013). Through analysis of flowering time of 182 F7 recombinant inbred lines developed from a cross between foxtail millet (*Setaria italica*) and its wild relative green foxtail (*Setaria viridis*) (Wang et al., 1998; Bennetzen et al., 2012), a total of 16 QTLs were deteced in eight trials conducted in greenhouses, field and growth chambers at different locations (Mauro-Herrera et al., 2013). Underlying QTL regions, flowering pathway genes were identified from rice (*Oryza sativa*), maize (*Zea mays*), sorghum, and *Arabidopsis* including PRR95, PRR59, GI involved in circadian clock and CONSTANS involved in the photoperiod pathway. Compared to the previous and current studies in switchgrass, the results in Setaria supported that some flowering genes such as PRR5 could play an important role in regulating flowering time across a range of grass species and other environmental factors.

Candidate genes related to plant growth and flowering were identified within 50 kb of significant SNPs. Such genes included BES1/BZR1 homolog 4, FLOWERING LOCUS T (FT), pseudo-response regulator 5 (PRR5), light-regulated zinc finger protein 1, UDP-Glycosyltransferase, hAT transposon superfamily protein, helix-loop-helix DNA-binding protein, and serine/threonine-protein kinase WNK (**Table 4** and **Supplementary Table S4**). The significant SNP related to heading on chromosome 5b was inside gene *Pavir.Eb00235*, which was a homolog of *BES1/BZR1* encoding a brassinosteroid signaling regulator. This SNP was deemed significant at Lafayette location across 2015 and 2016 (**Table 4**). Brassinosteroids (BRs) are a class of steroidal hormones essential for plant growth and development, including regulating flowering time (Li et al., 2010). BZR1/BES1 can bind directly to the promoter regions of the BR biosynthetic genes, *CPD* and *DWF4*, and inhibit their expression (Tanaka et al., 2005). BR biosynthetic mutants of *CPD* and *DWF4* had delayed flowering time (Li et al., 2010). In this study, higher expression levels of *BZR1* shown in the two late flowering genotypes of switchgrass compared to the early types supported that increased expression of this gene could inhibit BR biosynthetic genes, thus delaying flowering. These results support previous findings from expression analysis of flowering time genes in switchgrass (Tornqvist et al., 2017).

Candidate gene *Pavir.Ha01815* that encodes PRR5 was identified on chromosome 8a within 20 kb of SNP related to heading across Lafayette and DeKalb locations and years (**Table 4**). Through linkage mapping analysis in one of the populations used in this study, one PRR5 on chromosome 2a related to heading and anthesis date was detected in Lafayette or DeKalb location (Tornqvist et al., 2018). Although the two *PRR5* were located on different chromosomes, the results from GWAS and QTL mapping suggest that *PRR5* plays a role in regulating flower time. PRR5 has been shown to modulate light input into the circadian clock (Nakamichi et al., 2005). In *A. thaliana,* PRR5 regulates the period of free-running rhythms of certain clock-controlled genes including CIRCADIAN CLOCK ASSOCIATE 1 (CCA1) and ARABIDOPSIS PSEUDO-RESPONSE REGULATOR 1 (APRR1)

(Kamioka et al., 2016). The PRR5 mutant of *A. thaliana* showed late flowering under continuous light, late flowering under long days, and early flowering under short days (Yamamoto et al., 2003; Nakamichi et al., 2007; Niinuma et al., 2008). However, *PRR5*-overexpressing transgenic lines of *Arabidopsis* flowered earlier than the wild-type plants under both long and short day conditions (Sato et al., 2002). In barley (*Hordeum vulgare*), *PRR59* and *PRR95*, homologs of *At PRR5* or *AtPRR9*, respectively, exhibited higher expression abundances in late flowering genotypes compared to the early flowering genotypes under long day conditions (Campoli et al., 2012). The transcriptomic analysis showed that expression of *PRR5* was either up- or down-regulated or remained unchanged in early or late flowering genotypes at different growth stages of switchgrass (Tornqvist et al., 2017). In this study, one late flowering genotype had a much higher expression of *PRR5* relative to other early and late flowering genotypes. Collectively, it appears that expression of *PRR5* genes varies across plant species, genotypes and environmental conditions.

The candidate gene *Pavir.Ha01820*, a homolog of UDP-Glycosyltransferase was identified on chromosome 8a within 30 kb of a SNP associated with heading date across the Lafayette and DeKalb locations and years (**Table 4**). Glycosyltransferases (GTs) are the enzymes for the glycosylation of plant compounds (Bowles et al., 2005). GTs might play an important role in the maintenance of cell homeostasis and regulation of plant growth and defense responses (Jones and Vogt, 2001; Lim and Bowles, 2004). An *Arabidopsis* mutant of *UGT87A2*, encoding a putative family 1 GT, had delayed flowering time, while overexpression of *UGT87A2* caused much earlier flowering than the mutant (Wang et al., 2012). They further verified that *UGT87A2* regulated flowering time via the flowering repressor FLC (Wang et al., 2012). The transcriptomic analysis indicated up, down, or unchanged expression level of *UGPD* in early or late flowering switchgrass genotypes in response to different growth stages (Tornqvist et al., 2017). Our results supported that expressions of *UGPD* were not consistent across genotypes with higher level found in one early flowering (7071) and one late flowering genotype (1) of switchgrass.

Within 50 kb of a SNP on chromosome 2b related to heading date, *Pavir.Bb00118*, a homolog of serine/threonine-protein kinase WNK, was identified across Lafayette and DeKalb locations in 2016 (**Supplementary Table S4**). Protein kinases play important roles in controlling diverse cellular processes (Manning et al., 2002). In *Arabidopsis*, *AtWNK1*, *AtWNK2*, *AtWNK4*, and *AtWNK6* seem to be under the control of the circadian clock (Nakamichi et al., 2002). For instance, *AtWNK1* interacts with APRR3 and phosphorylates the APRR3 component of APRR1 /TOC1 in plants (Nakamichi et al., 2002). Further studies showed that *AtWNK2*, *AtWNK5*, and *AtWNK8* mutants caused early flowering, while in contrast, *AtWNK1* mutant delayed flowering time (Wang et al., 2008). In this study, gene *Pavir.Bb00118* is a homolog of *AT1G64625,* also named FEHLSTART (FST) and WNK10 in *Arabidopsis* (Li et al., 2015). The *fst-1* mutant had normal vegetative growth and floral organ development, but showed low fertility and shorter siliques with fewer seeds (Li et al., 2015). Higher expression of *WNK* in one early flowering genotype demonstrated that WNK could be involved in the early flowering response. The transcriptomic analysis indicated up- or down-regulated expression level of *WNK* in early or late flowering switchgrass genotypes (Tornqvist et al., 2017). These results suggest that expression of *WNK* varies with growth stage and genotype. In addition, the transcript levels of ELF4, TOC1, CO, and FT were altered in *AtWNK* mutants, indicating that *WNK* genes regulate flowering time by modulating the photoperiod pathway (Wang et al., 2008).

Within 10 kb of an identified SNP on chromosome 5b, *Pavir.Eb00237,* a homolog of FLOWERING LOCUS T (FT) related to heading date was identified at Lafayette location across 2015 and 2016 (**Table 4**). FT is a key component of the photoperiodic pathway. In *A. thaliana*, the photoperiodic pathway acts through FT to promote floral induction in response to day length (Andrés et al., 2015). Allelic variation in the FT gene was associated with flowering time in natural or seminatural populations of perennial ryegrass (*Lolium perenne*) (Skøt et al., 2011). Gene *Pavir.J05082*, a homolog of the major flowering time regulator FT, was associated with early flowering in switchgrass natural populations (Grabowski et al., 2017). However, the transcriptomic analysis indicated very low expression levels of FT genes in early and low flowering switchgrass genotypes (Tornqvist et al., 2017). Similarly, we were not able to detect expression of FT through qRT-PCR in early and late flowering samples. By blasting against *Arabidopsis* FT family, we found a total of 47 FT genes in switchgrass genome (**Supplementary Table S5**). Further studies could verify the role of these genes in regulating flowering time.

CONCLUSION

Plant flowering is regulated by a complex network of genetic and environmental signals. The work presented here elucidates the genetic mechanisms controlling heading and anthesis dates in four pseudo-F_2 populations (two pairs of reciprocal crosses) of switchgrass through GWAS. Five significant SNPs were detected and associated with heading or anthesis dates, and candidate genes for light signaling, reproductive structures, circadian clock rhythm, and flowering time were identified. The results indicated genetic complexities (i.e., multiple regions/components) related to floral development. Future research could verify gene function associated with heading and anthesis development, which has great potential to enhance breeding programs aimed at germplasm improvement of switchgrass.

AUTHOR CONTRIBUTIONS

MT and C-ET collected the phenotypic data. MT and XZ performed the statistical and GWAS analysis. JE and CB performed the genotyping using exome-capture sequencing technique and raw SNP calling. MT, GR, and MS conducted further SNP filtering. C-ET, XZ, PG, GR, MS, CB, RD, JM, JV, and MC participated in interpreting

the results and writing the manuscript. MC and YJ designed the experiments. MT and YJ led writing of the manuscript.

FUNDING

This research was supported by the US Department of Energy (DOE), Office of Biological and Environmental Research (BER), grant nos. DE-SC0010631 and DE-SC0008180, and the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494).

ACKNOWLEDGMENTS

The authors would like to thank Drs. Mingxi Liu and Zijian Sun and Ms. Yu Cui for assisting in planting in the field.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01250/full#supplementary-material

REFERENCES

Andrés, F., Romera-Branchat, M., Martínez-Gallegos, R., Patel, V., Schneeberger, K., Jang, S., et al. (2015). Floral induction in *Arabidopsis* by flowering locus T requires direct repression of blade-on-petiole genes by the homeodomain protein pennywise. *Plant Physiol.* 169, 2187–2199. doi: 10.1104/pp.15.00960

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., et al. (2012). Reference genome sequence of the model plant *Setaria*. *Nat. Biotechnol.* 30, 555–561. doi: 10.1038/nbt.2196

Blumel, M., Dally, N., and Jung, C. (2015). Flowering time regulation in crops - what did we learn from *Arabidopsis*? *Curr. Opin. Biotechnol.* 32, 121–129. doi: 10.1016/j.copbio.2014.11.023

Bowles, D., Isayenkova, J., Lim, E. K., and Poppenberger, B. (2005). Glycosyltransferases: managers of small molecules. *Curr. Opin. Plant Biol.* 8, 254–263. doi: 10.1016/j.pbi.2005.03.007

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

Buck-Sorlin, G. (2013). "Phytomer," in *Encyclopedia of Systems Biology*, eds W. Dubitzky, O. Wolkenhauer, K. H. Cho, and H. Yokota (New York, NY: Springer).

Campoli, C., Shtaya, M., Davis, S. J., and von Korff, M. (2012). Expression conservation within the circadian clock of a monocot: natural variation at barley Ppd-H1affects circadian expression of flowering time genes, but not clock orthologs. *BMC Plant Biol.* 12:97. doi: 10.1186/1471-2229-12-97

Casler, M. D. (2012). *Switchgrass Breeding, Genetics, and Genomics*. London: Springer. doi: 10.1007/978-1-4471-2903-5_2

Casler, M. D., and Boe, A. R. (2003). Cultivar × environment interactions in switchgrass. *Crop Sci.* 43, 2226–2233. doi: 10.2135/cropsci2003.2226

Casler, M. D., Vermerris, M., and Dixon, R. A. (2015). Replication concepts for bioenergy research experiments. *Bioenergy Res.* 8, 1–16. doi: 10.1007/s12155-015-9580-7

Casler, M. D., Vogel, K. P., Taliaferro, C. M., and Wynia, R. L. (2004). Latitudinal adaptation of switchgrass populations. *Crop Sci.* 44, 293–303. doi: 10.2135/cropsci2004.2930

Evans, J., Crisovan, E., Barry, K., Daum, C., Jenkins, J., Kunde-Ramamoorthy, G., et al. (2015). Diversity and population structure of northern switchgrass as revealed though exome capture sequencing. *Plant J.* 84, 800–815. doi: 10.1111/tpj.13041

Gimeno, J., Eattock, N., Van Deynze, A., and Blumwald, E. (2014). Selection and validation of reference genes for gene expression analysis in switchgrass (*Panicum virgatum*) using quantitative real-time RT-PCR. *PLoS One* 9:e91474. doi: 10.1371/journal.pone.0091474

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, 1178–1186. doi: 10.1093/nar/gkr944

Grabowski, P. P., Evans, J., Daum, C., Deshpande, S., Barry, K. W., Kennedy, M., et al. (2017). Genome-wide associations with flowering time in switchgrass using exome-capture sequencing data. *New Phytol.* 213, 154–169. doi: 10.1111/nph.14101

Helliwell, C. A., Wood, C. C., Robertson, M., James Peacock, W., and Dennis, E. S. (2006). The *Arabidopsis* FLC protein interacts directly in vivo with SOC1 and FT chromatin and is part of a high-molecular-weight protein complex. *Plant J.* 46, 183–192. doi: 10.1111/j.1365-313X.2006.02686.x

Johansson, M., and Staiger, D. (2015). Time to flower: interplay between photoperiod and the circadian clock. *J. Exp. Bot.* 66, 719–730. doi: 10.1093/jxb/eru441

Jones, P., and Vogt, T. (2001). Glycosyltransferases in secondary plant metabolism: tranquilizers and stimulant controllers. *Planta* 213, 164–174. doi: 10.1007/s004250000492

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

Khan, M. R. G., Ai, X. Y., and Zhang, J. Z. (2014). Genetic regulation of flowering time in annual and perennial plants. *Wiley Interdiscip. Rev. RNA.* 5, 347–359. doi: 10.1002/wrna.1215

Kim, D. H., Doyle, M. R., Sung, S., and Amasino, R. M. (2009). Vernalization: winter and the timing of flowering in plants. *Annu. Rev. Cell Dev. Biol.* 25, 277–299. doi: 10.1146/annurev.cellbio.042308.113411

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25. doi: 10.1186/gb-2009-10-3-r25

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., Dukowic-Schulze, S., Lindquist, I. E., Farmer, A. D., Kelly, B., Li, T., et al. (2015). The plant-specific protein FEHLSTART controls male meiotic entry, initializing meiotic synchronization in *Arabidopsis*. *Plant J.* 84, 659–671. doi: 10.1111/tpj.13026

Li, J., Li, Y., Chen, S., and An, L. (2010). Involvement of brassinosteroid signals in the floral-induction network of *Arabidopsis*. *J. Exp. Bot.* 61, 4221–4230. doi: 10.1093/jxb/erq241

Lim, E. K., and Bowles, D. J. (2004). A class of plant glycosyltransferases involved in cellular homeostasis. *EMBO J.* 23, 2915–2922. doi: 10.1038/sj.emboj.7600295

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

Manning, G., Whyte, D. B., Martinez, R., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science* 298, 1912–1934. doi: 10.1126/science.1075762

Martin, E. R., Kinnamon, D. D., Schmidt, M. A., Powell, E. H., Zuchner, S., and Morris, R. W. (2010). SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics* 26, 2803–2810. doi: 10.1093/bioinformatics/btq526

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Mauro-Herrera, M., Wang, X., Barbier, H., Brutnell, T. P., Devos, K. M., and Doust, A. N. (2013). Genetic control and comparative genomic analysis of flowering time in *Setaria* (Poaceae). *G3* 3, 283–295. doi: 10.1534/g3.112.005207

McMillan, C. (1965). Ecotypic differentiation within four North American prairie grasses. II. Behavioral variation within transplanted community fractions. *Am. J. Bot.* 57, 55–65. doi: 10.1002/j.1537-2197.1965.tb06757.x

Michaels, S. D., and Amasino, R. M. (2000). Memories of winter: vernalization and the competence to flower. *Plant Cell Environ.* 23, 1145–1153. doi: 10.1046/j.1365-3040.2000.00643.x

Nakamichi, N., Kita, M., Ito, S., Yamashino, T., and Mizuno, T. (2005). PSEUDO-RESPONSE REGULATORS, PRR9, PRR7 and PRR5, together play essential roles close to the circadian clock of *Arabidopsis thaliana*. *Plant Cell Physiol.* 46, 686–698. doi: 10.1093/pcp/pci086

Nakamichi, N., Kita, M., Niinuma, K., Ito, S., Yamashino, T., Mizoguchi, T., et al. (2007). *Arabidopsis* clock-associated pseudo-response regulators PRR9, PRR7 and PRR5 coordinately and positively regulate flowering time through the canonical CONSTANS-dependent photoperiodic pathway. *Plant Cell Physiol.* 48, 822–832. doi: 10.1093/pcp/pcm056

Nakamichi, N., Murakami-Kojima, M., Sato, E., Kishi, Y., Yamashino, T., and Mizuno, T. (2002). Compilation and characterization of a novel WNK family of protein kinases in *Arabidopsis thaliana* with reference to circadian rhythms. *Biosci. Biotech. Biochem.* 66, 2429–2436. doi: 10.1271/bbb.66.2429

Narasimhamoorthy, B., Saha, M. C., Swaller, T., and Bouton, J. H. (2008). Genetic diversity in switchgrass collections assessed by EST-SSR markers. *Bioenergy Res.* 1:136. doi: 10.1007/s12155-008-9011-0

Niinuma, K., Nakamichi, N., Miyata, K., Mizuno, T., Kamada, H., and Mizoguchi, T. (2008). Roles of *Arabidopsis* PSEUDO-RESPONSE REGULATOR (PRR) genes in the opposite controls of flowering time and organ elongation under long-day and continuous light conditions. *Plant Biotech.* 25, 165–172. doi: 10.5511/plantbiotechnology.25.165

Niu, L., Fu, C., Lin, H., Wolabu, T. W., Wu, Y., Wang, Z. Y., et al. (2016). Control of floral transition in the bioenergy crop switchgrass. *Plant Cell Environ.* 39, 2158–2171. doi: 10.1111/pce.12769

Nuñez, F. D. B., and Yamada, T. (2017). Molecular regulation of flowering time in grasses. *Agronomy* 7:17. doi: 10.3390/agronomy7010017

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

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5, 103–113. doi: 10.3835/plantgenome2012.06.0006

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Sato, E., Nakamichi, N., Yamashino, T., and Mizuno, T. (2002). Aberrant expression of the *Arabidopsis* circadian-regulated APRR5 gene belonging to the APRR1/TOC1 quintet results in early flowering and hypersensitiveness to light in early photomorphogenesis. *Plant Cell Physiol.* 43, 1374–1385. doi: 10.1093/pcp/pcf166

Skøt, L., Sanderson, R., Thomas, A., Skøt, K., Thorogood, D., Latypova, G., et al. (2011). Allelic variation in the perennial ryegrass FLOWERING LOCUS T gene is associated with changes in flowering time across a range of populations. *Plant Physiol.* 155, 1013–1022. doi: 10.1104/pp.110.169870

Song, Y. H., Ito, S., and Imaizumi, T. (2010). Similarities in the circadian clock and photoperiodism in plants. *Curr. Opin. Plant Biol.* 13, 594–603. doi: 10.1016/j.pbi.2010.05.004

Song, Y. H., Ito, S., and Imaizumi, T. (2013). Flowering time regulation: photoperiod- and temperature-sensing in leaves. *Trends Plant Sci.* 18, 575–583. doi: 10.1016/j.tplants.2013.05.003

Sukumaran, S., Li, X., Li, X., Zhu, C., Bai, G., Perumal, R., et al. (2016). QTL mapping for grain yield, flowering time, and stay-green traits in sorghum with genotyping-by-sequencing markers. *Crop Sci.* 56, 1429–1442. doi: 10.2135/cropsci2015.02.0097

Tanaka, K., Asami, T., Yoshida, S., Nakamura, Y., Matsuo, T., and Okamoto, S. (2005). Brassinosteroid homeostasis in *Arabidopsis* is ensured by feedback expressions of multiple genes involved in its metabolism. *Plant Physiol.* 138, 1117–1125. doi: 10.1104/pp.104.058040

Tornqvist, C.-E., Taylor, M., Jiang, Y., Evans, J., Buell, C. R., Kaeppler, S. M., et al. (2018). Quantitative trait locus mapping for flowering time in a lowland × upland switchgrass pseudo-F_2 population. *Plant Genome* 11:170093. doi: 10.3835/plantgenome2017.10.0093

Tornqvist, C.-E., Vaillancourt, B., Kim, J., Buell, C. R., Kaeppler, S. M., and Casler, M. D. (2017). Transcriptional analysis of flowering time in switchgrass. *Bioenergy Res.* 10, 700–713. doi: 10.1007/s12155-017-9832-9

Trevaskis, B., Bagnall, D. J., Ellis, M. H., Peacock, W. J., and Dennis, E. S. (2003). MADS box genes control vernalization-induced flowering in cereals. *Proc. Natl. Acad. Sci. U.S.A.* 100, 13099–13104. doi: 10.1073/pnas.1635053100

Van Esbroeck, G. A., Hussey, M. A., and Sanderson, M. A. (2003). Variation between Alamo and cave-in-rock switchgrass in response to photoperiod extension. *Crop Sci.* 43, 639–643.

Wang, B., Jin, S.-H., Hu, H.-Q., Sun, Y.-G., Wang, Y.-W., Han, P., et al. (2012). UGT87A2, an *Arabidopsis* glycosyltransferase, regulates flowering time via FLOWERING LOCUS C. *New Phytol.* 194, 666–675. doi: 10.1111/j.1469-8137.2012.04107.x

Wang, Y., Liu, K., Liao, H., Zhuang, C., Ma, H., and Yan, X. (2008). The plant WNK gene family and regulation of flowering time in *Arabidopsis*. *Plant Biol.* 10, 548–562. doi: 10.1111/j.1438-8677.2008.00072.x

Wang, Z., Devos, K. M., Liu, C., Wang, R., and Gale, M. D. (1998). Construction of RFLP-based maps of foxtail millet, *Setaria italica* (L.) P. Beauv. *Theor. Appl. Genet.* 96, 31–36. doi: 10.1007/s001220050705

Wickland, D. P., and Hanzawa, Y. (2015). The FLOWERING LOCUS T/TERMINAL FLOWER 1 gene family: functional evolution and molecular mechanisms. *Mol. Plant* 8, 1–15. doi: 10.1016/j.molp.2015.01.007

Woods, D. P., Bednarek, R., Bouché, F., Gordon, S. P., Vogel, J. P., Garvin, D. F., et al. (2017). Genetic architecture of flowering-time variation in *Brachypodium distachyon*. *Plant Physiol.* 173, 269–279. doi: 10.1104/pp.16.01178

Yamamoto, Y., Sato, E., Shimizu, T., Nakamich, N., Sato, S., Kato, T., et al. (2003). Comparative genetic studies on the APRR5 and APRR7 genes belonging to the APRR1/TOC1 quintet implicated in circadian rhythm, control of flowering time, and early photomorphogenesis. *Plant Cell Physiol.* 44, 1119–1130. doi: 10.1093/pcp/pcg148

Yan, L., Fu, D., Li, C., Blechl, A., Tranquilli, G., Bonafede, M., et al. (2006). The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc. Natl. Acad. Sci. U.S.A.* 103, 19581–19586. doi: 10.1073/pnas.0607142103

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 September 2018
doi: 10.3389/fpls.2018.01272

Genome-Wide Association Mapping of Seedling Heat Tolerance in Winter Wheat

*Frank Maulana[1], Habtamu Ayalew[1], Joshua D. Anderson[1], Tadele T. Kumssa[1], Wangqi Huang[1,2] and Xue-Feng Ma[1]**

[1] Noble Research Institute, Ardmore, OK, United States, [2] Institute of Food Crops, Yunnan Academy of Agricultural Sciences, Kunming, China

Heat stress during the seedling stage of early-planted winter wheat (*Triticum aestivum* L.) is one of the most abiotic stresses of the crop restricting forage and grain production in the Southern Plains of the United States. To map quantitative trait loci (QTLs) and identify single-nucleotide polymorphism (SNP) markers associated with seedling heat tolerance, a genome-wide association mapping study (GWAS) was conducted using 200 diverse representative lines of the hard red winter wheat association mapping panel, which was established by the Triticeae Coordinated Agricultural Project (TCAP) and genotyped with the wheat iSelect 90K SNP array. The plants were initially planted under optimal temperature conditions in two growth chambers. At the three-leaf stage, one chamber was set to 40/35°C day/night as heat stress treatment, while the other chamber was kept at optimal temperature (25/20°C day/night) as control for 14 days. Data were collected on leaf chlorophyll content, shoot length, number of leaves per seedling, and seedling recovery after removal of heat stress treatment. Phenotypic variability for seedling heat tolerance among wheat lines was observed in this study. Using the mixed linear model (MLM), we detected multiple significant QTLs for seedling heat tolerance on different chromosomes. Some of the QTLs were detected on chromosomes that were previously reported to harbor QTLs for heat tolerance during the flowering stage of wheat. These results suggest that some heat tolerance QTLs are effective from the seedling to reproductive stages in wheat. However, new QTLs that have never been reported at the reproductive stage were found responding to seedling heat stress in the present study. Candidate gene analysis revealed high sequence similarities of some significant loci with candidate genes involved in plant stress responses including heat, drought, and salt stress. This study provides valuable information about the genetic basis of seedling heat tolerance in wheat. To the best of our knowledge, this is the first GWAS to map QTLs associated with seedling heat tolerance targeting early planting of dual-purpose winter wheat. The SNP markers identified in this study will be used for marker-assisted selection (MAS) of seedling heat tolerance during dual-purpose wheat breeding.

Keywords: wheat, GWAS, QTL, heat stress, seedling stress

OPEN ACCESS

Edited by:
Yiwei Jiang,
Purdue University, United States

Reviewed by:
Xusheng Wang,
St. Jude Children's Research Hospital, United States
Jiasheng Wu,
Zhejiang Agriculture & Forestry University, China

****Correspondence:***
Xue-Feng Ma
xma@noble.org

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *16 May 2018*
Accepted: *14 August 2018*
Published: *04 September 2018*

Citation:
Maulana F, Ayalew H, Anderson JD, Kumssa TT, Huang W and Ma X-F (2018) Genome-Wide Association Mapping of Seedling Heat Tolerance in Winter Wheat. Front. Plant Sci. 9:1272. doi: 10.3389/fpls.2018.01272

INTRODUCTION

Wheat (*Triticum aestivum* L.) is one of the most important feed and food crops in the world and it covers more cultivable land globally than any other crop. Moreover, it provides food for 36% of the world's population (Cossani and Reynolds, 2012; Prerna et al., 2013; Kim and Anderson, 2015). In the southern Great Plains of the United States, including Oklahoma and Texas, dual-purpose wheat grown for cool season grazing is seeded at least 2–3 weeks earlier than wheat grown for grain only to increase fall to winter forage production. However, early planting in the fall often coincides with high temperatures that affect seed germination, seedling growth, and development, eventually resulting in reduced forage and grain yield. As the global climate continues to change, the severity and frequency of high temperature stress is likely to increase, thereby resulting in reduction of productivity of important crops including wheat. Climate predictions show that, by the end of the 21st century, the average global temperature is expected to increase by 1–4°C (Driedonks et al., 2016). Therefore, development of dual-purpose wheat cultivars with tolerance to heat stress during the seedling stage is crucial for early planting in the region.

Several studies have outlined the effects of heat stress on plant morphological, physiological, and biochemical processes at various growth stages of wheat (Cossani and Reynolds, 2012; Paliwal et al., 2012; Feng et al., 2014; Chaturvedi et al., 2017). The effects of heat stress during the seedling stage include reduction of photosynthesis, chlorophyll content, respiration rate, and death of the seedlings due to excessive dehydration of leaves beyond the permanent wilting point (Ristic et al., 2007; Cossani and Reynolds, 2012). Research findings in the past indicated that heat stress causes swelling of the thylakoid membrane and malfunction of photosystem II involved in photosynthetic activity (Ristic et al., 2007; Talukder et al., 2014). Chlorophyll is harbored in the thylakoid membrane and when this membrane is damaged by the stress, chlorophyll content is reduced (Ristic et al., 2008). However, phenotypic variability for heat tolerance among genotypes has been studied during the reproductive stage but limited information is available for the seedling stage.

High temperature stress at the grain filling stage has been reported to reduce the yield and quality of wheat (Wardlaw et al., 2002; Schapendonk et al., 2007; Stratonovitch and Semenov, 2015), sorghum [*Sorghum bicolor* (L.) Moench] (Prasad et al., 2008), maize (*Zea mays* L.) (Yang et al., 2015), and rice (*Oryza sativa* L.) (Shi et al., 2016). Heat stress has been found to reduce wheat yield by 33.6% and by more than 50% (Chatrath et al., 2007; Joshi et al., 2007). The yield potential of wheat is rarely attained, particularly when moderate heat stress occurs and alternates with periodic extreme heat stress (Mason et al., 2011). Heat tolerance is a polygenic trait that is controlled by many genes with minor effects on the phenotype. Therefore, selection of heat stress tolerance under field conditions is very challenging because of its genetic complexity, weather variability and the influence of genotype-by-environment interaction effect. In this regard, identification of QTLs and molecular markers associated with tolerance to heat stress is crucial for improving breeding efficiency using marker-assisted selection (MAS).

To date, dissection of QTLs for heat tolerance in wheat has been mainly conducted during the grain filling stage using bi-parental mapping populations (Mason et al., 2010; Vijayalakshmi et al., 2010; Talukder et al., 2014). These studies identified various major and minor QTLs for vegetative and reproductive stage traits on different wheat chromosomes. For example, five QTLs for heat tolerance in wheat were detected on chromosomes 1B, 1D, 2B, 6A, and 7A (Talukder et al., 2014). Similarly, two QTLs for heat tolerance have been detected on chromosomes 2B and 5B in a spring wheat mapping population (Butler, 2002). Again in other studies, QTLs for heat tolerance during the grain filling stage have been found on several chromosomal regions including 1A, 1B, 2B, 3B, 5A, and 6D (Mason et al., 2010). In addition, QTLs associated with yield components and physiological traits, such as stay green and senescence of wheat, were found on chromosomes 2A, 3A, 4A, 6A, 6B, and 7A (Vijayalakshmi et al., 2010).

Moreover, using a meta-analysis strategy, major QTLs associated with heat tolerance were detected on chromosomes 1B, 2B, 2D, 4A, 4D, 5A, and 7A (Acuña-Galindo et al., 2015). Similarly, another significant locus on chromosome 3B, associated with the heat susceptibility index of yield components, was identified using a bi-parental mapping population (Mason et al., 2010). Although linkage mapping using bi-parental mapping populations has successfully identified heat tolerance QTLs, it requires a large amount of resources and time to develop mapping populations such as recombinant inbred lines (RILs). In addition, it relies on recent recombination resulting in low mapping resolution and only alleles differing in the parents are considered with this approach. On the other hand, with GWAS, diverse individuals are used without developing new mapping populations, making it less expensive. Historic recombination events existing in the population are leveraged in GWAS, thereby resulting in high mapping resolution compared with linkage mapping. However, GWAS requires higher marker density than traditional QTL mapping because linkage disequilibrium (LD) is in general much lower in a GWAS population than in a bi-parental population.

The GWAS approach has been used to discover genes controlling both polygenic and monogenic traits. For example, QTLs associated with important traits such as disease resistance (Tadesse et al., 2015; Arruda et al., 2016; Liu et al., 2017), yield, and grain quality traits (Tadesse et al., 2015) in wheat have been discovered with GWAS. QTLs associated with heat tolerance in wheat have also been detected using GWAS (Mondal et al., 2015).

Although heat tolerance during the reproductive stage of wheat has been well characterized, heat stress during the seedling stage is not studied. Therefore, the objectives of this study were: (1) to map QTLs associated with seedling heat tolerance in wheat and (2) to identify SNP markers for MAS of seedling heat tolerance during dual-purpose wheat breeding in the southern Great Plains of the United States.

MATERIALS AND METHODS

Genetic Materials and Phenotyping

A set of 200 lines, selected based on genetic diversities from a hard red winter wheat association mapping panel consisting of 299 wheat lines from the Triticeae Coordinated Agricultural Project (TCAP[1]), was used in this study. The association mapping panel is composed of representative winter wheat lines across the Great Plains (Grogan et al., 2016). The experiments were conducted in two growth chambers. A high-temperature treatment (40/35°C day/night) to mimic heat stress was induced at the three-leaf stage for 14 days in one chamber, and an optimal-temperature treatment (25/20°C day/night) was used as a control in the other chamber. Photoperiod and light intensity in both growth chambers were set at 16 h and 400 μml m^{-2} s^{-1}, respectively. The plants were planted in 72-well flat trays in a randomized complete block design with three biological replicates of each line. The trays were randomly arranged and periodically moved around to avoid positional effect. Throughout the experimentation, plants were watered as needed in both growth chambers to ensure no drought stress.

Data were collected on leaf chlorophyll content, shoot length, number of leaves per seedling, and seedling recovery. Leaf chlorophyll content was measured using a self-calibrating SPAD chlorophyll meter (Model 502, Spectrum Technologies, Plainfield, IL, United States). Three measurements of leaf chlorophyll content were taken per line, and the average was used for statistical analysis. Shoot length was measured from the soil surface to the tip of the longest leaf. Leaf chlorophyll content and shoot length were measured 10 days after heat stress treatment. Number of leaves per seedling was recorded as the average number of leaves counted from three seedlings, 14 days after the seedlings were exposed to heat stress. Seedling recovery was the percentage of seedlings that were able to recover 7 days after removal of heat stress treatment. Heat stress response, referred to as trait relative difference (TRD), was calculated as the difference between trait performance at optimal and high temperatures, and then divided by performance at optimal temperature. The experiment was repeated six times (i.e., six runs) using the same two chambers.

Phenotypic Data Analysis

Analysis of variance of the phenotypic data was performed using the Statistical Analysis System (SAS) software V9.3 (SAS Institute, 2011) to assess the effects of genotype, run, and genotype-by-run interaction. All sources of variation were considered as random effects. All other variances besides genotype and experimental run were pooled as residuals.

SNP Genotyping

The wheat lines were genotyped using the wheat iSelect 90K SNP genotyping array (Wang et al., 2014; Guttieri et al., 2017), which generated 21,555 SNPs. After SNPs with minor allele frequency (MAF) of less than 5% and missing data of more than 10% were filtered out, a total of 15,574 SNPs remained and were used for analysis. The genetic positions of the SNP markers used in this study were based on the consensus map developed using eight wheat mapping populations (Wang et al., 2014).

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

Population Structure, Kinship, and Linkage Disequilibrium Analyses

The genetic structure of the panel was assessed using the STRUCTURE program, principal component analysis (PCA), and neighbor-joining (NJ) tree analysis. The STRUCTURE program version 2.3.4 (Falush et al., 2003) was used to estimate the number of groups (K) and the membership coefficients. A model-based Bayesian clustering approach was performed, where the number of assumed groups was set from k = 1 to 10. During STRUCTURE analysis, a Markov chain Monte Carlo (MCMC) of 15,000 burn-in replicates followed by 15,000 iterations was run and repeated five times using an admixture model. Due to lots of admixtures in the panel, the STRUCTURE results were verified by comparing the results to other analyses. The optimal number of groups in this panel was determined based on the point where the posterior probability [LnP(D)] began to plateau from the STRUCTURE analysis (Casa et al., 2008) and the NJ tree analysis. The principal components (PCs) were calculated using the R function *princomp*, while the NJ tree analysis was performed in TASSEL version 5.2.28 (Bradbury et al., 2007). To determine the number of PCs to use in clustering and GWAS analysis, a scree plot was generated by plotting the percentage of variances explained by the first 10 PCs against the number of PCs. Based on this, the optimal number of PCs (where the "*elbow*" point occurred) was selected. The analysis of K between lines was performed following the identity-by-state method (Endelman and Jannink, 2012).

Linkage disequilibrium among pairs of SNP markers was performed with the TASSEL software using 3,484 tag SNPs selected using the R package SNPRelate (Zheng, 2013). LD for within and across the three wheat genomes (A, B, and D) was estimated as a squared allele frequency correlation (r^2) between SNP marker pairs. All SNP marker pairs with p-values of less than 0.001 were considered to be in significant LD. LD decay distance was estimated by plotting the scatterplot of LD r^2 values between marker pairs and the genetic distance (in cM) using the R package SNPRelate (Zheng, 2013), while the trend line was fitted by second-degree LOESS (Cleveland, 1979). To determine whether significant SNP markers associated with the trait on each chromosome were in LD with the highest $-\log_{10}$(p-value) SNP hit, LD analysis was performed on every chromosome where significant QTLs were detected.

Genome-Wide Association Mapping Analysis

Genome-wide association mapping was performed with the Genome Association and Prediction Integrated Tool (GAPIT) (Lipka et al., 2012). For the Q model, three PCs that were selected based on scree plot generated from PCA were included in the model as fixed-effect covariate (Zhao et al., 2007) to correct for population structure. In the K model, the K matrix between individuals was calculated and included in the model as

random-effect covariate. For mixed linear model (MLM), both the population structure (PCs) and K matrix were included in the model as fixed and random-effect covariates, respectively (Yu et al., 2006).

For the Q model, the following equation was used:

$$Y = X\beta + e$$

Y is the vector of phenotypic values, X is the design matrix, β is the vector consisting of SNP markers and population structure (PCs) included in the model as fixed effects, and e is the random error.

For the K and MLM models, the following equation was used:

$$Y = X\beta + Z\mu + e,$$

where Z is the design matrix and μ is the vector comprising additive genetic effects considered as random. In the K-model, β contains only markers and μ contains the K-matrix, while in the MLM, β has both markers and population structure (PCs), and μ has the K matrix. Significant QTLs were initially tested based on a false discovery rate (FDR)-adjusted p-value of 0.05 following a step-wise procedure (Benjamini and Hochberg, 1995), which is very stringent (Müller et al., 2011). However, a lower threshold, unadjusted significance p-value <0.001, was eventually used to declare significance since the FDR is too stringent in the current study. Visualization of the significant QTLs and SNPs was done using Manhattan plots, generated using the R package *qqman* (Turner, 2018).

Candidate Gene Analysis

A BLAST search was performed against the newly released wheat reference sequence hosted by the URGI-INRA[2] and the National Center for Biotechnology Information (NCBI) database to identify candidate genes or related proteins with DNA sequences similar to the SNPs significantly associated with seedling heat tolerance-related traits detected in this study.

RESULTS

Phenotypic Data Analysis

Phenotypic variation was observed among genotypes for all traits in both temperature regimes (**Table 1**). Frequency distribution of the lines for the investigated traits at optimal and heat-stressed growth conditions are presented in **Figure 1**. Mean leaf chlorophyll content at optimal temperature was 38.3 with a range from 31.8 to 44.9, while for heat-stressed plants, mean leaf chlorophyll content was 26.7, ranging from 17.0 to 37.1. At optimal temperature, mean shoot length was 44.9 cm, ranging from 35.0 to 56.5 cm, whereas at heat-stressed growth condition, the mean value was 33.8 cm, and the range was from 23.5 to 44.4 cm. Mean number of leaves per seedling was six at optimal temperature compared with four at heat-stressed growth condition. For the number of leaves per seedling, phenotypic variation among lines was very small as shown in **Figure 1** because almost all plants were at three-leaf stage when the experiment started. As a result, variation in number of leaves per seedling among lines was very small by the end of 14-day temperature treatment. As for seedling recovery, on average, 52.3% of seedlings were able to recover after the removal of heat stress treatment (**Table 1**). Overall, heat stress reduced leaf chlorophyll content, shoot length and number of leaves per seedling by 30.3, 25.0, and 32.2%, respectively.

Population Structure Analysis

Three different clustering methods, PCA, NJ tree analysis, and STRUCTURE analysis, were compared to assess their agreement in the pattern of structuring of this panel. PCA divided the panel into four main groups with lots of admixture (**Figure 2**). However, the PCA revealed that the population structure in this panel is very low since the first three PCs collectively explained only about 19.4% of the total variance. The first principal component (PC1) explained about 9.4%, while the second (PC2) and the third (PC3) explained about 6.2 and 3.8% of the total variance, respectively (**Figure 2**). According to the NJ tree analysis, this panel can also be divided into four major groups (G1, G2, G3, and G4), based mainly on geographic origins and pedigree information (**Supplementary Figure S1A**). For example, in the first main group (G1), majority of the lines were from the Oklahoma State University and the Texas A&M University. Most lines with a common parent in their pedigree tended to cluster into the same group. For example, the majority of the lines assigned to G1 had "Jagger" as one of the parents in their pedigree. The largest number of lines forming G2 originated from the University of Nebraska breeding program, followed by the Kansas State University and the Colorado State University. Group G3 was dominated by wheat lines from the AgriPro Syngenta followed by those from the University of Nebraska

[2]https://urgi.versailles.inra.fr/

TABLE 1 | Seedling trait performance of the winter wheat lines under optimal and high temperatures in the present study.

Trait	Optimal temperature (25/20°C)			Heat stress (40/35°C)		
	Mean	Range	SD	Mean	Range	SD
Leaf chlorophyll content (SPAD)	38.30	31.82–44.89	2.60	26.70	17.03–37.14	3.85
Shoot length (cm)	44.87	35.00–56.50	4.28	33.83	23.50–44.38	3.64
Number of leaves per seedling	6	3–9	1.14	4	3–5	0.25
Seedling recovery (%)	N/A	N/A	N/A	52.32	6.25–89.59	18.48

SD, standard deviation; N/A, not applicable.

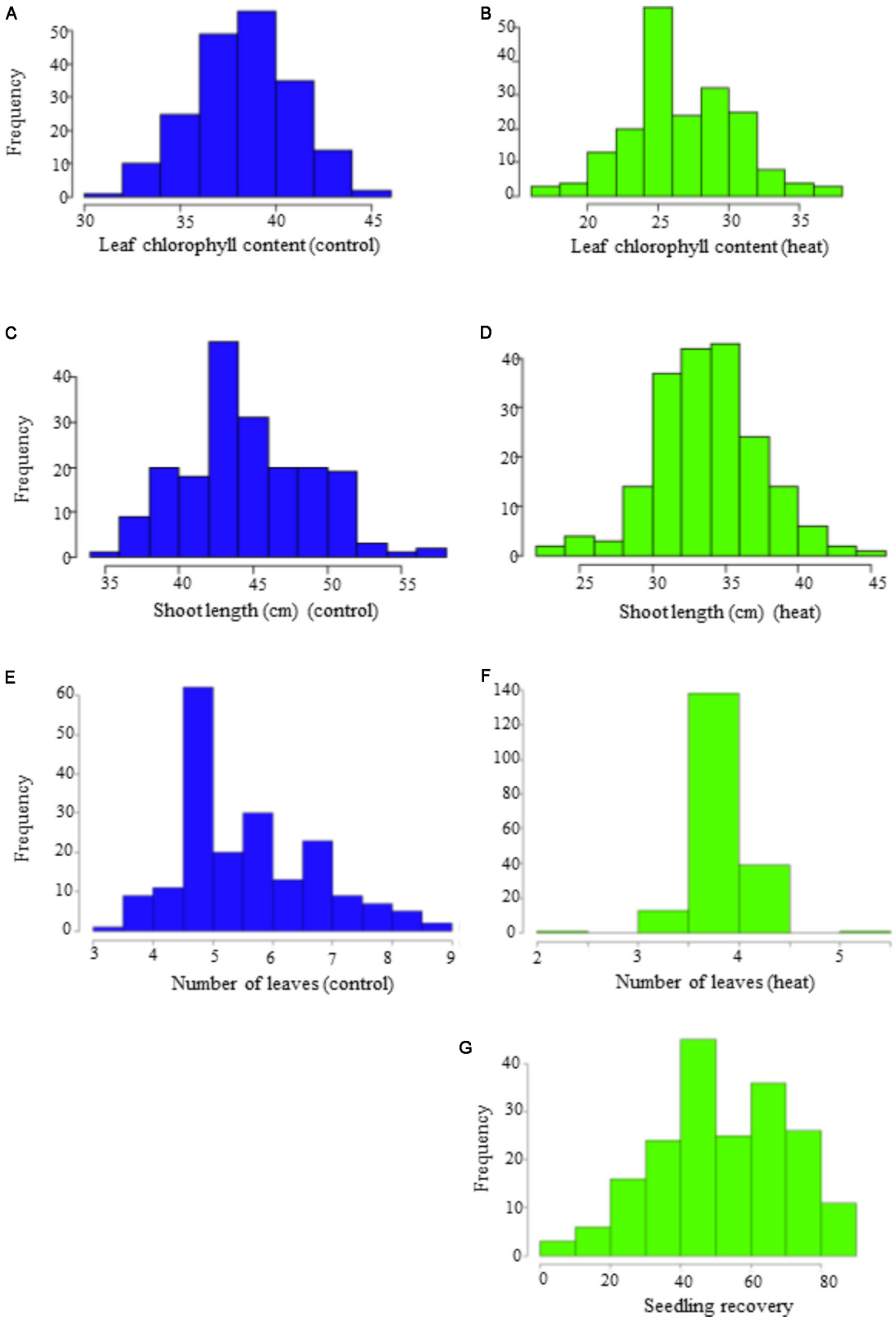

FIGURE 1 | Frequency distribution of the seedling traits observed at optimal (OT) and heat-stressed (HS) growth conditions in the panel. **(A,B)** Leaf chlorophyll content at optimum and heat-stressed growth conditions; **(C,D)** shoot length (cm) at optimum and heat-stressed growth conditions; **(E,F)** number of leaves at optimum and heat-stressed growth conditions; **(G)** seedling recovery after removal of the heat stress.

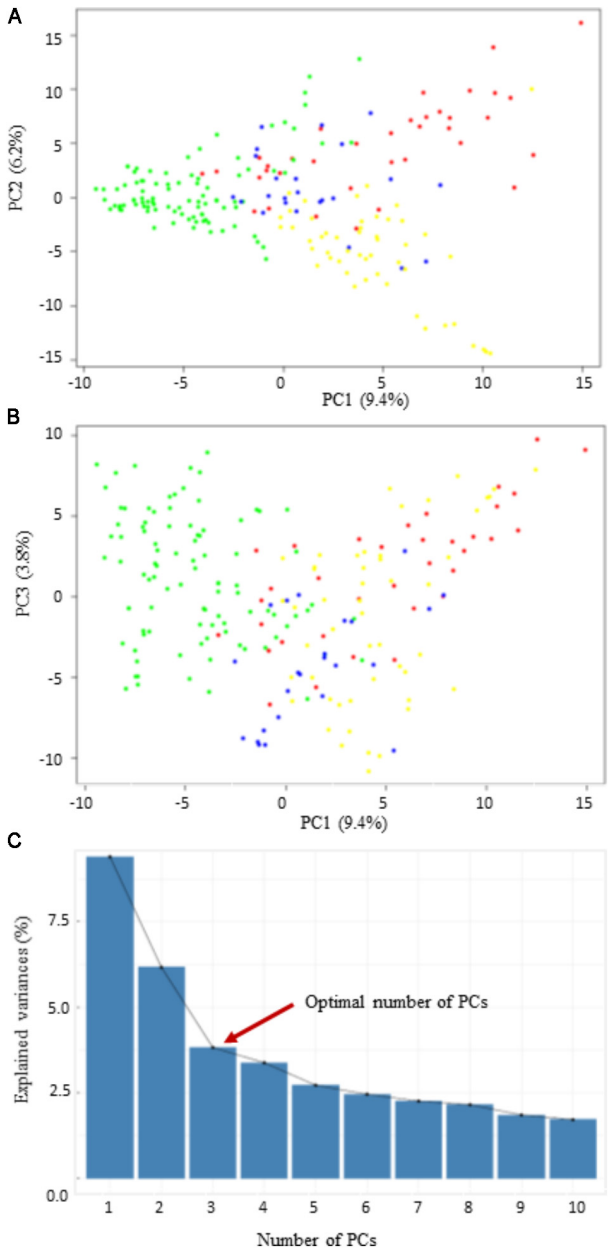

FIGURE 2 | Scatter plots of the first three principal components (PCs). **(A,B)** The distribution of the 200 winter wheat lines in PC1 vs. PC2, and PC1 vs. PC3; **(C)** the scree plot shows the variance explained by the first 10 PCs.

wheat breeding program. Finally, the largest number of lines in G4 came from the Texas A&M University, followed by those from the Oklahoma State University.

The STRUCTURE program also stratified panel into four groups but with a lot of admixtures (**Supplementary Figure S1B**). The lack of a distinct clustering pattern observed in this panel is because there is a high degree of relatedness among lines included in this study due to sharing genetic materials among wheat breeding programs. For GWAS analysis, we used the three PCs from the PCA as a fixed-effect covariate in the Q and MLM to correct for population structure.

Linkage Disequilibrium Analysis

After filtering using the R package SNPRelate (Zheng, 2013), 3,484 tag SNPs were obtained for LD analysis. The majority of SNP markers were distributed across the wheat A and B genomes with 41% (1439 SNPs) and 45% (1568 SNPs), respectively, while the D genome had the lowest number (343) of SNPs (10%). In addition, the B genome had the highest number of SNP markers per cM (7.19), seconded by A genome (5.97) and D genome (1.53). On A genome, 29.5% of SNP marker pairs were in significant LD ($p < 0.001$), while on B and D genomes, 32.8 and 14.0% of SNP marker pairs were in significant LD (**Supplementary Table S1**). The scatter plots of the allele frequency correlations (r^2) between the SNP marker pairs and the genetic distance (in cM) within each of the three wheat genomes (A, B, and D) are presented in **Supplementary Figure S2**. The data showed that LD decayed to <0.1 at 9.7 cM in A genome, 9.8 cM in B genome, and 10.9 cM in D genome.

Genome-Wide Association Mapping Analysis

Compared to Q and K models, MLM has high statistical power for controlling false positives. Therefore, in this study MLM was chosen as the appropriate model for reporting QTL mapping results. The quantile–quantile (Q–Q) plots of *p*-values comparing the uniform distribution of the expected $-\log_{10}(p)$ to the observed $-\log_{10}(p)$ of all evaluated traits are presented as **Supplementary Figure S3**. Genome-wide association mapping analysis results for all traits using the MLM are presented in **Figures 3–6**. The QTLs and the SNP markers significantly associated with seedling traits at optimal and heat-stressed growth conditions, as well as heat stress responses of all traits are presented in **Supplementary Table S2**. Although, no QTLs were declared significant at a FDR of 0.05, some SNPs were significant at unadjusted significance *p*-value <0.001 at optimal and/or heat-stressed growth conditions.

For leaf chlorophyll content at the optimal temperature, five QTLs, represented by 15 SNPs, were detected significant based on unadjusted significance *p*-value <0.001 in chromosomes 1B, 2B, 3B, 5B, and 6B (**Figure 3A** and **Supplementary Table S2**). The first QTL (*QLCCOT.nri-1B*) region was represented by six SNPs, which were mapped within genetic distance of 78–82 cM on chromosome 1B, and together accounted for 42.9% of the total phenotypic variation in leaf chlorophyll content at the optimal temperature. The second QTL region (*QLCCOT.nri-2B*), represented by four SNPs, was mapped at the genetic position of 119 cM on chromosome 2B. The four markers together explained 23.3% of the phenotypic variation in leaf chlorophyll content. On chromosome 5B, one QTL (*QLCCOT.nri-5B*) was mapped at 171–184 cM, which explained 18.6% of the phenotypic variation. On chromosomes 3B and 6B, two QTLs, *QLCCOT.nri-3B* (124 cM) and *QLCCOT.nri-6B* (121 cM) were detected collectively accounted for 11.7% of the phenotypic variation. Overall, the most significant SNPs for the trait were IWB9175

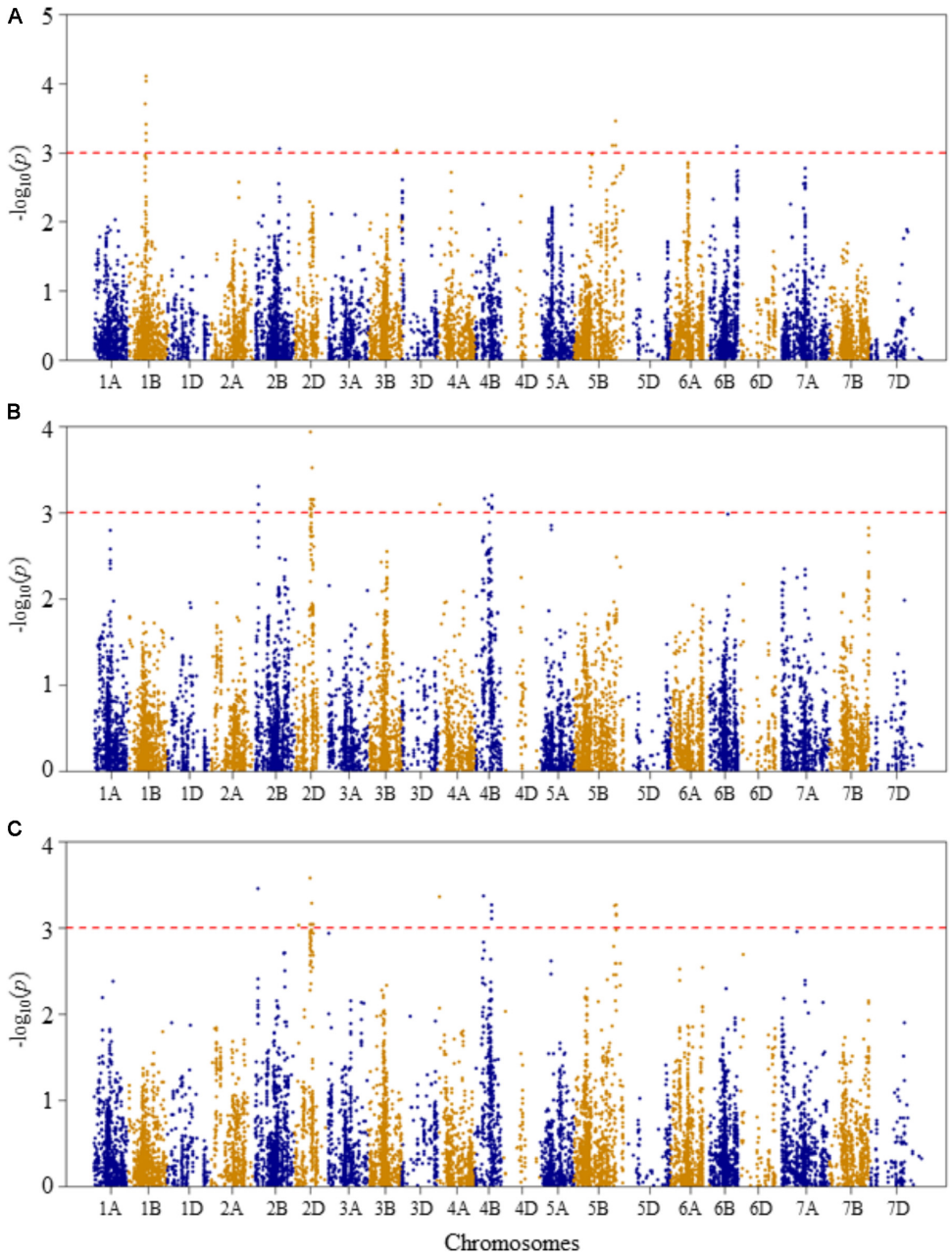

FIGURE 3 | Manhattan plots of GWAS conducted on leaf chlorophyll content of the association mapping panel. **(A)** Optimal temperature; **(B)** heat stressed temperature; and **(C)** heat stress response using the trait relative difference between the two temperature treatments.

(80 cM), IWB14950 (80 cM), and IWB27292 (78 cM) on chromosome 1B, which collectively explained about 23.8% of the total phenotypic variation in leaf chlorophyll content under optimum growth temperature.

For leaf chlorophyll content at the heat-stressed growth condition, six QTLs were detected on chromosomes 2B, 2D, 4A, and 4B (**Figure 3B** and **Supplementary Table S2**). The first QTL (*QLCCHS.nri-2B*) was located on chromosome 2B, and explained 12.1% of the phenotypic variation of the trait. On chromosome 2D, one QTL, *QLCCHS.nri-2D* (71–86 cM) was detected. This QTL was represented by 37 SNPs, explaining phenotypic variation in leaf chlorophyll content at heat-stressed growth condition ranging from 5.7 to 7.8%. On chromosome 4A, one QTL (*QLCCHS.nri-4A*) was found and mapped at 9 cM. This QTL explained about 5.8% of the phenotypic variation. In addition, three QTLs (*QLCCHS.nri-4B.1*, *QLCCHS.nri-4B.2*, and *QLCCHS.nri-4B.3*) were detected at 42, 60, and 76 cM on chromosome 4B, respectively. The phenotypic variation explained by these QTLs ranged from 5.8 to 17.5%. The most significant SNP markers associated with leaf chlorophyll

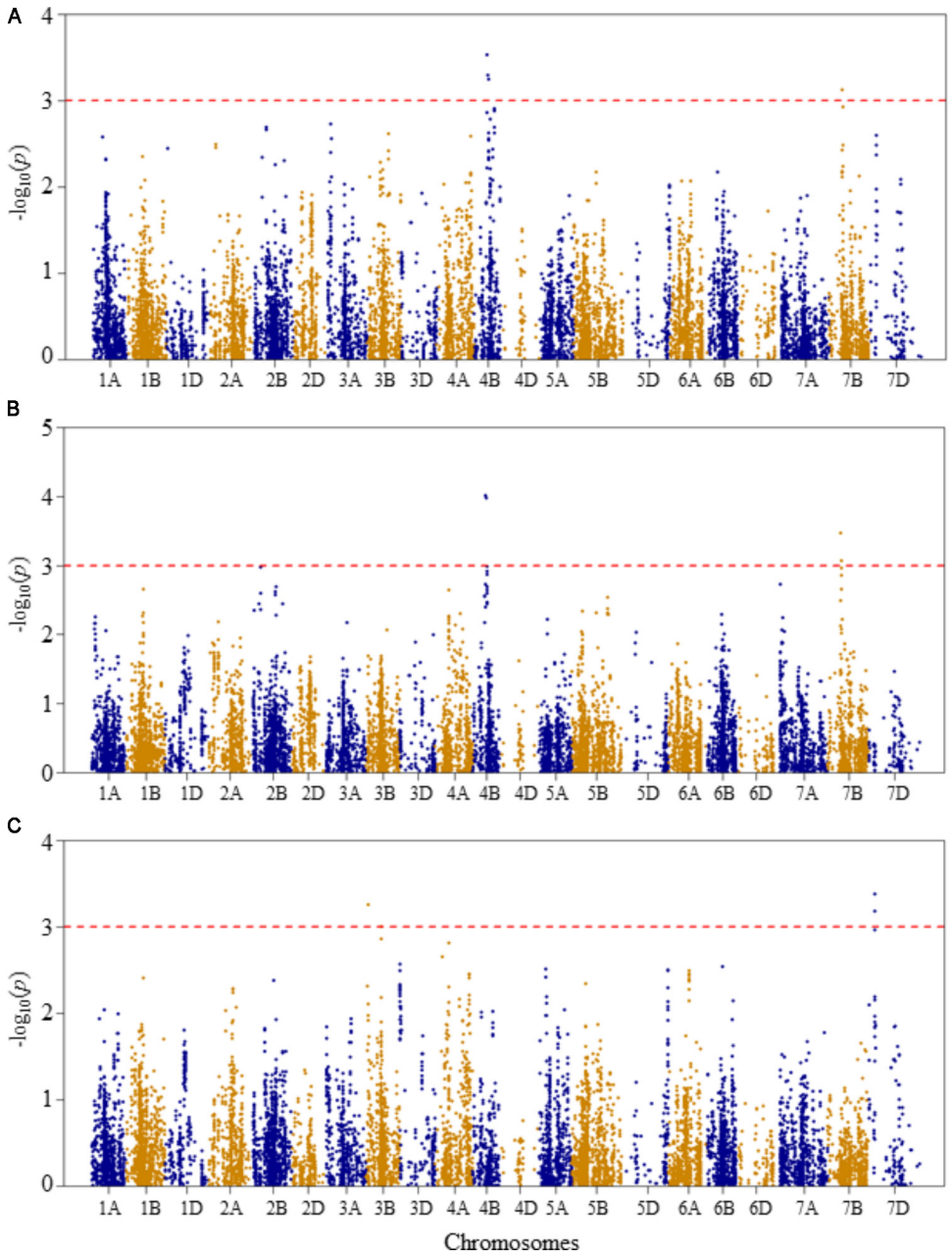

FIGURE 4 | Manhattan plots of GWAS conducted on shoot length of the association mapping panel. **(A)** Optimal temperature; **(B)** heat stressed temperature; and **(C)** heat stress response using the trait relative difference between the two temperature treatments.

content under heat-stressed growth condition were IWB28109 (71 cM) and IWB65632 (77 cM) on chromosome 2D, and IWB55435 (27 cM) on chromosome 2B (**Supplementary Table S2**). These three SNP markers together accounted for 21.4% of the phenotypic variation of leaf chlorophyll content at heat-stressed growth condition.

For heat stress response of the leaf chlorophyll content, i.e., the relative difference under the two growth temperatures, seven QTLs were identified on chromosomes 2B, 2D, 4A, 4B, and 5B (**Figure 3C** and **Supplementary Table S2**). The QTLs were represented by 39 SNPs significantly associated with heat stress response. A single QTL (*QLCCHR.nri-2B*) was detected on chromosome 2B, and it was mapped at a genetic position of 27 cM. The phenotypic variation explained by this QTL was 6.8%. On chromosome 2D, two QTLs: *QLCCHS.nri-2D.1* (22 cM) and *QLCCHS.nri-2D.2* (71–85 cM) were identified. The two QTLs on 2D were represented by 29 SNPs, which accounted for 5.8–7.1% of the total phenotypic variation in heat stress response of leaf chlorophyll content. Furthermore, one QTL was mapped at 9 cM on chromosome 4A, and it accounted for 6.6% of the

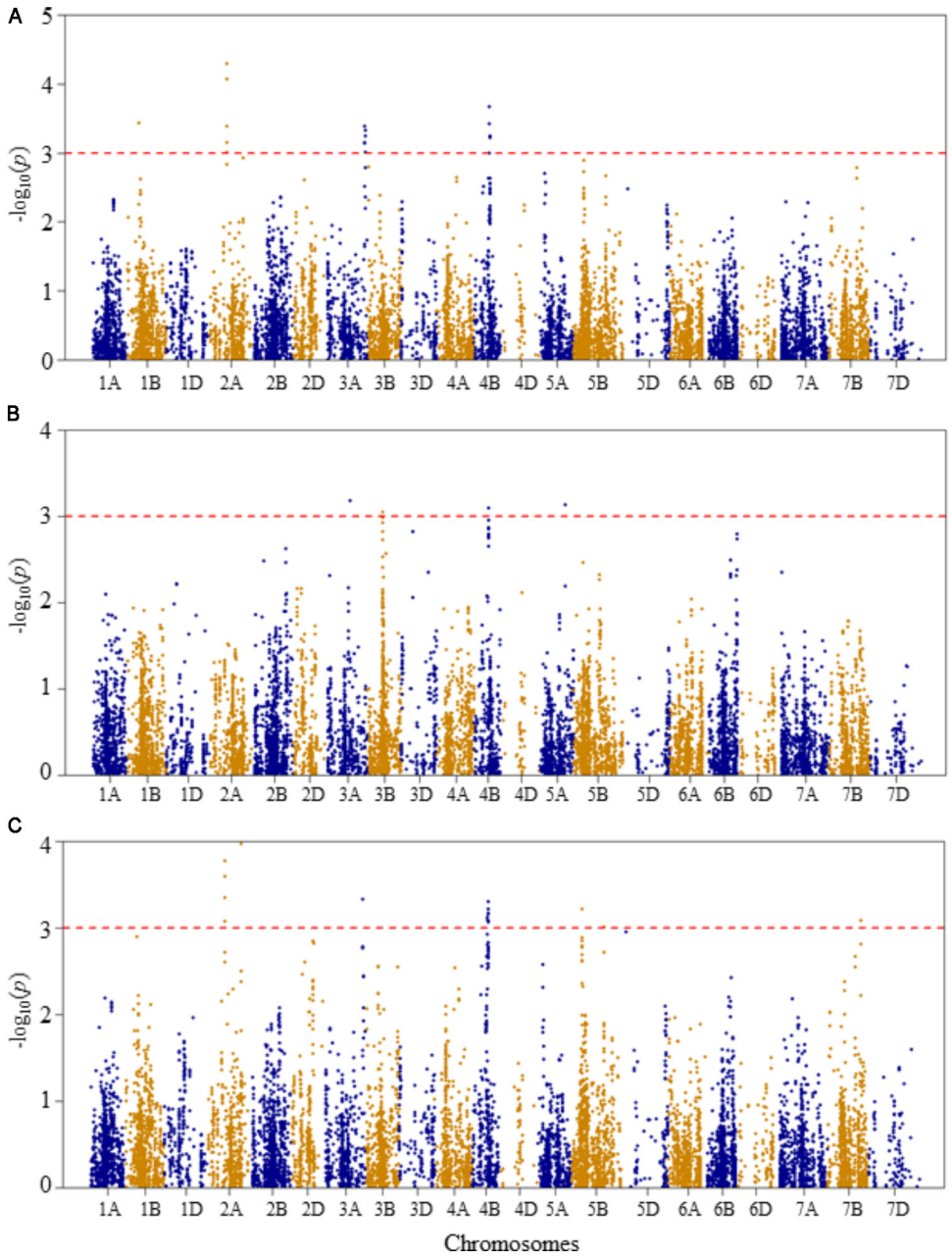

FIGURE 5 | Manhattan plots of GWAS conducted on number of leaves per seedling of the association mapping panel. **(A)** Optimal temperature; **(B)** heat stressed temperature; and **(C)** heat stress response using the trait relative difference between the two temperature treatments.

phenotypic variation in heat stress response of leaf chlorophyll content. On chromosome 4B, two QTLs (*QLCCHR.nri-4B.1* and *QLCCHR.nri-4B.2*) were detected at genetic positions of 40 and 76 cM, respectively. Similarly, on chromosome 5B, one QTL (*QLCCHR.nri-5B*) was mapped at 182–189 cM. The QTL on 5B was represented by four SNPs which together explained about 25.1% of total phenotypic variation in heat stress response of the trait (**Figure 3C** and **Supplementary Table S2**). The most significant SNPs were IWB28109 at 71 cM on 2D, IWB55435 at 27 cM on 2B and IWB48055 at 40 cM on 4B. These SNP markers accounted for 6.6–7.1% of the phenotypic variation in heat stress response of leaf chlorophyll content.

Overall, the data suggest that the leaf chlorophyll content QTLs associated with heat stress or heat response are located on chromosomes 2B, 2D, 4A, 4B, and 5B based on the QTLs detected for heat response of the trait, or the QTLs detected under heat-stressed but not under the optimum condition (**Table 2**).

For shoot length at the optimal growth temperature, two QTLs represented by four SNPs were detected significant at unadjusted *p*-value <0.001 (**Figure 4A** and **Supplementary Table S2**).

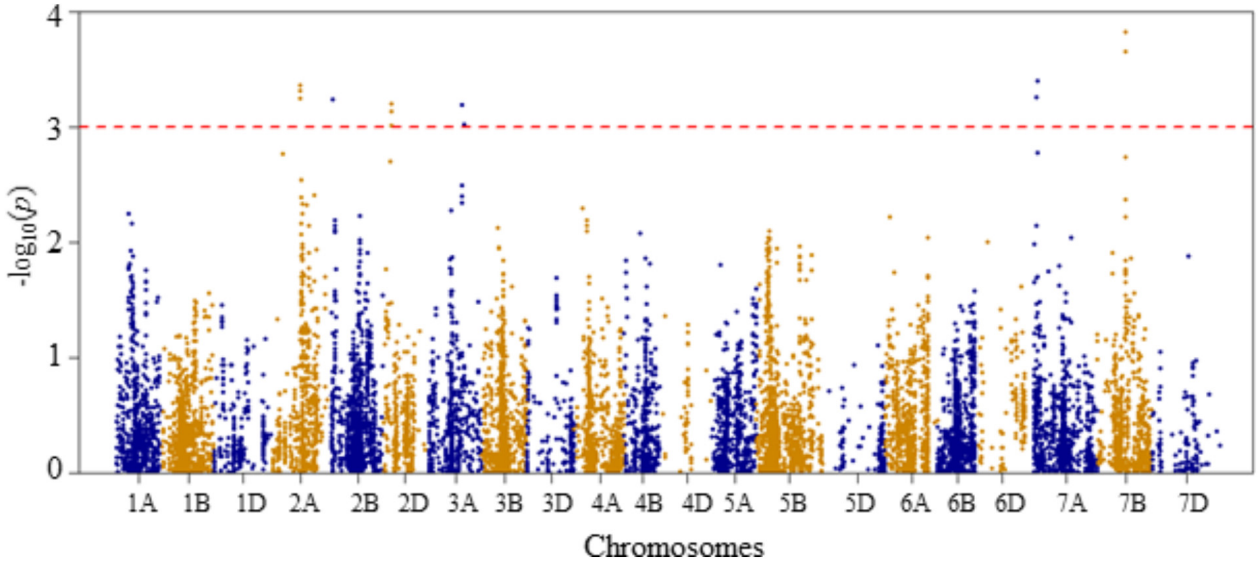

FIGURE 6 | Manhattan plot of GWAS conducted on seedling recovery after removal of heat stress treatment of the association mapping panel.

TABLE 2 | Heat stress responding QTL at the seedling stage of wheat in the present study.

Trait	Chr	Position (cM)	QTL for heat stress only[1]		QTL for heat response[2]	
			QTL Name	R^2 (%)	QTL Name	R^2 (%)
Leaf chlorophyll content	2B	27.2	QLCCHS.nri-2B	12.13	QLCCHR.nri-2B	6.80
Leaf chlorophyll content	2D	22.46			QLCCHR.nri-2D.1	5.80
Leaf chlorophyll content	2D	70.65–85.97	QLCCHS.nri-2D	24.64	QLCCHR.nri-2D.2	18.78
Leaf chlorophyll content	4A	8.61	QLCCHS.nri-4A	5.82	QLCCHR.nri-4A	6.58
Leaf chlorophyll content	4B	39.93–41.65	QLCCHS.nri-4B.1	5.96	QLCCHR.nri-4B.1	6.59
Leaf chlorophyll content	4B	59.94	QLCCHS.nri-4B.2	5.83		
Leaf chlorophyll content	4B	75.65	QLCCHS.nri-4B.3	17.53	QLCCHR.nri-4B.2	18.51
Leaf chlorophyll content	5B	182.15–188.58			QLCCHR.nri-5B	24.85
Shoot length (cm)	3B	9.7			QSLHR.nri-3B.1	6.21
Shoot length (cm)	3B	67.17			QSLHR.nri-3B.2	5.64
Shoot length (cm)	7D	26.92			QSLHR.nri-7D	12.55
Shoot length (cm)	2A	150.11			QLNHR.nri-2A.2	8.29
Number of leaves per seedling	3A	111.62	QLNHS.nri-3A	6.22		
Number of leaves per seedling	3A	177.24			QLNHR.nri-3A	6.71
Number of leaves per seedling	3B	65.72	QLNHS.nri-3B	5.91		
Number of leaves per seedling	4B	68.45–71.46			QLNHR.nri-4B	13.06
Number of leaves per seedling	5A	114.97	QLNHS.nri-5A	6.12		
Number of leaves per seedling	5B	49.02			QLNHR.nri-5B.1	12.88
Number of leaves per seedling	5B	144.26			QLNHR.nri-5B.2	5.93
Number of leaves per seedling	7B	145.29			QLNHR.nri-7B	6.11
Seedling recovery (%)	2A	95.75	QSRHS.nri-2A	32.98	N/A	
Seedling recovery (%)	2B	19.16	QSRHS.nri-2B	6.48	N/A	
Seedling recovery (%)	2D	26.05	QSRHS.nri-2D	18.55	N/A	
Seedling recovery (%)	3A	123.05–128.87	QSRHS.nri-3A	12.33	N/A	
Seedling recovery (%)	7A	42.08–43.47	QSRHS.nri-7A	13.41	N/A	
Seedling recovery (%)	7B	89.82	QSRHS.nri-7B	23.35	N/A	

(1) QTL detected under heat stressed temperature but not at optimum temperature; (2) QTL detected using heat response, which is the trait relative difference between the two temperature treatments; QTL were declared significant based on unadjusted p-value <0.001; R^2 is phenotypic variance explained by each QTL expressed as a percentage.

The first QTL (*QSLOT.nri-4B*), represented by three SNPs, was mapped at 57–63 cM on chromosome 4B, explaining 17.4% of the phenotypic variation of shoot length at the optimal growth temperature. The other QTL (*QSLOT.nri-7B*) was mapped at 54 cM on chromosome 7B with about 5.3% of the phenotypic variation in shoot length.

On the other hand, at heat-stressed growth condition, the same two QTLs for shoot length were also found on chromosomes 4B and 7B (**Figure 4B** and **Supplementary Table S2**). On chromosome 4B, the QTL (*QSLHS.nri-4B*) was mapped at genetic position ranging from 57 to 60 cM. This QTL explained 12.8% of the phenotypic variation in shoot length. The QTL (*QSLHS.nri-7B*) on chromosome 7B was represented by two SNPs and mapped within 54–58 cM. Together, the two SNP markers explained 10% of the phenotypic variation in shoot length at heat-stressed growth condition. The most significant markers were the same markers that were detected at optimal growth condition, located on chromosomes 4B and 7B, indicating that the detected shoot length QTLs are expressed under both optimum and heat-stressed growth conditions, thus they are not necessarily related to heat stress.

For heat response of shoot length, three QTLs were detected on chromosomes 3B and 7D (**Figure 4C** and **Supplementary Table S2**). On chromosome 3B, two QTLs (*QSLHR.nri-3B.1 and QSLHR.nri-3B.2*) were found, one mapped at 10 cM and the second one at 67 cM, together explaining 11.8% of the phenotypic variation in heat stress response of shoot length. The third QTL (*QSLHR.nri-7D*) was located at 27 cM on chromosome 7D. This QTL was represented by two SNPs, which collectively explained 12.8% of the phenotypic variation in heat stress response. In short, as the same QTLs were detected under optimal and heat-stressed growth conditions, shoot length QTLs responding to heat stress were only found by mapping heat stress response of the trait on chromosomes 3B and 7D (**Table 2**).

At optimal growth condition, four QTLs associated with the number of leaves per seedling were detected at genetic positions of 56, 77–78, 177–181, and 68–72 cM on chromosomes 1B, 2A, 3A, and 4B, respectively (**Figure 5A** and **Supplementary Table S2**). The SNP markers representing the QTLs explained 5.8–8.9% of total phenotypic variation in the number of leaves per seedling at the optimal growth condition. The two most significant SNP markers (IWB40186 and IWB25267) were co-localized at 78 cM on chromosome 2A, explaining 17.2% of the phenotypic variation in the number of leaves per seedling. The third most significant SNP was mapped at 68 cM on chromosome 4B, which accounted for 7.4% of phenotypic variation.

At heat-stressed growth condition, four QTLs significantly associated with number of leaves per seedling were detected (**Figure 5B** and **Supplementary Table S2**). The first QTL (*QLNHS.nri-1B*) was mapped at 112 cM on chromosome 3A, and explained about 6.2% of the phenotypic variation. The second (*QLNHS.nri-3B*), third (*QLNHS.nri-4B*), and fourth QTLs (*QLNHS.nri-5A*) were located at genetic positions of 66, 64, and 115 cM on chromosomes 3B, 4B, and 5A, respectively, and collectively explained 24.2% of the phenotypic variation in number of leaves per seedling at heat-stressed growth condition.

For heat stress response of number of leaves per seedling, seven QTLs, represented by 26 significant SNPs, were detected on chromosomes 2A, 3A, 4B, 5B, and 7B (**Figure 5C** and **Supplementary Table S2**). On chromosome 2A, two QTLs were found; one (*QLNHR.nri-2A.1*) mapped at 77–78 cM, and the other QTL (*QLNHR.nri-2A.2*) was located at 150 cM. The QTL (*QLNHR.nri-3A*) on 3A was located at 177 cM, while the one on 4B (*QLNHR.nri-4B*) was mapped at genetic position of 68–71 cM. Furthermore, two QTLs, *QLNHR.nri-5B.1* and *QLHR.nri-5B.2* were located at 49 and 144 cM, respectively, on chromosome 5A, while one QTL, *QLNHR.nri-7B* was found at 145 cM on chromosome 7B. The most significant SNP markers were IWB40186 and IWB25267, which were co-localized at 78 cM on chromosome 2A, and IWB61157, which was mapped at 150 cM on the same chromosome. The two markers mapped at 78 cM together explained 15.2% of the phenotypic variation, while the marker located at 150 cM accounted for 8.3% of the phenotypic variation in heat stress response of number of leaves per seedling. Overall, the data suggest that heat stress or heat response QTLs associated with the number of leaves per seedling are located on chromosomes 2A, 3A, 3B, 4B, 5A, 5B, and 7B according to QTLs detected for heat stress response of the trait, or by comparing the QTLs detected under heat-stressed vs. the optimum condition (**Table 2**).

For seedling recovery after removal of heat stress treatment, six QTLs were detected on chromosomes 2A, 2B, 2D, 3A, 7A, and 7B, and these were represented by 16 SNPs (**Figure 6** and **Supplementary Table S2**). The phenotypic variation explained by these SNPs varied from 6.5 to 8.5%. On chromosome 2A, one QTL (*QSLHS.nri-2A*) was located at genetic position of 96 cM. This QTL was represented by five SNPs, which collectively explained 33% of the phenotypic variation in seedling recovery after heat stress. The second QTL (*QSLHS.nri-2B*) was found on chromosome 2B at 19 cM, which accounted for 6.5% of the phenotypic variation. Another QTL (*QSLHS.nri-2D*) was found on chromosome 2D at the genetic distance of 26 cM, and it was represented by three SNP markers, which together explained 18.6% of the phenotypic variation. On chromosome 3A, one QTL (*QSLHS.nri-3A*) was detected and mapped at 123–129 cM. The QTL on 3A explained 12.4% of the phenotypic variation in seedling recovery after removal of heat stress treatment. In addition, one QTL (*QSLHS.nri-7A*) was identified on chromosome 7A at position 42–43 cM, while another one (*QSLHS.nri-7B*) was found at 90 cM on chromosome 7B. The QTLs on 7A and 7B accounted for 13.4 and 23.3% of the phenotypic variation in seedling recovery, respectively.

DISCUSSION

Wheat is one of the most important food and feed crops in the world. In the Southern Plains of the United States including Oklahoma and Texas where livestock and forage production are the largest contributors to agricultural income, winter wheat is often used for cool season grazing, which needs early planting for increased fall to winter forage production. Winter wheat under a dual-purpose management system could be planted as early

as the end of August, when the temperature is often still very high for the crop to establish. Therefore, improving seedling heat tolerance for winter wheat grown for forage and grain production will have a huge economic impact in the region. We conducted a GWAS to map QTLs and identify SNP markers associated with seedling heat tolerance for MAS of seedling heat tolerance during wheat breeding. Identification of QTLs associated with seedling heat tolerance will facilitate the introgression of heat tolerance alleles into elite wheat cultivars through MAS.

In the present study, the association mapping panel showed significant phenotypic variation in leaf chlorophyll content, shoot length, number of leaves per seedling at optimal and high temperature regimes, and seedling recovery after removal of heat stress treatment. In addition, variation in heat stress response, i.e., relative performance difference between the two temperatures, for all traits was also observed. These results suggest that there is a great potential that these lines can be used to mine alleles for seedling heat tolerance for introgression into elite winter wheat lines for seedling heat tolerance improvement.

Population structure and familial relatedness can result in false positives in GWAS (Crossa et al., 2007; Matthies et al., 2012). Therefore, when GWAS is conducted, these parameters need to be considered in the model. In the present study, the level of genetic structure of the panel was assessed by the PCA, NJ tree, and STRUCTURE analyses. Results from the three clustering methods showed that this panel is structured into four major groups. Our results agree with previous GWAS done using winter wheat lines selected from the same hard red winter wheat association mapping panel (Ayana, 2017). In their study, they used 294 lines of the association mapping panel to molecular characterize spot blotch and bacterial leaf streak resistance in bread wheat, and the STRUCTURE analysis revealed four major groups existing in this panel, although admixtures were also observed. In the present study, the stratification was mainly based on geographical regions and pedigree relation. Genetic structuring of winter wheat lines along geographic regions has also been previously reported (Li et al., 2016; Liu et al., 2017).

In this study, we observed that some lines with common parents in their pedigrees tended to cluster in the same subgroup within the main group. For instance, some lines with "Jagger" wheat line as one of the parents in their pedigrees formed one subgroup. This result corroborates the GWAS of powdery mildew disease using a different set of winter wheat lines, in which the authors found that the lines were structured along pedigree information (Liu et al., 2017). Specifically, they found that 13 accessions with the common parent, "Jagger" in their pedigrees clustered in one group. However, in general, we observed that the level of genetic stratification was low, as revealed by the modest contribution of the three PCs (19.4%) to the total genetic variance. This reduced and loose population stratification is because of historical admixture resulting from sharing genetic materials among different wheat breeding programs in the hard red winter wheat region of the United States.

Linkage disequilibrium is one of the most important factors in association mapping studies because it determines the power of association between QTLs and phenotype. In this study, we estimated the LD decay distances of the three wheat genomes including A, B, and D genomes. Our results suggest that D genome had the highest LD decay distance (10.9 cM) compared to A (9.7 cM) and B (9.8 cM) genomes. As only 200 representative lines were selected from the original panel in the current study, the LD distances are changed compared to a previous study involving the same panel (Ayana, 2017). In general, our results corroborate previous studies done in wheat (Zhang et al., 2010; Hao et al., 2011; Ayana, 2017). However, other studies have reported much higher LD than estimated in our study when using different ecotype wheat lines. For example, LD decay distance of 23 cM in European hexaploid wheat lines has been reported (Nielsen et al., 2014).

In this study, three statistical models were compared to assess their ability to map QTLs and identify SNPs associated with seedling heat tolerance. We decided to do this because previous studies have shown that the best model can vary depending on the trait (Gurung et al., 2014). Finally, we selected the MLM, which accounts for both population structure (PCs) and K matrix, because of its statistical power to control false positives. To the best of our knowledge, rare QTL studies have been done for heat tolerance during the seedling stage of wheat. However, there have been a lot of QTLs studies in heat tolerance during the flowering stage or grain filling stage of wheat. For example, QTLs for heat tolerance during the grain filling stage of wheat have been reported (Vijayalakshmi et al., 2010; Talukder et al., 2014). Similarly, in other cereal crops such as sorghum (Chen et al., 2017; Chopra et al., 2017) and rice (Lafarge et al., 2017), QTL studies for heat tolerance have been conducted. The focus of previous studies was either on the vegetative stage or the flowering stage because heat stress during the flowering stage has been one of the most important limiting factors contributing to yield losses in many crop species. However, heat stress during the seedling stage of winter wheat has been a common issue in the southern Great Plains of the United States due to early planting, particularly in a dual-purpose management system, in which case the crop is planted very early in the fall. Therefore, this study was primarily conducted to unravel QTLs or genes associated with seedling heat tolerance in winter wheat purposely grown for forage as well as grain production.

Using the MLM, we identified multiple significant QTLs for wheat seedling traits at optimum and heat-stressed growth conditions. QTLs associated with seedling heat stress or heat response were found by comparing the QTLs detected under heat-stressed vs. the optimum condition, or mapping heat response QTLs using the relative phenotypic trait difference between the two growth conditions. QTLs associated with leaf chlorophyll content at heat-stressed growth condition but not at optimum temperature were found on chromosomes 2B, 2D, 4A, and 4B, while QTLs for heat stress response of the trait were detected on chromosomes 2B, 2D, 4A, 4B, and 5B (**Table 2**). We believe that these are the true chromosomes that harbor leaf chlorophyll content QTLs responding to heat stress since they were only detected under heat stressed temperature and/or mapped using heat response of the trait. Previous studies also identified QTLs for heat stress tolerance traits, specifically at grain filling stage of wheat on chromosomes 2B, 2D, and 4A (Paliwal et al., 2012; Acuña-Galindo et al., 2015;

Bhusal et al., 2017). However, previous QTL studies conducted in wheat also identified QTLs for leaf chlorophyll content under heat stress on other chromosomes including 1B, 1D, 6A, and 7A (Talukder et al., 2014), which were not detected at the seedling stage in the current study. Moreover, in another QTL study, leaf chlorophyll content QTLs under heat stress mapped on chromosomes 1A and 6B were reported (Tahmasebi et al., 2016). Again, these QTLs were not detected in the present study. However, the QTL for heat-stress response of the leaf chlorophyll content detected on 5B was not reported in other studies mentioned above. Although QTLs detected at optimal temperature are not related to heat stress, interestingly in this study one SNP marker (IWB14950) on 1B, which associated with leaf chlorophyll content at optimal growth condition, has high sequence similarity with kDa class VI heat shock protein, known to be involved in heat stress tolerance. However, this SNP marker was not detected at heat-stressed growth condition as well as heat stress response.

We also conducted BLAST search on NCBI to unravel candidate genes using sequences of the SNPs detected in the present study (**Supplementary Table S2**). The results showed that some of the significant SNP markers have high sequence similarities with candidate genes, known to be involved in plant stress responses in different crops including wheat. For example, on chromosome 2D, the significant SNP IWB28728 for leaf chlorophyll content responding to heat stress has 89% sequence similarity with putative plastid-lipid-associated protein 13. The putative plastid lipid-associated protein 13 has been reported to play an important role in improving plant performance under stress conditions. In addition, it actively participates in thylakoid function from biogenesis to senescence, suggesting that it is a precursor of the chloroplast thylakoid membranes (Rottet et al., 2015). Similarly, on chromosome 4B, significant SNP IWB42264 for leaf chlorophyll content at heat-stressed growth condition and heat response of the trait, has 94% sequence similarity with K (+) efflux antiporter 5 isoform X1, which contains potassium (K^+), a major osmoticum of plant cells. The accumulation of potassium (K^+) in the plant vacuole is important for plants under high-salt stressed conditions (Assaha et al., 2017). In addition, the significant SNP IWB18745 for heat stress response of leaf chlorophyll content on chromosome 2D has 97% sequence similarity with IAA-amino acid hydrolase ILR1-like, which is able to hydrolyze certain amino acid conjugates of the plant growth regulator indole-3-acetic acid (IAA) (LeClere et al., 2002). Moreover, on chromosome 2D, the heat responding SNP IWB4541 has a DNA sequence with 100% similarity to that of the heat shock N-terminal domain-containing protein found in maize, which is essentially involved in plant responses to various environmental stress including heat.

For shoot length, the same significant QTLs were detected at both optimal and heat-stressed growth conditions. Generally, QTLs associated with a trait under optimal conditions usually controls the trait under stressed-conditions (Mathews et al., 2008; Mwadzingeni et al., 2017). In the present study, this scenario was observed for shoot length QTLs on chromosomes 4B and 7B indicating that the detected QTLs were associated with shoot length itself as a plant architecture trait, and not related with heat stress tolerance *per se*. These results suggest that the effects of these QTLs are not influenced by temperature changes. Therefore, such kind of QTLs may be useful in marker-assisted breeding (MAB) of crops with broad environmental adaptation. On the other hand, shoot length QTLs were detected for heat response on 3B and 7D, which were also reported previously to harbor QTLs for heat tolerance traits at vegetative and grain filling stages of wheat (Vijayalakshmi et al., 2010; Paliwal et al., 2012).

Although some of the markers associated with shoot length were significant at both growth conditions, BLAST search revealed that some of the identified SNPs have high sequence similarities with candidate genes known for plant stress response. For example, the DNA sequence of SNP IWB35611 on chromosome 4B has high sequence similarity with serine/threonine protein kinase STE 20-like, which has been reported to play an important role in salt tolerance in plants (Liang et al., 2011). Another SNP IWB12856 on chromosome 4B has high sequence similarity with inositol-tetrakisphosphate 1-kinase 3, transcript variant X1, which has been reported to confer plant stress tolerance (Yang et al., 2008). The two SNP markers on 4B were located 2.45 cM apart from each other. In addition, the SNP marker IWB1428 on 3B, which was found to be significantly associated with heat stress response of shoot length, showed 83% sequence similarity with G-type lectin S-receptor-like serine/threonine protein kinase. Research done in the past showed that the G-type lectin S-receptor-like serine/threonine protein kinase acts as a positive regulator of plant tolerance to salt stress (Deng et al., 2009; Sun et al., 2013).

Similarly, for the number of leaves per seedling and seedling recovery, some of the QTLs detected in this study were located in the same chromosomes that were reported in other heat stress studies at various adult plant stages (Mason et al., 2010; Vijayalakshmi et al., 2010; Paliwal et al., 2012; Talukder et al., 2014; Acuña-Galindo et al., 2015). However, BLAST search against sequences of SNPs associated with the number of leaves per seedling and seedling recovery did not reveal any candidate genes that are known responding to abiotic stress.

In summary, some QTLs for seedling heat tolerance-related traits identified in this study were found on the same chromosomes previously reported to harbor QTLs for heat tolerance, although the growth stages reported in the previous studies are different from the growth stage investigated in the present study. Our results suggest that some of heat tolerance QTLs detected during the seedling and the flowering stages of wheat may be co-localized. In addition, other QTLs identified in the seedling stage in the present study have not been reported in those studies conducted at the flowering time or grain filling stages. Moreover, BLAST search using DNA sequences of some of the significant loci found in this study revealed candidate genes known to be involved in plant stress responses in wheat and other crop species. To the best of our knowledge, this is the first GWAS to map QTLs and identify SNP markers significantly associated with seedling heat tolerance-related traits targeting early planting of dual-purpose winter wheat. Significant SNP markers identified in this study will be used for MAS of seedling heat tolerance to facilitate selection of the trait during wheat breeding.

AUTHOR CONTRIBUTIONS

FM phenotyped the association mapping panel, analyzed both phenotypic and genotypic data, and drafted the manuscript. HA helped in the candidate gene search and review of the manuscript. JA assisted in experiment implementation and review of the manuscript. TK helped in data collection and review of the manuscript. WH helped in review of manuscript. X-FM supervised the study and finalized the manuscript. All authors read and approved the manuscript.

FUNDING

This study was funded by the Samuel Roberts Noble Foundation.

ACKNOWLEDGMENTS

The authors sincerely thank Andrea Mongler for critical reading of the manuscript. The authors also thank Triticeae Coordinated Agricultural Project (TCAP) for making the genotypic data publicly available.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01272/full#supplementary-material

FIGURE S1 | Structure analysis of the 200 winter wheat lines from the association mapping panel. **(A)** Neighbor-joining (NJ) tree; **(B)** population structure.

FIGURE S2 | Scatter plots showing the linkage disequilibrium (LD) decay curves of the three subgenomes estimated in the association mapping panel. The LD estimates (r^2) for pairs of SNP markers were plotted against the genetic distance in cM.

FIGURE S3 | The quantile-quantile (Q-Q) plots of the mixed liner model applied to the investigated traits. **(A–C)** Leaf chlorophyll content at optimum temperature, heat-stressed growth condition, and heat response of the trait; **(D–F)** shoot length (cm) at optimum temperature, heat-stressed growth condition, and heat response of the trait; **(G-I)** Number of leaves at optimum temperature, heat-stressed growth condition, and heat response of the trait; **(J)** Seedling recovery after removal of the heat stress.

TABLE S1 | Linkage disequilibrium analysis of 200 lines in the hard red winter association mapping panel.

TABLE S2 | The QTL and the significant SNP markers associated with seedling traits of wheat at optimal and heat-stressed growth conditions.

REFERENCES

Acuña-Galindo, M. A., Mason, R. E., Subramanian, N. K., and Hays, D. B. (2015). Meta-analysis of wheat QTL regions associated with adaptation to drought and heat stress. *Crop Sci.* 55, 477–492. doi: 10.2135/cropsci2013.11.0793

Arruda, M. P., Brown, P., Brown-Guedira, G., Krill, A. M., Thurber, C., Merrill, K. R., et al. (2016). Genome-wide association mapping of fusarium head blight resistance in wheat using genotyping-by-sequencing. *Plant Genome* 9, 1–14. doi: 10.3835/plantgenome2015.04.0028

Assaha, D. V. M., Ueda, A., Saneoka, H., Al-Yahyai, R., and Yaish, M. W. (2017). The role of Na(+) and K(+) transporters in salt stress adaptation in glycophytes. *Front. Physiol.* 8:509. doi: 10.3389/fphys.2017.00509

Ayana, G. (2017). *Molecular Characterization of Spot Blotch and Bacterial Leaf Streak Resistance in Bread Wheat Electronic*. Ph.D. thesis, South Dakota State University, Brookings, SD.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300.

Bhusal, N., Sarial, A. K., Sharma, P., and Sareen, S. (2017). Mapping QTLs for grain yield components in wheat under heat stress. *PLoS One* 12:e0189594. doi: 10.1371/journal.pone.0189594

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

Butler, J. (2002). *Quantitative Trait Locus Evaluation for Agronomic and Morphological Traits in a Spring Wheat Population*. Ph.D. thesis, Colorado State University, Fort Collins, CO.

Casa, A. M., Pressoir, G., Brown, P. J., Mitchell, S. E., Rooney, W. L., Tuinstra, M. R., et al. (2008). Community resources and strategies for association mapping in sorghum all rights reserved. no part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher. *Crop Sci.* 48, 30–40. doi: 10.2135/cropsci2007.02.0080

Chatrath, R., Mishra, B., Ferrara, G. O., Singh, S., and Joshi, A. (2007). Challenges to wheat production in South Asia. *Euphytica* 157, 447–456. doi: 10.1007/s10681-007-9515-2

Chaturvedi, A. K., Bahuguna, R. N., Shah, D., Pal, M., and Jagadish, S. V. K. (2017). High temperature stress during flowering and grain filling offsets beneficial impact of elevated CO(2) on assimilate partitioning and sink-strength in rice. *Sci. Rep.* 7:8227. doi: 10.1038/s41598-017-07464-6

Chen, J., Chopra, R., Hayes, C., Morris, G., Marla, S., Burke, J., et al. (2017). Genome-wide association study of developing leaves' heat tolerance during vegetative growth stages in a sorghum association panel. *Plant Genome* 10, 1–15. doi: 10.3835/plantgenome2016.09.0091

Chopra, R., Burow, G., Burke, J. J., Gladman, N., and Xin, Z. (2017). Genome-wide association analysis of seedling traits in diverse Sorghum germplasm under thermal stress. *BMC Plant Biol.* 17:12. doi: 10.1186/s12870-016-0966-2

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 74, 829–836. doi: 10.1080/01621459.1979.10481038

Cossani, C. M., and Reynolds, M. P. (2012). Physiological traits for improving heat tolerance in wheat. *Plant Physiol.* 160, 1710–1718. doi: 10.1104/pp.112.207753

Crossa, J., Burgueño, J., Dreisigacker, S., Vargas, M., Herrera-Foessel, S. A., Lillemo, M., et al. (2007). Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. *Genetics* 177, 1889–1913. doi: 10.1534/genetics.107.078659

Deng, K., Wang, Q., Zeng, J., Guo, X., Zhao, X., Tang, D., et al. (2009). A lectin receptor kinase positively regulates aba response during seed germination and is involved in salt and osmotic stress response. *J. Plant Biol.* 52:493. doi: 10.1007/s12374-009-9063-5

Driedonks, N., Rieu, I., and Vriezen, W. H. (2016). Breeding for plant heat tolerance at vegetative and reproductive stages. *Plant Reprod.* 29, 67–79. doi: 10.1007/s00497-016-0275-9

Endelman, J. B., and Jannink, J.-L. (2012). Shrinkage estimation of the realized relationship matrix. *G3* 2, 1405–1413. doi: 10.1534/g3.112.004259

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.

Feng, B., Liu, P., Li, G., Dong, S. T., Wang, F. H., Kong, L. A., et al. (2014). Effect of heat stress on the photosynthetic characteristics in flag leaves at the grain-filling stage of different heat-resistant winter wheat varieties. *J. Agron. Crop Sci.* 200, 143–155. doi: 10.1111/jac.12045

Grogan, S. M., Anderson, J., Baenziger, P. S., Frels, K., Guttieri, M. J., Haley, S. D., et al. (2016). Phenotypic plasticity of winter wheat heading date and grain yield

across the US great plains. *Crop Sci.* 56, 2223–2236. doi: 10.2135/cropsci2015.06.0357

Gurung, S., Mamidi, S., Bonman, J. M., Xiong, M., Brown-Guedira, G., and Adhikari, T. B. (2014). Genome-wide association study reveals novel quantitative trait loci associated with resistance to multiple leaf spot diseases of spring wheat. *PLoS One* 9:e108179. doi: 10.1371/journal.pone.0108179

Guttieri, M. J., Frels, K., Regassa, T., Waters, B. M., and Baenziger, P. S. (2017). Variation for nitrogen use efficiency traits in current and historical great plains hard winter wheat. *Euphytica* 213:87. doi: 10.1007/s10681-017-1869-5

Hao, C., Wang, L., Ge, H., Dong, Y., and Zhang, X. (2011). Genetic diversity and linkage disequilibrium in chinese bread wheat (*Triticum aestivum* L.) revealed by SSR markers. *PLoS One* 6:e17279. doi: 10.1371/journal.pone.0017279

Joshi, A., Mishra, B., Chatrath, R., Ferrara, G. O., and Singh, R. P. (2007). Wheat improvement in India: present status, emerging challenges and future prospects. *Euphytica* 157, 431–446. doi: 10.1007/s10681-007-9385-7

Kim, K.-S., and Anderson, J. D. (2015). Forage yield and nutritive value of winter wheat varieties in the southern Great Plains. *Euphytica* 202, 445–457. doi: 10.1007/s10681-014-1325-8

Lafarge, T., Bueno, C., Frouin, J., Jacquin, L., Courtois, B., and Ahmadi, N. (2017). Genome-wide association analysis for heat tolerance at flowering detected a large set of genes involved in adaptation to thermal and other stresses. *PLoS One* 12:e0171254. doi: 10.1371/journal.pone.0171254

LeClere, S., Tellez, R., Rampey, R. A., Matsuda, S. P. T., and Bartel, B. (2002). Characterization of a family of iaa-amino acid conjugate hydrolases from Arabidopsis. *J. Biol. Chem.* 277, 20446–20452. doi: 10.1074/jbc.M111955200

Li, G., Xu, X., Bai, G., Carver, B. F., Hunger, R., Bonman, J. M., et al. (2016). Genome-wide association mapping reveals novel QTL for seedling leaf rust resistance in a worldwide collection of winter wheat. *Plant Genome* 9, 1–12. doi: 10.3835/plantgenome2016.06.0051

Liang, C., Zhang, X., Chi, X., Guan, X., Li, Y., Qin, S., et al. (2011). Serine/threonine protein kinase spkg is a candidate for high salt resistance in the unicellular cyanobacterium *Synechocystis* sp. PCC 6803. *PLoS One* 6:e18718. doi: 10.1371/journal.pone.0018718

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, N., Bai, G., Lin, M., Xu, X., and Zheng, W. (2017). Genome-wide association analysis of powdery mildew resistance in U.S. Winter Wheat. *Sci. Rep.* 7:11743. doi: 10.1038/s41598-017-11230-z

Mason, R. E., Mondal, S., Beecher, F. W., and Hays, D. B. (2011). Genetic loci linking improved heat tolerance in wheat (*Triticum aestivum* L.) to lower leaf and spike temperatures under controlled conditions. *Euphytica* 180, 181–194. doi: 10.1007/s10681-011-0349-6

Mason, R. E., Mondal, S., Beecher, F. W., Pacheco, A., Jampala, B., Ibrahim, A. M., et al. (2010). QTL associated with heat susceptibility index in wheat (*Triticum aestivum* L.) under short-term reproductive stage heat stress. *Euphytica* 174, 423–436. doi: 10.1007/s10681-010-0151-x

Mathews, K. L., Malosetti, M., Chapman, S., McIntyre, L., Reynolds, M., Shorter, R., et al. (2008). Multi-environment QTL mixed models for drought stress adaptation in wheat. *Theor. Appl. Genet.* 117, 1077–1091. doi: 10.1007/s00122-008-0846-8

Matthies, I. E., van Hintum, T., Weise, S., and Röder, M. S. (2012). Population structure revealed by different marker types (SSR or DArT) has an impact on the results of genome-wide association mapping in European barley cultivars. *Mol. Breed.* 30, 951–966. doi: 10.1007/s11032-011-9678-3

Mondal, S., Mason, R. E., Huggins, T., and Hays, D. B. (2015). QTL on wheat (*Triticum aestivum* L.) chromosomes 1B, 3D and 5A are associated with constitutive production of leaf cuticular wax and may contribute to lower leaf temperatures under heat stress. *Euphytica* 201, 123–130. doi: 10.1007/s10681-014-1193-2

Müller, B. U., Stich, B., and Piepho, H. P. (2011). A general method for controlling the genome-wide type I error rate in linkage and association mapping experiments in plants. *Heredity* 106, 825–831. doi: 10.1038/hdy.2010.125

Mwadzingeni, L., Shimelis, H., Rees, D. J. G., and Tsilo, T. J. (2017). Genome-wide association analysis of agronomic traits in wheat under drought-stressed and non-stressed conditions. *PLoS One* 12:e0171692. doi: 10.1371/journal.pone.0171692

Nielsen, N. H., Backes, G., Stougaard, J., Andersen, S. U., and Jahoor, A. (2014). Genetic diversity and population structure analysis of european hexaploid bread wheat (*Triticum aestivum* L.) Varieties. *PLoS One* 9:e94000. doi: 10.1371/journal.pone.0094000

Paliwal, R., Röder, M. S., Kumar, U., Srivastava, J., and Joshi, A. K. (2012). QTL mapping of terminal heat tolerance in hexaploid wheat (*T. aestivum L.*). *Theor. Appl. Genet.* 125, 561–575. doi: 10.1007/s00122-012-1853-3

Prasad, P. V. V., Staggenborg, S. A., and Ristic, Z. (2008). "Impacts of drought and/or heat stress on physiological, developmental, growth, and yield processes of crop plants," in *Response of Crops to Limited Water: Understanding and Modeling Water Stress Effects on Plant Growth Processes*, eds L. R. Ahuja, V. R. Reddy, S. A. Saseendran, and Q. Yu (Madison, WI: American Society of Agronomy), 301–355.

Prerna, A., Kumar, A., and Sengar, R. (2013). Evaluation of heat and drought tolerance of wheat cultivars through physiological, biochemical and molecular approaches. *Res. J. Agric. Sci.* 4, 139–145.

Ristic, Z., Bukovnik, U., Momčilović, I., Fu, J., and Prasad, P. V. (2008). Heat-induced accumulation of chloroplast protein synthesis elongation factor, EF-Tu, in winter wheat. *J. Plant Physiol.* 165, 192–202.

Ristic, Z., Bukovnik, U., and Prasad, P. V. V. (2007). Correlation between heat stability of thylakoid membranes and loss of chlorophyll in winter wheat under heat stress all rights reserved. no part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. permission for printing and for reprinting the material contained herein has been obtained by the publisher. *Crop Sci.* 47, 2067–2073. doi: 10.2135/cropsci2006.10.0674

Rottet, S., Besagni, C., and Kessler, F. (2015). The role of plastoglobules in thylakoid lipid remodeling during plant development. *Biochim. Biophys. Acta Bioenerg.*1847, 889–899. doi: 10.1016/j.bbabio.2015.02.002

SAS Institute (2011). *The SAS System for Windows. V.*9.3. Cary, NC: SAS Institute.

Schapendonk, A. H. C. M., Xu, H. Y., Van Der Putten, P. E. L., and Spiertz, J. H. J. (2007). Heat-shock effects on photosynthesis and sink-source dynamics in wheat (*Triticum aestivum* L.). *Njas Wageningen J. Life Sci.* 55, 37–54. doi: 10.1016/S1573-5214(07)80003-0

Shi, P., Zhu, Y., Tang, L., Chen, J., Sun, T., Cao, W., et al. (2016). Differential effects of temperature and duration of heat stress during anthesis and grain filling stages in rice. *Environ. Exp. Bot.* 132, 28–41. doi: 10.1016/j.envexpbot.2016.08.006

Stratonovitch, P., and Semenov, M. A. (2015). Heat tolerance around flowering in wheat identified as a key trait for increased yield potential in Europe under climate change. *J. Exp. Bot.* 66, 3599–3609. doi: 10.1093/jxb/erv070

Sun, X.-L., Yu, Q.-Y., Tang, L.-L., Ji, W., Bai, X., Cai, H., et al. (2013). GsSRK, a G-type lectin S-receptor-like serine/threonine protein kinase, is a positive regulator of plant tolerance to salt stress. *J. Plant Physiol.* 170, 505–515. doi: 10.1016/j.jplph.2012.11.017

Tadesse, W., Ogbonnaya, F., Jighly, A., Sanchez-Garcia, M., Sohail, Q., Rajaram, S., et al. (2015). Genome-wide association mapping of yield and grain quality traits in winter wheat genotypes. *PLoS One* 10:e0141339. doi: 10.1371/journal.pone.0141339

Tahmasebi, S., Heidari, B., Pakniyat, H., and McIntyre, C. L. (2016). Mapping QTLs associated with agronomic and physiological traits under terminal drought and heat stress conditions in wheat (*Triticum aestivum* L.). *Genome* 60, 26–45. doi: 10.1139/gen-2016-0017

Talukder, S. K., Babar, M. A., Vijayalakshmi, K., Poland, J., Prasad, P. V. V., Bowden, R., et al. (2014). Mapping QTL for the traits associated with heat tolerance in wheat (*Triticum aestivum* L.). *BMC Genet.* 15:97. doi: 10.1186/s12863-014-0097-4

Turner, S. D. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* 3:731. doi: 10.21105/joss.00731

Vijayalakshmi, K., Fritz, A. K., Paulsen, G. M., Bai, G., Pandravada, S., and Gill, B. S. (2010). Modeling and mapping QTL for senescence-related traits in winter wheat under high temperature. *Mol. Breed.* 26, 163–175. doi: 10.1007/s11032-009-9366-8

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787–796. doi: 10.1111/pbi.12183

Wardlaw, I. F., Blumenthal, C., Larroque, O., and Wrigley, C. W. (2002). Contrasting effects of chronic heat stress and heat shock on kernel weight and flour quality in wheat. *Funct. Plant Biol.* 29, 25–34. doi: 10.1071/PP00147

Yang, H., Lu, D., Shen, X., Cai, X., and Lu, W. (2015). Heat stress at different grain filling stages affects fresh waxy maize grain yield and quality. *Cereal Chem.* 92, 258–264. doi: 10.1094/CCHEM-07-14-0146-R

Yang, L., Tang, R., Zhu, J., Liu, H., Mueller-Roeber, B., Xia, H., et al. (2008). Enhancement of stress tolerance in transgenic tobacco plants constitutively expressing AtIpk2β, an inositol polyphosphate 6-/3-kinase from *Arabidopsis thaliana*. *Plant Mol. Biol.* 66, 329–343. doi: 10.1007/s11103-007-9267-3

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, D., Bai, G., Zhu, C., Yu, J., and Carver, B. F. (2010). Genetic diversity, population structure, and linkage disequilibrium in U.S. Elite winter wheat. *Plant Genome* 3, 117–127. doi: 10.3835/plantgenome2010.03.0004

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., et al. (2007). An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* 3:e4. doi: 10.1371/journal.pgen.0030004

Zheng, X. (2013). *A Tutorial for the R Package SNPRelate*. Washington, DC: University of Washington.

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.2018.01166

Genetic Diversity and Population Structure of the USDA Sweetpotato (*Ipomoea batatas*) Germplasm Collections Using GBSpoly

Phillip A. Wadl[1], Bode A. Olukolu[2,3]*, Sandra E. Branham[1], Robert L. Jarret[4], G. Craig Yencho[2] and D. Michael Jackson[1]*

[1] United States Vegetable Laboratory, United States Department of Agriculture, Agricultural Research Service, Charleston, SC, United States, [2] Department of Horticultural Science, North Carolina State University, Raleigh, NC, United States, [3] Department of Entomology and Plant Pathology, University of Tennessee, Knoxville, Knoxville, TN, United States, [4] Plant Genetic Resources Conservation Unit, United States Department of Agriculture, Agricultural Research Service, Griffin, GA, United States

OPEN ACCESS

Edited by:
Yiwei Jiang,
Purdue University, United States

Reviewed by:
Don LaBonte,
Louisiana State University, United States
Heena Ambreen,
University of Delhi, India
Barbara Pipan,
Agricultural Institute of Slovenia, Slovenia

***Correspondence:**
Phillip A. Wadl
Phillip.Wadl@ars.usda.gov
Bode A. Olukolu
bolukolu@utk.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *13 April 2018*
Accepted: *23 July 2018*
Published: *21 August 2018*

Citation:
Wadl PA, Olukolu BA, Branham SE, Jarret RL, Yencho GC and Jackson DM (2018) Genetic Diversity and Population Structure of the USDA Sweetpotato (Ipomoea batatas) Germplasm Collections Using GBSpoly. Front. Plant Sci. 9:1166. doi: 10.3389/fpls.2018.01166

Sweetpotato (*Ipomoea batatas*) plays a critical role in food security and is the most important root crop worldwide following potatoes and cassava. In the United States (US), it is valued at over $700 million USD. There are two sweetpotato germplasm collections (Plant Genetic Resources Conservation Unit and US Vegetable Laboratory) maintained by the USDA, ARS for sweetpotato crop improvement. To date, no genome-wide assessment of genetic diversity within these collections has been reported in the published literature. In our study, population structure and genetic diversity of 417 USDA sweetpotato accessions originating from 8 broad geographical regions (Africa, Australia, Caribbean, Central America, Far East, North America, Pacific Islands, and South America) were determined using single nucleotide polymorphisms (SNPs) identified with a genotyping-by-sequencing (GBS) protocol, GBSpoly, optimized for highly heterozygous and polyploid species. Population structure using Bayesian clustering analyses (STRUCTURE) with 32,784 segregating SNPs grouped the accessions into four genetic groups and indicated a high degree of mixed ancestry. A neighbor-joining cladogram and principal components analysis based on a pairwise genetic distance matrix of the accessions supported the population structure analysis. Pairwise F_{ST} values between broad geographical regions based on the origin of accessions ranged from 0.017 (Far East – Pacific Islands) to 0.110 (Australia – South America) and supported the clustering of accessions based on genetic distance. The markers developed for use with this collection of accessions provide an important genomic resource for the sweetpotato community, and contribute to our understanding of the genetic diversity present within the US sweetpotato collection and the species.

Keywords: Convolvulaceae, genotyping-by-sequencing, GBSpoly, polyploid, SNPs, sweetpotato, USDA germplasm

INTRODUCTION

Sweetpotato, *Ipomoea batatas* (L.) Lam. (Convolvulaceae), is the sixth most important food crop worldwide, following rice, wheat, potatoes, maize, and cassava (International Potato Center, 2018). This important root crop plays a critical role in food security, especially in developing countries. China is the largest producer of sweetpotato, accounting for over 70% of the world's production, followed by Sub-Saharan Africa. While global sweetpotato production has been relatively stable for the past 45 years (Padmaja, 2009), production and consumption in the US has increased considerably since 2000 (United States Department of Agriculture, Economic Research Service [USDA-ERS], 2016). The US produced 3.1 billion tons of sweetpotatoes in 2015 and is ranked in the top 10 countries in annual worldwide production of the crop (United States Department of Agriculture, Economic Research Service [USDA-ERS], 2016). One of the reasons for increased production and consumption appears to be increased public awareness of the health benefits of sweetpotatoes and the production of a wide array of value-added products from the crop. Sweetpotatoes not only provide a source of carbohydrates, but are a major source of vitamins A (carotenoids from the orange-fleshed types), C, B1, B2 (riboflavin), B3 (niacin), B6, E, biotin, and pantothenic acid, as well as dietary fiber, potassium, copper, manganese, and iron; additionally, they are low in fat and cholesterol (Hill et al., 1992; Wang et al., 2016).

The origin of cultivated sweetpotato is unclear (Rajapakse et al., 2004; Roullier et al., 2013b). The most recent proposed origin of sweetpotato is of autopolyploid origin with *I. trifida* as the sole relative (Munoz-Rodriguez et al., 2018). Another hypothesis has proposed that *I. batatas* is an allo-autohexaploid ($2n = 6x = 90$), with a $B_1B_1B_2B_2B_2B_2$ genome composition resulting from an initial crossing between a tetraploid ancestor and a diploid progenitor followed by a whole genome duplication event (Magoon et al., 1970; Yang et al., 2016). It has also been proposed that sweetpotato originated from hybridization by unreduced gametes of diploid *I. trifida* and a tetraploid *I. batatas* (Shiotani, 1987; Orjeda et al., 1989; Oracion et al., 1990; Freyre et al., 1991), or that the species is derived from *I. trifida* and *I. triloba* (Austin, 1988).

Regardless of its origin, the sweetpotato genome is hexaploid and highly heterozygous, and this genetic complexity has slowed genome sequencing, assembly, and annotation over the past 10 years. Nevertheless, the available molecular resources for sweetpotato are rapidly expanding and include *de novo* assembled transcriptomes of sweetpotato and several of its predicted wild-type relatives (Schafleitner et al., 2010; Wang et al., 2010; Tao et al., 2012; Xie et al., 2012; Effendy et al., 2013; Firon et al., 2013; Solis et al., 2014, 2016; Ponniah et al., 2017), microsatellite, specific length amplified fragment (SLAF) and amplified fragment length polymorphism (AFLP) markers to characterize genetic diversity (Bruckner, 2004; Techen et al., 2009; Schafleitner et al., 2010; Roullier et al., 2011, 2013a,b; Su et al., 2017), recently released draft genome assemblies (Yang et al., 2016; Zhou et al., 2017) and whole chloroplast genomes of sweetpotato and wild relatives (Munoz-Rodriguez et al., 2018). Diploid reference genome assemblies based on progenitor wild relatives, *I. trifida* and *I. triloba*, are now available for hexaploid *I. batatas*. Assembled genomes together with gene annotations (32,301 annotated high confidence gene models of *I. trifida* and 31,426 of *I. triloba*) and aligned RNAseq data is available via the Genomic Tools for Sweetpotato Improvement Project, Sweetpotato Genomics Resource at Michigan State University (2018).

Genetic improvement of sweetpotato through traditional plant breeding is difficult due to its polyploid nature, genetic complexity, and high variability with regard to flower production and incompatibility (Jones et al., 1986). Generating additional (and leveraging existing) genomic resources would aid efforts to identify the molecular basis of phenotypic variation and advance the design of efficient and effective marker-assisted breeding strategies (Yoon et al., 2015). Marker-assisted breeding allows assessment of young plants at the seedling stage for multiple traits of interest, greatly reducing costs associated with growing the plants to maturity. This approach is especially valuable for sweetpotato, where the expense of long-term field evaluation is a major limiting factor in breeding efforts and where it is not feasible to conduct backcrossing breeding to introgress simple or oligogenic traits. Furthermore, genomic data provide a foundation to elucidate genetic relationships among parental lines and potentially identify new sources of genetic variation associated with environmental tolerance, pest and disease resistance, and other high-value traits. Genomic selection may also facilitate the assessment of hardiness, resistance to emerging diseases and insect pests, and changing consumer preferences. Inexpensive genome sequencing, innovative methods for the construction of nucleic acid libraries, improved mapping methodologies, and advanced computational approaches make genomics-based breeding a very attractive and powerful option for the improvement of sweetpotato.

The US sweetpotato germplasm collection is maintained by the USDA, ARS, Plant Genetic Resources Conservation Unit (PGRCU) in Griffin, Georgia, United States. This genebank maintains a diverse collection of *Ipomoea* spp. and provides clonal propagules of sweetpotato that are maintained as *in vitro* cultures. Available clones of hexaploid *I. batatas* were acquired over many decades, often in collaboration with various national and international programs and organizations. The collection provides the genetic foundation that supports ongoing research in breeding and genetics programs for the improvement of sweetpotato.

The USDA sweetpotato breeding program was initiated more than 45 years ago with the goal of developing germplasm resistant to soil insect pests while maintaining good horticultural characteristics. The germplasm used for improvement is primarily from materials developed and maintained at the USDA, ARS, US Vegetable Laboratory (USVL) in Charleston, South Carolina, United States and secondarily from the PGCRU collection. In general, resistance to insects is not associated with undesirable root quality traits in sweetpotato (Jones and Cuthbert, 1973). The program is based on recurrent mass selection using an open polycross system of 15-25 parental lines, and it relies on natural populations of bees for cross-pollination (Jones et al., 1986). To ensure that plant breeders

continue to develop improved germplasm, there is a need for comprehensively phenotyped and genotyped germplasm collections. There has recently been significant progress in the phenotyping of storage root, foliage, and growth characteristics for over 700 accessions in the PGCRU collection (Jackson et al., 2018). There appears to be ample phenotypic diversity for root and vegetative phenotypic characteristics within the PGCRU and USVL collections, but there is a lack of knowledge with regard to the level of genetic diversity within both collections.

The objective of our study was to provide information on the level of genetic diversity contained within the combined USDA (PGRCU and USVL) sweetpotato collections. A major obstacle to utilizing the recently developed genomics resources for sweetpotato and other polyploid crops has been the lack of bioinformatics tools designed specifically to handle polyploid genetic data. Here, we take advantage of recently developed resources for polyploid genotyping and analysis at the genomic level. GBSpoly, an optimized genotyping-by-sequencing protocol for highly heterozygous and polyploid genomes, and GBSapp, a SNP calling and filtering bioinformatics pipeline, were used to identify 32,784 segregating SNP markers within a collection of 417 accessions from the combined collections. The markers developed for the analysis of this collection of accessions will provide an important genomic resource to the sweetpotato community.

MATERIALS AND METHODS

Plant Materials

A total of 417 *Ipomoea batatas* accessions randomly selected from the PGCRU and USVL were examined in this study. Of the 417 accessions, 303 were from PGRCU and 114 accessions from USVL (**Table 1** and **Supplementary Table S1**). Eleven cultivars recommended for production in the southeastern US were part of the set of accessions, including Beauregard, Bayou Belle, Bellevue, Bonita, Burgundy, Carolina Ruby, Hayman White, Hernandez, Jewel, O'Henry, and Orleans. These materials originated from over 30 countries in 8 broad geographical regions (Africa, Australia, Caribbean, Central America, Far East, North America, Pacific Islands, and South America). Accessions were planted in field plots at the USVL and phenotyped for percent dry weight, periderm color, and stele color according to the methods of Jackson et al. (2018). Fresh leaf tissue was collected, placed into a labeled Ziploc® style bag, stored on ice during the field collection process, and then immediately freeze-dried. Freeze-dried leaf tissue was stored at −20°C until used.

TABLE 1 | Collection location of 417 sweetpotato (*Ipomoea batatas*) genotypes analyzed using GBSPoly.

Region (no. of accessions)	Countries (no. of accessions)
Africa (*n* = 15)	Nigeria (*n* = 14), Uganda (*n* = 1)
Australia (*n* = 2)	Australia (*n* = 2)
Caribbean (*n* = 18)	Cuba (*n* = 4), Puerto Rico (*n* = 14)
Central America (*n* = 26)	Costa Rica (*n* = 1), Guatemala (*n* = 20), Mexico (*n* = 5)
Far East (*n* = 47)	China (*n* = 16), Indonesia (*n* = 1), Japan (*n* = 11), Korea (*n* = 2), Philippines (*n* = 5), Taiwan (*n* = 10), Thailand (*n* = 1), Vietnam (*n* = 1), Unknown (*n* = 11)
North America (*n* = 210)	Canada (*n* = 1), United States (*n* = 209)
Pacific Islands (*n* = 42)	Cook Islands (*n* = 1), Fiji (*n* = 1), New Caledonia (*n* = 1), New Zealand (*n* = 12), Northern Mariana Islands (*n* = 1), Papua New Guinea (*n* = 17), Samoa (*n* = 4), Solomon Islands (*n* = 5)
South America (*n* = 44)	Brazil (*n* = 1), Columbia (*n* = 1), Ecuador (*n* = 1), Peru (*n* = 34), Uruguay (*n* = 4), Venezuela (*n* = 3)

Genotyping, SNP Calling, and Dosage Calling

Total genomic DNA was isolated from freeze dried leaf tissue using the DNeasy Plant Mini Kit (Qiagen). The integrity, purity, and concentration of the isolated genomic DNA was determined by 2% agarose gel electrophoresis and a NanoDrop 2000 spectrophotometer (ThermoFisher). A modified genotyping-by-sequencing (GBSpoly) protocol optimized for highly heterozygous and polyploid genomes was implemented (Sweetpotato Genomics Resource at Michigan State University, 2018). A complementary bioinformatic pipeline, GBSapp, was used for SNP/dosage calling and data filtering. DNA concentrations were adjusted to 50 ng/μl. A double-digest was performed using 1 μg of DNA in a total volume of 30 μl with 5 units of *Cvi*AII at 25°C for 3 h and then with 5 units of *Tse*I at 65°C for 3 h in NEB CutSmart buffer (New England Biolabs). The digested DNA samples were purified with AMPure XP magnetic beads (ThermoFisher), quantified using a picogreen assay and then diluted to a concentration of 10 ng/μl. The resulting fragments were ligated to barcoded adapters which were designed to contain an 8-bp buffer sequence positioned upstream of the variable length (6–9 bp) barcode sequence (multiplexed for 96 pooled samples). Barcode design accounted for substitution and indel errors using the levenshtein/edit distance metric (Faircloth and Glenn, 2012). The buffer sequence ensures that the barcode sequence lies within a high-quality base call region of the sequence read. Aliquots of the samples were pooled and then a secondary double-digest with *Cvi*AII and *Tse*I (same enzymes and reactions conditions above) was performed to eliminate chimeric sequence ligations. The pools were again purified with AMPure XP magnetic beads and size-selected for 300–400 bp fragments using the Blue Pippin Prep system (Sage Science). PCR amplifications were performed (18 cycles) using NEB Phusion high-fidelity polymerase (New England Biolabs). The resulting libraries were size-selected again and then sequenced on an Illumina HiSeq 2500 system.

Raw Fastq files were processed by the following steps within the GBSapp pipeline, which integrates various software tools. The steps within the pipeline included:

(i) Using the FASTx-Toolkit[1], QC plots of the raw reads were generated to ensure that the base calls within the barcodes had high quality scores.

[1]http://hannonlab.cshl.edu/fastx_toolkit/

(ii) Using the FASTx-Toolkit, the buffer sequence and any base position with low quality scores at the proximal end of the reads were trimmed to a quality score of at least Q36 for the lower whisker (minimum) in the boxplot (i.e., approximately 99.99% base calling accuracy, **Supplementary Figure S1**).

(iii) Demultiplexing of samples was performed with the FASTX-Toolkit.

(iv) Using BWA-MEM (Li, 2013), reads derived from each sample were mapped (using default parameters) to the two purported ancestral diploid reference genomes (*I. trifida* and *I. triloba*; Sweetpotato Genomics Resource at Michigan State University, 2018) of the hexaploid sweetpotato. Reads uniquely matching *I. trifida* and *I. triloba*, respectively, produced the 4x (tetraploid) and 2x (diploid) genotypes, while reads aligned to both genomes produced 6x (hexaploid) genotypes. On average, 90, 3.7, and 3.8% of the reads produced 6x, 4x, and 2x genotype calls, respectively.

(v) Additional processing of alignment files was performed with SAMtools (Li et al., 2009; Li, 2011) and Picard Tools[2].

(vi) The GATK (version 3.7) HaplotypeCaller was used to call SNP, copy number variation (CNV), and Indel variants (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013). Using VCFtools v.0.1.14 (Danecek et al., 2011), genotype calls and read depth information were extracted from the output VCF files for data filtering in R v 3.4.1 (R Core Team, 2012).

(vii) Data filtering in R was performed to identify high quality variants and bi-allelic variants. Thresholds were set for minor allele frequency (MAF) at 0.05 and missing data at no more than 20% missing. The optimal read depth for calling 2x (diploid), 4x (tetraploid), and 6x (hexaploid) genotype calls were empirically determined using a diagnostic tool written in R (R Core Team, 2012). This was accomplished by resampling (without replacement) reads from some individuals that were sequenced multiple times (replicated in multiplexed pools) to achieve very high read depth across all loci (i.e., as much as 500-2,000X coverage). Resampling was performed to capture 0.1% to 0.9% (increments of 0.1%) and 1–100% (increments of 1%) of the total reads. This simulated multiple passes of Illumina sequencing to capture each locus at various read depths. The stability of each genotype was measured across different read depths by comparing genotypes at all read depths for each locus to read depth at 100% resampling. At a 95% confidence limit, nulliplex (000000 or 111111) markers required a 1X-coverage threshold, simplex (000001 or 0111111) markers required a 35X-coverage threshold, while duplex (000011 or 0011111) and triplex (000111) markers required a 100X-coverage threshold. Aligning the reads underlying each variant back to each of the reference genomes revealed that the stable genotype calls were derived from unique sequences mapping to a single locus in the genome, while unstable genotype calls tended to map to multiple regions in the genome, suggesting that the sequence context was due to paralogs or repetitive sequences.

[2]https://broadinstitute.github.io/picard/

Data Analysis

Marker summary statistics by broad geographical region, including allele frequencies and region-specific alleles, were calculated with SVS version 8.7.0 (Golden Helix). The markers were filtered by linkage disequilibrium to generate a set of unlinked SNPs to meet assumptions for population structure analyses in STRUCTURE. Linkage disequilibrium was calculated via the expectation-maximum method (Excoffier and Slatkin, 1995) using the LD prune function of SVS with the filtering parameters set to an r^2 of 0.5, window size of 50, and window increment of 5. To determine the extent of LD decay in sweetpotato, LD was assessed by estimating r^2 values (Hill and Robertson, 1968) for all marker pairs using R base, while a plot of LD (r^2) against marker intervals (physical distance in bp) was implemented with ggplot2, a R package (R Core Team, 2012). The analyses were performed with the genotype data set that retained allelic dosage information as well as a diploidized genotype format The first five principle components of the reduced dataset were calculated in SVS using an additive model to visualize the genetic diversity across the collection (N = 417). To examine the population structure of sweetpotato accessions, they were clustered into populations with the program STRUCTURE v2.3.4 using the admixture model with correlated allele frequencies (Pritchard et al., 2000; Falush et al., 2003, 2007; Hubisz et al., 2009) using the set of 32,784 segregating SNPs. Population numbers (K) of 1 to 10 were run 10 times each with 35,000 burn-in iterations and 35,000 Markov Chain Monte Carlo repetitions. Estimation of the best K value was determined using STRUCTURE Harvester (Earl, 2012), which identifies the appropriate number of clusters (k) using the *ad hoc* statistic delta K (Evanno et al., 2005). This is based on the second order rate of change in the log probability of the data between successive values of k.

Pairwise genetic distance matrices using polymorphic markers were created in MEGA v6.06 with the p-distance model (Tamura et al., 2013). An unrooted neighbor joining phylogenetic tree was constructed in MEGA v6.06 (Tamura et al., 2013) using a pairwise genetic distance matrix of 417 accessions. The interior-branch test method (1,000 bootstrap replications) determined branch support and branches of less than 50% confidence were collapsed. FigTree v1.4.2 was used to transform the phylogenetic tree to a cladogram (Rambaut, 2014).

RESULTS

SNP Calling and Allele Dose-Dependent Genotypes

The raw fastq files generated on Illumina HiSeq2500 platform were processed to evaluate the distribution of quality scores at each base position along the sequence reads (**Supplementary Figure S1**). The quality scores were empirically determined based

on aligning reads to the Illumina's PhiX control. All the bases within the barcode region and the genomic insert had a high quality score of 38 or approximately 99.984% accuracy (i.e., median, quartiles and minimum in boxplot), except for the last base call which had a minimum score of 28 and was trimmed off (first quartile ranging from 28 to 34, an inter-quartile range of 34–38, and a median of 38). This high accuracy in the base calls of the barcode sequence is particularly important for accurate de-multiplexing of the reads derived from pooled samples. The first 5–6 bp of this 8 bp buffer sequence had slightly lower quality score (32–34), while the next 2–3 base calls had a quality score of 38. These 8 bases were trimmed off before de-multiplexing.

After aligning the raw reads to the two ancestral reference sub-genomes, on average, 96.05% of the raw reads mapped to both reference genomes. On average, 85.7% of reads matched to both sub-genomes (genotypes with 6 alleles across 6 homeologous chromosomes), 3.1% were specific to the *I. trifida* subgenome (genotypes with 4 alleles across 4 of the 6 homeologous chromosomes), and 7.2% were specific to the *I. triloba* subgenome (genotypes with 2 alleles across 4 of the 6 homeologous chromosomes). Reads specific to the *I. triloba* subgenome tended to have higher proportions than those specific to *I. trifida*-specific (**Supplementary Figure S2**). Even though more subgenome specific reads matched *I. triloba*, the *I. trifida* produced more unfiltered SNPs (496,157 against 311,409). This suggests more reads specific to the *I. triloba* subgenome might be enriched and derived from repetitive sequences. Distribution of read depth was relatively uniformly distributed across the sweetpotato accessions and across genomic loci, except for a few individuals that were underrepresented in the library (**Supplementary Figure S3**).

To empirically determine the optimal read depth threshold for accurately calling the dose-dependent genotypes, four individuals with high sequencing coverage/read depth were resampled (as described in the methods). The accuracy (or dose-dependent genotype stability) associated with the genotypic classes was determined by this resampling method. To achieve a 95% accuracy in 6x genotypes, the read depth threshold required for simplex and multi-dose genotypes was 35 and 100, respectively, while an 85% accuracy required a read depth threshold of 20 and 45. Since the data set is comprised of different dose-dependent genotypic classes that were not known *a priori*, a read threshold of 45 was used. A read depth threshold of 45 indicates almost 100% accuracy for nulliplex genotypes, over 95% accuracy for simplex genotypes and about 85% accuracy for multi-dose genotypes (**Figure 1**). Most of the genotypes were nulliplex and simplex (**Figure 2**).

Genetic Diversity and Population Structure in the Germplasm Accessions From the PGCRU and USVL

A total of 417 *I. batatas* accessions from 8 broad geographical regions (Africa, Australia, Caribbean, Central America, Far East, North America, Pacific Islands, and South America) were genotyped at 43,105 diploidized SNPs derived from GBSpoly (**Table 1** and **Supplementary Table S1**; SRA Accession SRP152827). Pruning by linkage disequilibrium resulted in a set of 32,784 segregating markers used for all analyses (**Supplementary Dataset S1**). Genetic distances between accessions (measured as the proportion of variant alleles) ranged from 0.035 to 0.41, with a mean of 0.31 and a standard deviation of 0.028 (**Supplementary Table S2**). The distribution of pairwise genetic distances had a narrow spread with only 0.1% of the comparisons less than 0.1 and 0.02% greater than or equal to 0.4. The cultivars (n = 11) recommended for production in the southeastern US had lower genetic diversity with a mean genetic distance of 0.25. The most similar cultivars, Bayou Belle and Burgundy, only differed at 4.5% of their alleles. Jewel was the most genetically distinct cultivar and was present in all of the

FIGURE 1 | Plot showing read depth thresholds and the associate SNP calling accuracy (or genotype stability with varying read depth).

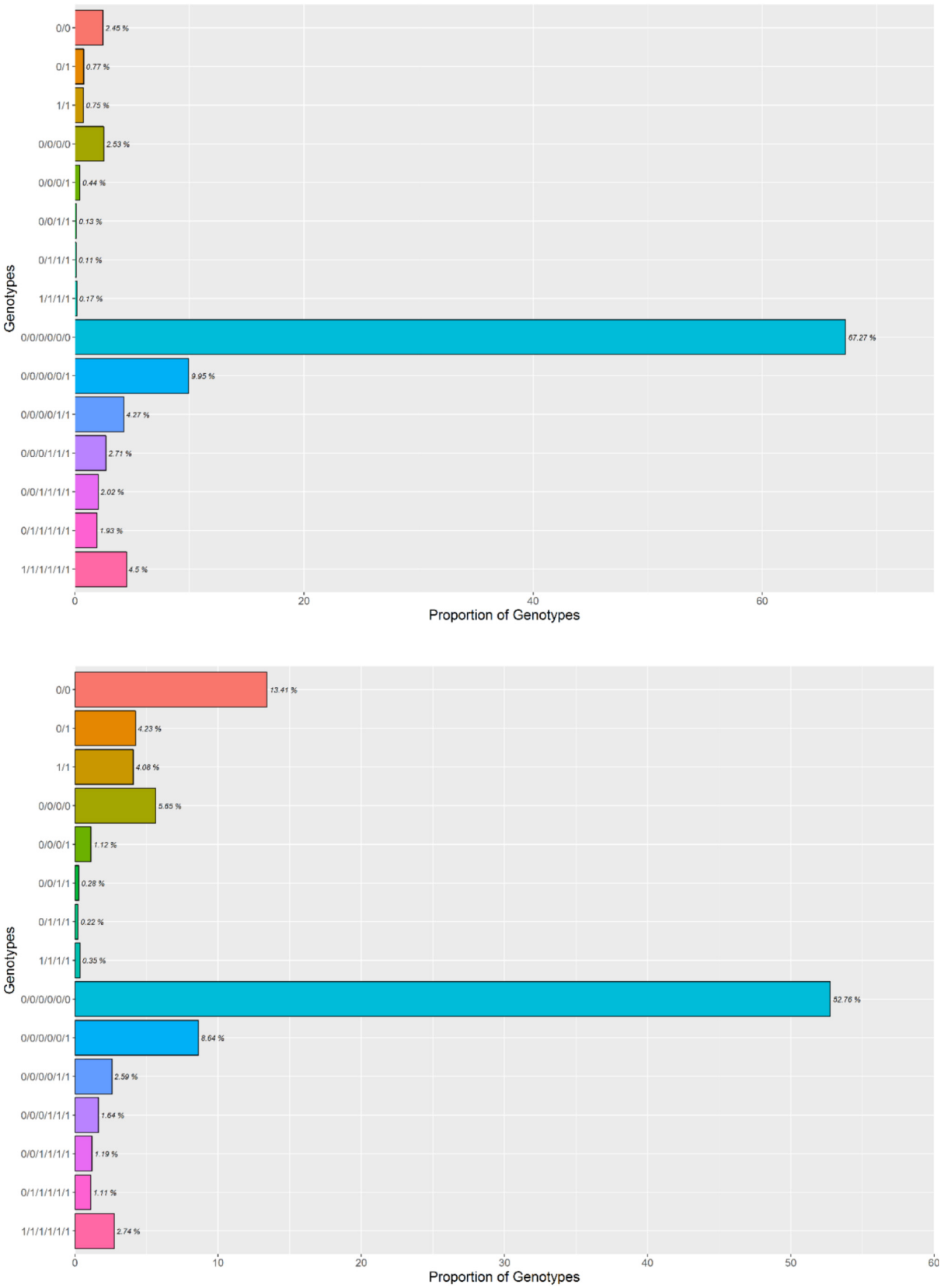

FIGURE 2 | Proportions of dose-dependent genotype calls at two different read depth thresholds of 6, 20, and 45 for 2x, 4x, and 6x genotypes, respectively, **(top)** and 6, 35, and 100 for 2x, 4x, and 6x genotypes, respectively **(bottom)**.

TABLE 2 | Minor allele frequency (MAF) for the 417 *Ipomoea batatas* accessions that were grouped by geographical region.

Broad geographical region (No. accessions)	Mean MAF	Range MAF
Africa (*n* = 15)	0.220	0–1.000
Australia (*n* = 2)	0.227	0–1.000
Caribbean (*n* = 18)	0.223	0–0.933
Central America (n = 26)	0.237	0–0.956
Far East (*n* = 55)	0.224	0–0.887
North America (*n* = 210)	0.211	0–0.730
Pacific Islands (*n* = 47)	0.221	0–0.930
South America (*n* = 44)	0.204	0–0.907
Total (*n* = 417)	0.219	0.041–0.050

top 10 most genetically distant pairwise comparisons within the cultivars (0.32–0.34). The remaining cultivar comparisons did not exceed 0.30. The overall mean MAF was 0.219 and ranged from 0.204 (South America) – 0.237 (Central America) by geographical regions (**Table 2**). Allele frequencies and Hardy-Weinberg equilibrium values for each SNP by region are provided in **Supplementary Table S3**.

To determine population structure, STRUCTURE 2.3.4 was used. The largest delta *K* peak was observed at *K* = 2 and a smaller peak was seen at *K* = 4 (**Figure 3**). We adopted the grouping for *K* = 4, because this represents a more accurate estimate of the gene pools given the historical movement of sweetpotato germplasm and that developed through crop improvement programs. There was a high degree of admixture observed in cluster assignments of the accessions with both population numbers but it was higher for *K* = 4 (**Figure 4**). When *K* = 2, the accessions clustered into two groups (South America and the other regions, **Figure 4B**). Alternatively, when *K* = 4 is considered, there is a much higher degree of admixture and assignment of accessions to clusters into four gene pools (**Figure 4**). Despite the high degree of admixture

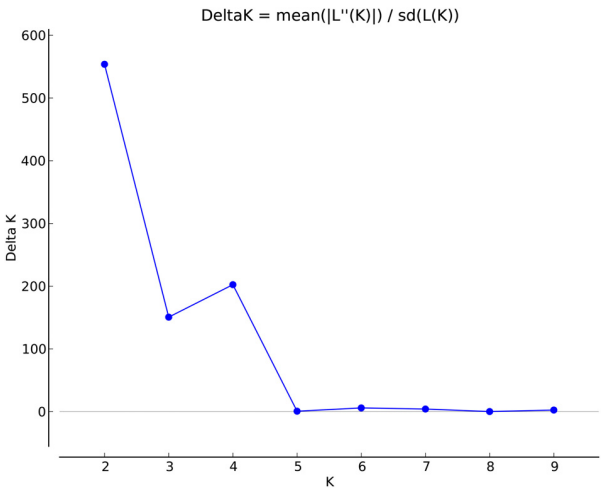

FIGURE 3 | Delta *K* values for different numbers of populations (K) assumed in STRUCTURE analysis.

within the accessions, our analyses supported dividing the 417 PIs into four clusters using a *q*-value threshold of 0.65 (**Figure 2** and **Supplementary Figure S4**), where C1 is red, C2 is green, C3 is blue, C4 is yellow, and gray is admixed (**Figure 4** and **Supplementary Figure S4**). In total, 247 out of 417 accessions (60%) were assigned to 1 of the 4 clusters. Clusters 1 through 4 consisted of 96 (23%), 49 (12%), 29 (7%), and 73 (18%) accessions, respectively. The remaining 170 accessions (40%) were categorized as having admixed ancestry from the clusters (**Supplementary Table S1**). Cluster 1 consisted mostly of accessions from Central America (*n* = 19), the Far East (*n* = 26), and the Pacific Islands (*n* = 25); cluster 2 from North America (*n* = 26) and the Pacific Islands (*n* = 9); cluster 3 from South America (*n* = 20); and cluster 4 consisted only of accessions from the North America [United States (*n* = 73)]. Accessions with mixed ancestry were found from all the studied regions: Africa (53.3%), Australia (50%), Caribbean (77.8%), Central America (15.4%), Far East (44.7%), North America (46.7%), Pacific Islands (26.2%), and South America (29.6%).

The results from the Bayesian clustering method (STRUCTURE) were further supported by population differentiation and genetic diversity analyses. Population differentiation, measured as pairwise F_{ST} values between geographical regions (**Table 3**), supported the cluster assignments for *K* = 4. The greatest differentiation was between South American and other regional groups, as all values were $\geq$ 0.063 and ranged from 0.063 (Far East) to 0.110 (Australia). Values differentiating the North American group were the greatest between it and the South American (0.103) and Central American (0.074) groups and the lowest between it and the Caribbean group (0.031). In general, differentiation was low between regions, as 17 of the 28 pairwise F_{ST} values were $\leq$ 0.05. The lowest differentiation was between the Far East and Pacific Islands (0.017) and the Far East and Australia (0.018). Further corroboration for the Bayesian clustering method was observed in the neighbor-joining (NJ) cluster analysis (**Figure 5**) and the principal components analysis [PCA (**Figure 6**)]. The NJ cladogram clearly separated the accessions into four groups, primarily based on region of collection (North American 1 and 2, South American, and the remaining regions). The PCA also separated the accessions into four groups and was consistent with the individual assignments made with STRUCTURE (**Figure 4** and **Supplementary Figure S4**). There was no apparent pattern of the accessions in the PCA by dry weight or periderm (skin) color, but there was clustering by stele (flesh) color (**Supplementary Figure S5**).

Linkage Disequilibrium in Sweetpotato

Knowing the extent of linkage disequilibrium decay in a crop is crucial for determining if genome-wide association analysis in diverse germplasm can be used for fine-mapping QTL to genic resolution. We estimated r^2 values for all marker pairs within and between chromosomes in order to empirically determine LD decay. LD analysis based on the 417 sweetpotato accessions revealed that linkage disequilibrium blocks (LD decays between 0.6 and 1.2 kb at a r^2 threshold of 0.1 and 0.2, respectively) are small enough for fine-mapping to the gene-level (**Supplementary**

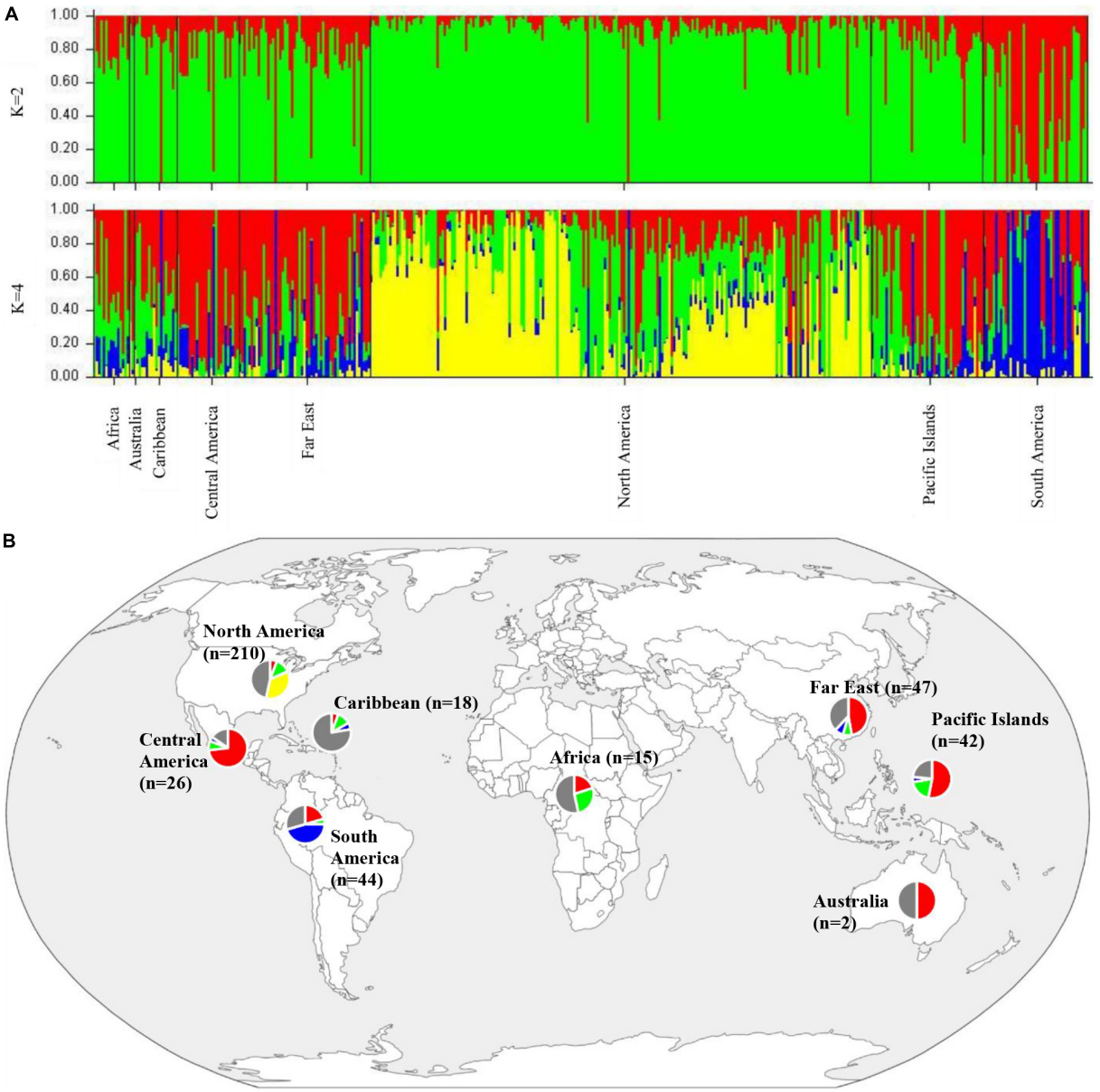

FIGURE 4 | (A) Bar plots of Bayesian assignment probabilities for each *Ipomoea batatas* accession analyzed with 32,784 SNPs using the program STRUCTURE for 2 (K = 2) or 4 clusters (K = 4). The x-axis indicates accession and the y-axis indicates the assignment probability of that accession to each of the four clusters. Each vertical line represents an individual's probability of belonging to one of K clusters (represented by different colors) or a combination of clusters if ancestry is mixed. **(B)** Map of the sampled regions for 417 *Ipomoea batatas* accessions. Pie charts correspond to the population assignment for the four genetic groups defined by the Bayesian assignment of STRUCTURE. Accessions were assigned to a cluster based on probabilities calculated in STRUCTURE, where C1 is red, C2 is green, C3 is blue, and C4 is yellow. A q-value threshold of 0.65 was used to divide the accessions into one of the four clusters or as admixed (gray section of pie charts).

Figure S6). Genotypes with allelic dosage information and the diploidized genotype data set both revealed very similar trends.

DISCUSSION

Sweetpotato is widely produced throughout the tropical regions of the world and plays a critical role in food security. Within the US, increased consumption and exports of sweetpotato have spurred substantial economic growth for producers as the value of the crop has increased by over $500 million USD between 2000 to 2015 (United States Department of Agriculture, Economic Research Service [USDA-ERS], 2016). The US sweetpotato germplasm collections maintained by the PGRCU and USVL have a general lack of genetic information, which poses challenges for germplasm curators, breeders and geneticists, entomologists, horticulturalists, and plant pathologists. In our study, two US

TABLE 3 | Pairwise F_{ST} values between broad geographical regions based on the collection location of *Ipomoea batatas* accessions.

	Africa	Australia	Caribbean	Central America	Far East	North America	Pacific Islands	South America
Africa	0.000							
Australia	0.050	0.000						
Caribbean	0.021	0.035	0.000					
Central America	0.054	0.033	0.039	0.000				
Far East	0.032	0.018	0.031	0.033	0.000			
North America	0.056	0.049	0.043	0.074	0.052	0.000		
Pacific Islands	0.031	0.030	0.026	0.034	0.017	0.050	0.000	
South America	0.084	0.110	0.068	0.093	0.063	0.103	0.087	0.000

sweetpotato germplasm collections (PGCRU and USVL) were genotyped with 32,784 SNPs to characterize genetic diversity and population structure. We found that genetic diversity and population structure were associated with geographic region and that collections had high levels of mixed ancestry (40%).

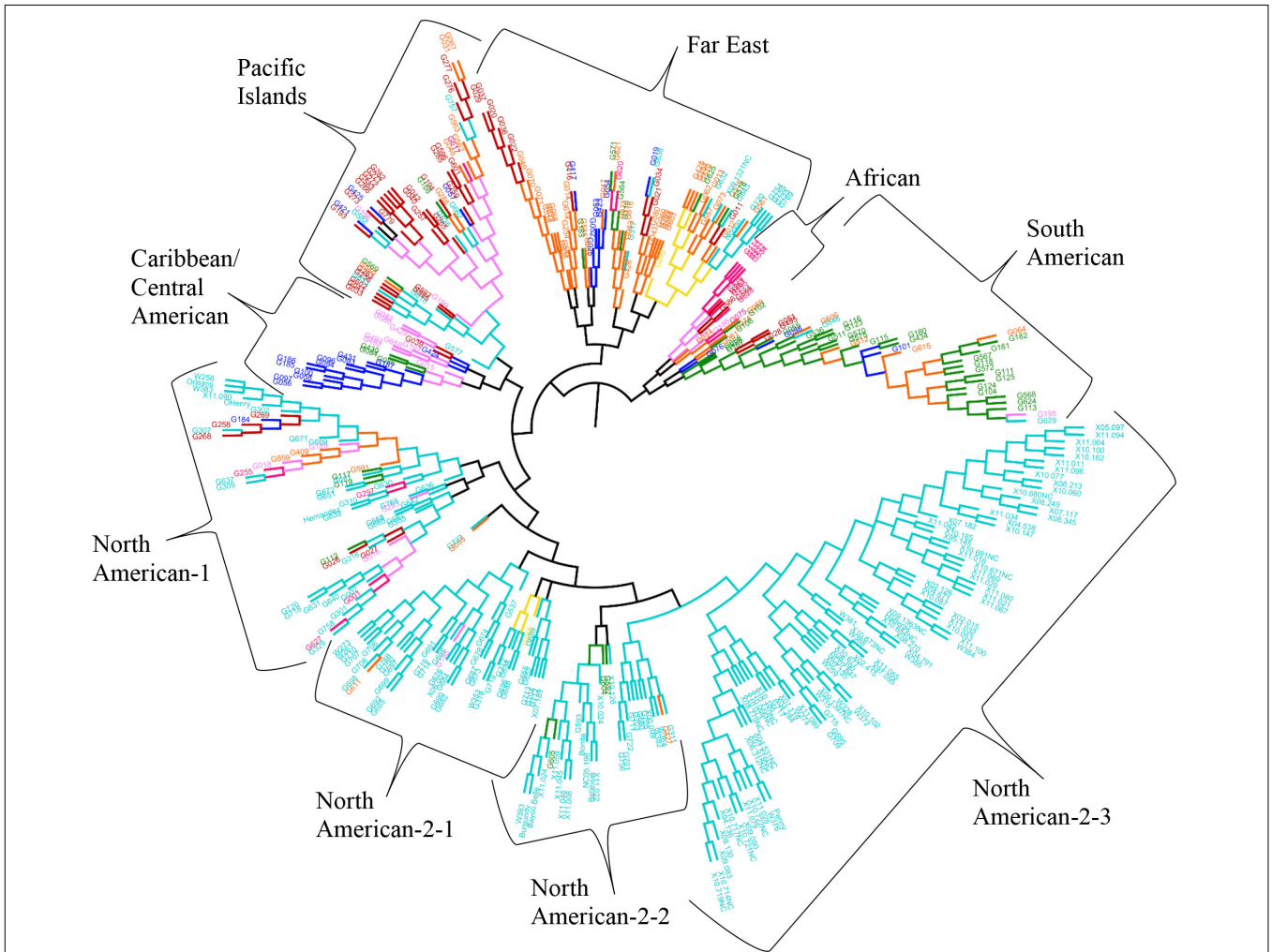

FIGURE 5 | Neighbor-joining tree based on genetic distances for 417 *Ipomoea batatas* accessions from 8 geographical regions using 32,784 single nucleotide polymorphisms. The interior-branch test method (1,000 bootstrap replications) determined branch support and branches of less than 50% confidence were collapsed. Each accession is colored according to their geographical origin, with the RGB color value used in FigTree listed in parentheses. Africa = pink (F70A7B), Australia = yellow (F7DA10), Caribbean = purple (F77AF8), Central America = blue (030EF8), Far East = orange (F96509), North America = turquoise (00CCCC), Pacific Islands = red (BC0302), South America = green (0F890E).

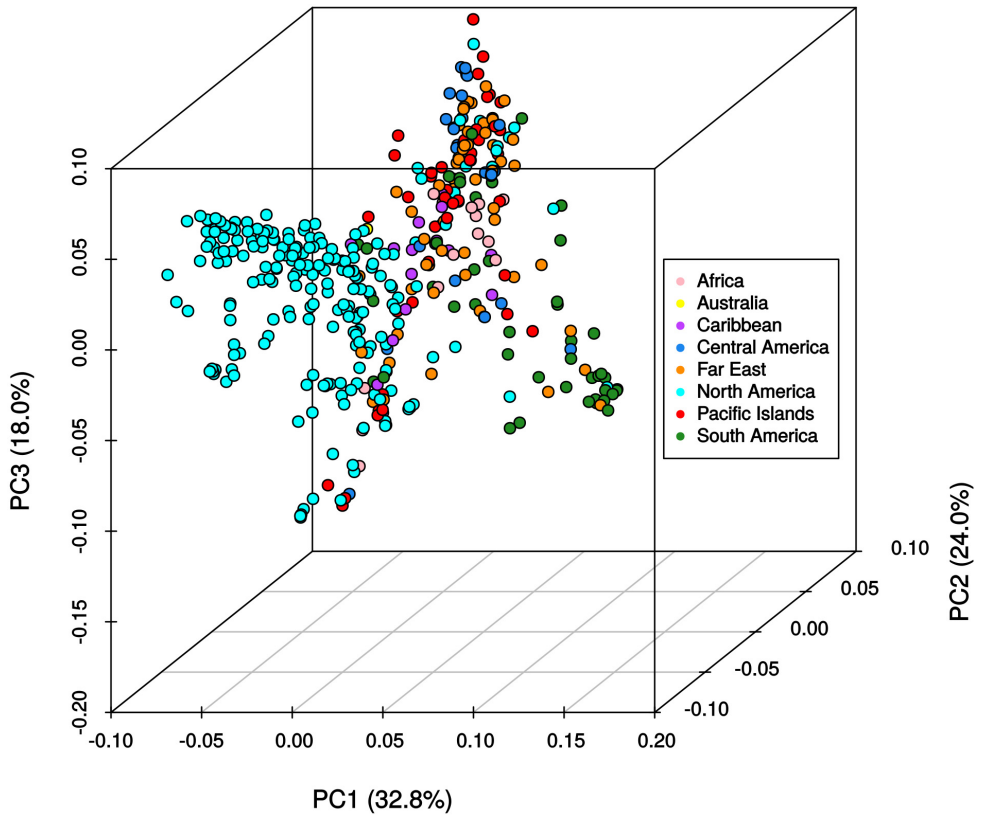

FIGURE 6 | Principle components analysis of 417 *Ipomoea batatas* accessions using 32,784 single nucleotide polymorphisms. The percent variation explained by each principle component is indicated in parentheses. Geographical regions are color-coded according to the legend.

Our analyses are the first to report the use of SNPs to characterize US sweetpotato germplasm collections and the second reported for any sweetpotato collection. Su et al. (2017) assessed diversity and population structure of 197 accessions that were mostly from China (n = 178) and found high levels of genetic diversity and support for 3 groups. Our mean minor allele frequency (0.219) was similar to their reported value (0.216). We were unable to compare the degree of admixture as it was not discussed by Su et al. (2017). However, we found a few instances of clustering of accessions into groups that did not match the geographic collection location. This can likely be explained as the result of exchange of germplasm between regions and/or incomplete or inaccurate information about these accessions in the germplasm database.

Molecular markers (AFLPs and RAPDs) have been used to examine the diversity of sweetpotato in Oceania (Zhang et al., 1998, 2004; Gichuki et al., 2003; Bruckner, 2004). These studies demonstrated that materials from the Pacific were more genetically related to materials from Mesoamerica (Mexico). More recently, chloroplast and nuclear microsatellite markers were used to demonstrate the existence of a northern and a southern gene pool of sweetpotato from Tropical America (Roullier et al., 2011). The northern gene pool was composed of materials from the Caribbean and Central America, whereas the materials in the southern gene pool were from the Peru-Ecuador region of South America. Munoz-Rodriguez et al. (2018) using whole chloroplast genomes, demonstrated the existence of two distinct sweetpotato lineages that supported the findings of Roullier et al. (2013a) using chloroplast microsatellites. Our results support the existence of multiple gene pools within the USDA sweetpotato collections. The least admixture was seen in accessions from Central America and South America and appears to support the occurrence of a northern and a southern gene pool as indicated by Roullier et al. (2011) and Munoz-Rodriguez et al. (2018). Additionally, our findings provide evidence to support the K = 3 that was reported by Roullier et al. (2013a, Figure 2 in Appendix S1) in the sweetpotato clones from tropical America. We obtained similar results for accessions collected from this region when Bayesian clustering methods were used. We also provide evidence for a third gene pool within our samples from tropical America (**Figure 4** and **Supplementary Figure S4**, green cluster). The existence of a third gene pool would account for the subset of accessions not developed by the USVL that were clustered within the materials from North America. The majority of the accessions from the Far East (55%) have probability assignments in STRUCTURE similar to the Central American accessions (**Supplementary Table S1**, **Figure 4**, and **Supplementary**

Figure S4). We speculate that this third gene pool from tropical America (Roullier et al., 2011) is the source of genetic material that has been introgressed into US sweetpotato accessions and that can be traced to the cultivar 'Porto Rico' (individual #234 in **Supplementary Figure S4**, PI 566646). 'Porto Rico' was imported into the US from Puerto Rico in 1906 and the sweetpotato industry was subsequently built around it (Harmon et al., 1970). Additional accessions that have been used in the development of the North American germplasm are PIs 153655, 208029, 399163, 399164, 153907, 296116, 344124, 566636, 286619, 286621, 308196, 318848, 318855, 318858, 324885, and 344140 (Harmon et al., 1970; Jones et al., 1991, **Supplementary Table S1** and **Supplementary Figure S4**) and introgression of these are evident within the North American accessions. We also found support of a second gene pool within the North American accessions (**Figure 4** and **Supplementary Figure S4**). We suspect that this gene pool is the result of the heavy selection pressure that has been used in the development of this germplasm (USVL collection). The W-lines and cultivars developed by the USDA sweetpotato breeding program are the result of mass selection using parents of diverse origins and this is reflected in the high level of mixed ancestry within these materials (75%). These individuals were selected for multiple insect, nematode, and disease resistance traits in combination with other desirable production and market quality traits (Jones et al., 1991). We suspect that the USVL-lines (**Supplementary Table S1** and **Supplementary Figure S4**) show a less diverse genetic base due to this material being developed through recurrent selection for insect resistance where the most resistant selections were used as parents in subsequent breeding cycles as compared to the mass selection techniques used in the development of the USVL W-lines (Jones and Dukes, 1976; Jones et al., 1986, 1991). Within the group of USVL-Lines, the introgression of material from other gene pools is evident. For example, individuals with STRUCTURE IDs 32, 41, 59, 64, 65, 70, 71, 72, and 73 in **Supplementary Figure S4** can trace their ancestry to the Uruguayan and the Louisiana State University sweetpotato breeding programs as material from these programs were used as parents in crossing blocks. Two individuals (29 and 79, **Supplementary Figure S4**) exhibit a significantly different background than the other USVL-lines. We believe that this is due to incomplete or erroneous information regarding the origins of the parents of those lines.

Germplasm collections are critical for providing genetic materials needed to ensure a continued global supply of food. Plant breeding and the associated disciplines require well characterized and readily available germplasm resources to develop crops with resistance/tolerance to pests, disease, and environmental stress. Our results indicate that there is high genetic diversity within the US sweetpotato collection and now there is the potential to utilize genotype data from our study and corresponding phenotype data (Jackson et al., 2018) for selection of a core germplasm collection. The markers developed for use with this collection of accessions provide an important genomic resource for the sweetpotato community and contribute to our understanding of the genetic diversity present in the US sweetpotato germplasm.

AUTHOR CONTRIBUTIONS

PW, BO, SB, RJ, GY, and DJ: conceived and designed the experiments. PW, BO, and SB: performed the experiments. PW, BO, SB, RJ, GY, and DJ: analyzed the data. PW, BO, SB, RJ, and GY: contributed reagents, materials, and analysis tools. PW, BO, and SB: wrote the paper. PW, BO, SB, RJ, GY, and DJ: read and approved the manuscript. PW, BO, and SB: contributed equally to this project.

FUNDING

This project was partially funded by the USDA, REE, ARS, Office of National Programs, Crop Production and Protection as Germplasm Evaluation Project No. 6080-22000-027-00D ("Genotyping-by-Sequencing of the Sweetpotato Germplasm Collection"). This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by North Carolina State University, the University of Tennessee, or the USDA. USDA is an equal opportunity provider and employer.

ACKNOWLEDGMENTS

We thank Ty Phillips, E. Parker Richardson, Craig Robertson, Sarah Moon, and Merrelyn Spinks for technical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01166/full#supplementary-material

FIGURE S1 | QC Boxplot showing distribution of quality scores of raw reads in a multiplexed library containing 96 *Ipomoea batatas* accessions. Buffer sequence lie within the first 8 base calls, while variable barcodes (6–9 bp) lie at position 14–17 bp.

FIGURE S2 | Proportion raw reads matching both reference subgenomes (6x genotypes) and those specific to each of the subgenomes (4x and 2x genotypes derived *Ipomoea trifida* and *I. triloba*, respectively).

FIGURE S3 | Boxplot shows relatively uniform read depth across individual samples and genomic loci after de-multiplexing pool samples. Only genotypes with 6 alleles/dose are shown here.

FIGURE S4 | Bar plots of Bayesian assignment probabilities for each *Ipomoea batatas* accession analyzed with segregating 32,784 SNPs using the program STRUCTURE for $K = 4$. The *x*-axis indicates accession and the *y*-axis indicates the assignment probability of that accession to each of the four clusters. Each vertical line represents an individual's probability of belonging to one of K clusters (represented by different colors) or a combination of if ancestry is mixed. The

asterisk (*) indicates the cultivar Porto Rico, which is a foundational line of the sweetpotato industry in the US. The plus sign (+) indicates that this accession was used as parental material in the mass selection populations developed by Jones et al. (1991). The USDA, ARS, US Vegetable Laboratory (USVL) W-lines and USVL-lines originate from the mass selection populations. Information for all accessions is found in **Supplementary Table S1**.

FIGURE S5 | Linkage disequilibrium estimates (r^2) of all genome-wide marker pairs plotted against corresponding interval between marker pairs. Curve (blue line) based on game smoothing method function shows distribution of all data points. Top and middle plot based on genotype data with allelic dosage information, while bottom plot is based on diploidized genotypes.

TABLE S1 | Information of *Ipomoea batatas* accessions analyzed by GBSpoly.

TABLE S2 | Pairwise genetic distance matrix between *Ipomoea batatas* accessions.

TABLE S3 | Information for individual SNPs used for data analyses.

DATASET S1 | Structure data file for *Ipomoea batatas* accessions.

REFERENCES

Austin, D. F. (1988). "The taxonomy, evolution and genetic diversity of sweet potatoes and related wild species," in *Proceedings of the Exploration, Maintenance and Utilization of Sweet Potato Genetic Resources: report of the First Sweet Potato Planning Conference 1987* (Lima: International Potato Center), 27–59.

Bruckner, A. W. (2004). *AFLP-based Genetic Diversity Assessment of Global Sweetpotato* (*Ipomoea batatas* (L.) Lam.) *Germplasm Resources: Progress Toward the Development of a Sweetpotato Core Collection*. Master Thesis, North Carolina State University, Raleigh, NC, USA.

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

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., and et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Earl, D. A. (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

Effendy, J., LaBonte, D., and Biasakh, N. (2013). Identifcation and expression of skinning injury-responsive genes in sweetpotato. *J. Am. Soc. Hortic. Sci.* 138, 210–216.

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

Excoffier, L., and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 12, 921–927.

Faircloth, B. C., and Glenn, T. C. (2012). Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One* 7:e42543. doi: 10.1371/journal.pone.0042543

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.

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Resour.* 7, 574–578. doi: 10.1111/j.1471-8286.2007.01758.x

Firon, N., LaBonte, D., Villordon, A., Kfir, Y., Solis, J., Lapis, E., et al. (2013). Transcriptional profiling of sweetpotato (*Ipomoea batatas*) roots indicates down-regulation of lignin biosynthesis and up-regulation of starch biosynthesis at an early stage of storage root formation. *BMC Genomics* 14:460. doi: 10.1186/1471-2164-14-460

Freyre, R., Iwanaga, M., and Orjeda, G. (1991). Use of *Ipomoea trifida* (HBK.) G. Don germplasm for sweet potato improvement. Part 2. Fertility of synthetic hexaploids and triploids with 2n gametes of I. trifida and their interspecific crossability with sweet potato. *Genome* 34, 209–214. doi: 10.1139/g91-033

Gichuki, S. T., Berenyi, M., Zhang, D., Hermann, M., Schmidt, J., Glössl, J., et al. (2003). Genetic diversity in sweetpotato [*Ipomoea batatas* (L.) Lam.] in relationship to geographic sources as assessed with RAPD markers. *Genetic Resour. Crop Evol.* 50, 429–437. doi: 10.1023/A:1023998522845

Harmon, S. A., Hammett, H. L., Hernandez, T., and Pope, D. T. (1970). "Progress in the breeding and development of new varieties", in *Thirty Years of Cooperative Sweetpotato Research 1939 – 1969*, ed. T.P. Hernandez (Baton Rouge, LA: Southern Cooperative Series Bulletin No. 369, Louisiana Agricultural Experiment Station), 8–17.

Hill, W. A., Mortley, D. G., MacKowiak, C. L., Loretan, P. A., Tibbitts, T. W., Wheeler, R. M., et al. (1992). Growing root and tuber crops hydroponically. *Adv. Space Res.* 12, 125–131. doi: 10.1016/0273-1177(92)90018-S

Hill, W. G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38, 226–231. doi: 10.1007/BF01245622

Hubisz, M. J., Falush, D., Stephens, M., and Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* 9, 1322–1332. doi: 10.1111/j.1755-0998.2009.02591.x

International Potato Center (2018). *Sweetpotato Facts and Figures*. Available at: https://cipotato.org/crops/sweetpotato/sweetpotato-facts-and-figures/ [accessed March 27, 2018].

Sweetpotato Genomics Resource at Michigan State University (2018). *GT4SP Project Ipomoea Genome Pseudomolecules and Annotation – Version 3*. Available at: http://sweetpotato.plantbiology.msu.edu/gt4sp_download.shtml [accessed March 27, 2018].

Jackson, D. M., Harrison, H. F., Jarret, R. L., and Wadl, P. A. (2018). Color analysis of storage roots from the USDA, ARS sweetpotato (*Ipomoea batatas*) germplasm collection. *Genet. Resour. Crop Evol.* 65, 1217–1236. doi: 10.1007/s10722-018-0609-6

Jones, A., and Cuthbert, P. F. (1973). Associated effects of mass selection for soil insect-resistance in sweet potato. *J. Am. Soc. Hortic. Sci.* 98, 480–482.

Jones, A., and Dukes, P. D. (1976). Some seed, seedling and maternal characters as estimates of commercial performance in sweet potato breeding. *J. Amer. Soc. Hort. Sci.* 101, 385–388.

Jones, A., Dukes, P. D., and Schalk, J. M. (1986). "Sweet potato breeding", in *Breeding Vegetable Crops*, ed. M. J. Bassett (Westport, CT: AVI Publishing Co.,), 1–35.

Jones, A., Dukes, P. D., Schalk, J. M., and Hamilton, M. G. (1991). 1/13 and J/8 sweetpotato mass selection populations. *HortScience* 26, 929–930.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:*1303.3997v1. Available at: https://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Magoon, M. L., Krishnar, R., and Bai, K. V. (1970). Cytological evidence on the origin of sweet potato. *Theor. Appl. Genet.* 10, 360–366. doi: 10.1007/BF00285415

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., and 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

Munoz-Rodriguez, P., Carruthers, T., Wood, J. R. I., Williams, B. R. M., Weitemier, K., Kronmiller, B., et al. (2018). Reconciling conflicting phylogenies in the origin of sweet potato and dispersal to Polynesia. *Curr. Biol.* 28, 1–11. doi: 10.1016/j.cub.2018.03.020

Oracion, M. Z., Niwa, K., and Shiotani, I. (1990). Cytological analysis of tetraploid hybrids between sweet potato and diploid *Ipomoea trifida* (H.B.K.) Don. *Theor. Appl. Genet.* 80, 617–624. doi: 10.1007/BF00224220

Orjeda, G., Freyre, R., and Iwanaga, M. (1989). Production of 2m pollen in diploid *Ipomoea trifida*, a putative wild ancestor of sweet potato. *Heredity* 81, 462–467. doi: 10.1093/oxfordjournals.jhered.a111026

Padmaja, G. (2009). "Uses and nutritional data of sweet potato", in *The Sweetpotato*, eds G. Loebenstein and G. Thottapilly (New York, NY: Springer), 189–234. doi: 10.1007/978-1-4020-9475-0_11

Ponniah, S. K., Thimmapuram, J., Bhide, K., Kalavacharla, V., and Manoharan, M. (2017). Comparative analysis of the root transcriptomes of cultivated sweetpotato (*Ipomoea batatas* [L.] Lam) and its wild ancestor (*Ipomoea trifida* [Kunth] G. Don). *BMC Plant Biol.* 17:9. doi: 10.1186/s12870-016-0950-x

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

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.r-project.org/

Rajapakse, S., Nilmalgoda, S. D., Molnar, M., Ballard, R. E., Austin, D. F., and Bohac, J. R. (2004). Phylogenetic relationships of the sweetpotato in *Ipomoea* series *Batatas* (Convolvulaceae) based on nuclear β-amylase gene sequences. *Mol. Phylogenet. Evol.* 30, 623–632. doi: 10.1016/S1055-7903(03)00249-5

Rambaut, A. (2014). *FigTree, a Graphical Viewer of Phylogenetic Trees*. Available at: http://tree.bio.ed.ac.uk/software/figtree/ [accessed March 27, 2018].

Roullier, C., Benoit, L. B., McKey, D. B., and Lebot, V. (2013a). Historical collections reveal patterns of diffusion of sweet potato in Oceania obscured by modern plant movements and recombination. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2205–2210. doi: 10.1073/pnas.1211049110

Roullier, C., Duputié, A., Wennekes, P., Benoit, L., Fernandez Bringas, V. M., Rossel, G., et al. (2013b). Disentangling the origins of cultivated sweet potato (*Ipomoea batatas* (L.) Lam.). *PLoS One* 8:e62707. doi: 10.1371/journal.pone.0062707

Roullier, C., Rossel, G., Tay, D., McKey, D., and Lebot, V. (2011). Combining chloroplast and nuclear microsatellites to investigate origin and dispersal of New World sweet potato landraces. *Mol. Ecol.* 20, 3963–3977. doi: 10.1111/j.1365-294X.2011.05229.x

Schafleitner, R., Tincopa, L. R., Palomino, O., Rossel, G., Robles, R. F., and et al. (2010). A sweetpotato gene index established by de novo assembly of pyrosequencing and Sanger sequences and mining for gene-based microsatellite markers. *BMC Genomics* 11:604. doi: 10.1186/1471-2164-11-604

Shiotani, I. (1987). "Genomic structure and the gene flow in sweet potato and related species", in *Exploration, Maintenance, and Utilization of Sweet Potato Genetic Resources*, ed P. Gregory (Lima: Report 1st Sweet Potato Planning Conference, CIP), 61–73.

Solis, J., Baisakh, N., Villordon, A., and LaBonte, D. (2016). Transcriptome profiling of beach morning glory (*Ipomoea imperati*) under salinity and its comparative analysis with sweetpotato. *PLoS One* 11:e0147398. doi: 10.1371/journal.pone.0147398

Solis, J., Villordon, A., Baisakh, N., LaBonte, D., and Firon, N. (2014). Effect of drought on storage root development and gene expression profile of sweetpotato under greenhouse and field conditions. *J. Am. Soc. Hortic. Sci.* 139, 317–324.

Su, W., Wang, L., Lei, J., Chai, S., and Liu, Y. (2017). Genome-wide assessment of population structure and genetic diversity and development of a core germplasm set for sweet potato based on specific length amplified fragment (SLAF) sequencing. *PLoS One* 12:e0172066. doi: 10.1371/journal.pone.0172066

Tamura, K., Stecher, G., Peterson, D. A., and Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tao, X., Gu, Y. H., Wang, H. Y., Zheng, W., Li, X., Zhao, C. W., et al. (2012). Digital gene expression analysis based on integrated de novo transcriptome assembly of sweet potato [*Ipomoea batatas* (L.) Lam]. *PLoS One* 7:e36234. doi: 10.1371/journal.pone.0036234

Techen, N., Arias, R. S., Glynn, N. C., Pan, Z., Khan, I. A., and Scheffler, B. E. (2009). Optimized construction of microsatellite-enriched libraries. *Mol. Ecol. Resour.* 10, 508–515. doi: 10.1111/j.1755-0998.2009.02802.x

United States Department of Agriculture, Economic Research Service [USDA-ERS] (2016). *Production of Sweet Potatoes – a Thanksgiving Favorite – Is on the Rise*. Available at: https://www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=80859 [accessed on March 27, 2018].

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., and del Angel, G. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformat.* 43, 11.10.1–11.10.33. doi: 10.1002/0471250953.bi1110s43

Wang, S., Nie, S., and Zhu, F. (2016). Chemical constituents and health effects of sweet potato. *Food Res. Int.* 89, 90–116. doi: 10.1016/j.foodres.2016.08.032

Wang, Z., Fang, B., Chen, J., Zhang, X., Luo, Z., and et al. (2010). De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). *BMC Genomics* 11:726. doi: 10.1186/1471-2164-11-726

Xie, F., Burklew, C. E., Yang, Y., Liu, M., Xiao, P., Zhang, B., et al. (2012). De novo sequencing and a comprehensive analysis of purple sweet potato (*Ipomoea batatas* L.) transcriptome. *Planta* 236, 101–113. doi: 10.1007/s00425-012-1591-4

Yang, J., Moeinzadeh, M. H., Kuhl, H., Helmuth, J., Xiao, P., Liu, G., et al. (2016). The haplotype-resolved genome sequence of hexaploid Ipomoea batatas reveals its evolutionary history. *bioRxiv* [Preprint]. doi: 10.1101/064428

Yoon, J.B., Kwon, S.W., Ham, T.H., Kim, S., Thomson, M., Hechanova, S.L., et al. (2015). "Marker-Assisted Breeding", in *Current Technologies in Plant Molecular Breeding*, eds H. J. Koh, S. Y. Kwon, and M. Thomson (Dordrecht: Springer), 95–144. doi: 10.1007/978-94-017-9996-6_4

Zhang, D., Ghislain, M., Huam, Z., Golmirzaie, A., and Hijmans, R. (1998). RAPD variation in sweetpotato (*Ipomoea batatas* (L.) Lam.) cultivars from South America and Papua New Guinea. *Gen. Resour. Crop Evol.* 45, 271–277. doi: 10.1023/A:1008642707580

Zhang, D., Rossel, G., Kriegner, A., and Hijmans, R. (2004). AFLP assessment of diversity in sweet-potato from Latin America and the Pacific region: its implications on the dispersal of the crop. *Genetic Resour. Crop Evol.* 51, 115–120. doi: 10.1023/B:GRES.0000020853.04508.a0

Zhou, C., Olukolu, B., Gemenet, D., Wu, S., Gruneberg, W., Cao, M. D., et al. (2017). Assembly of whole-chromosome pseudomolecules for polyploid plant genomes using outcrossed mapping populations. *bioRxiv* [Preprint]. doi: 10.1101/119271

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 August 2018
doi: 10.3389/fpls.2018.01220

Genomic Prediction for 25 Agronomic and Quality Traits in Alfalfa (*Medicago sativa*)

Congjun Jia[1†], Fuping Zhao[1†], Xuemin Wang[1], Jianlin Han[2,3], Haiming Zhao[4], Guibo Liu[4] and Zan Wang[1]*

[1] Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China, [2] CAAS-ILRI Joint Laboratory on Livestock and Forage Genetic Resources, Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China, [3] International Livestock Research Institute (ILRI), Nairobi, Kenya, [4] Institute of Dryland Farming, Hebei Academy of Agriculture and Forestry Sciences, Hengshui, China

Agronomic and quality traits in alfalfa are very important to forage industry. Genomic prediction (GP) based on genotyping-by-sequencing (GBS) data could shorten the breeding cycles and accelerate the genetic gains of these complex traits, if they display moderate to high prediction accuracies. The aim of this study was to investigate the predictive potentials of these traits in alfalfa. A total of 322 genotypes from 75 alfalfa accessions were used for GP of the agronomic and quality traits, which were related to yield and nutrition value, respectively, using BayesA, BayesB, and BayesCπ methods. Ten-fold cross validation was used to evaluate the accuracy of GP represented by the correlation between genomic estimated breeding value (GEBV) and estimated breeding value (EBV). The accuracies ranged from 0.0021 to 0.6485 for different traits. For each trait, three GP methods displayed similar prediction accuracies. Among 15 quality traits, mineral element Ca had a moderate and the highest prediction accuracy (0.34). NDF digestibility after 48 h (NDFD 48 h) and 30 h (NDFD 30 h) and mineral element Mg had prediction accuracies varying from 0.20 to 0.25. Other traits, for example, fat and crude protein, showed low prediction accuracies (0.05 to 0.19). Among 10 agronomic traits, however, some displayed relatively high prediction accuracies. Plant height (PH) in fall (FH) had the highest prediction accuracy (0.65), followed by flowering date (FD) and plant regrowth (PR) with accuracies at 0.52 and 0.51, respectively. Leaf to stem ratio (LS), plant branch (PB), and biomass yield (BY) reached to moderate prediction accuracies ranging from 0.25 to 0.32. Our results revealed that a few agronomic traits, such as FH, FD, and PR, had relatively high prediction accuracies, therefore it is feasible to apply genomic selection (GS) for these traits in alfalfa breeding programs. Because of the limitations of population size and density of SNP markers, several traits displayed low accuracies which could be improved by a bigger reference population, higher density of SNP markers, and more powerful statistic tools.

Keywords: alfalfa, genomic prediction, agronomic trait, quality trait, estimated breeding value

OPEN ACCESS

Edited by:
Yiwei Jiang,
Purdue University, United States

Reviewed by:
Quanzhen Wang,
Northwest A&F University, China
Lan Zhu,
Oklahoma State University, United States

****Correspondence:***
Zan Wang
wangzan@caas.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *14 March 2018*
Accepted: *30 July 2018*
Published: *20 August 2018*

Citation:
Jia C, Zhao F, Wang X, Han J, Zhao H, Liu G and Wang Z (2018) Genomic Prediction for 25 Agronomic and Quality Traits in Alfalfa (Medicago sativa). Front. Plant Sci. 9:1220. doi: 10.3389/fpls.2018.01220

INTRODUCTION

Alfalfa (*Medicago sativa* L) is the first most-important forage legume in the world, because of its high biomass yield (BY) and good nutritional quality. To meet the future demand of quantity and quality, the main objectives in alfalfa breeding programs are biomass related agronomic traits and nutrition value related quality traits. Though yield and quality of alfalfa have been improved by phenotypic selection, the genetic gain are relatively low compared to other crops, owing to many reasons, such as low heritability, complex genetic architecture, and high genotype-environment interaction (Annicchiarico et al., 2015a). Therefore, it is emergent that new breeding strategies should be introduced into alfalfa breeding programs to accelerate the genetic gain of targeted traits and thus to meet the increasing demands of forage production.

Breeding value (BV), known as genetic merit of an individual which cannot be measured directly, is always the key issue in plant breeding programs. However, accurately estimated breeding value (EBV) is impossible to be achieved in complex traits by using phenotypic data alone. To improve the accuracy of prediction, incorporating information of genetic markers, known as marker-assisted selection (MAS), is an optional strategy. The superiority of MAS than phenotypic selection is determined by the percentage of the genetic variance accounted for by the QTLs associated with the markers (Meuwissen and Goddard, 1996). Unfortunately, the proportion of variation in complex traits explained by significant markers is usually very small (Hayes and Goddard, 2010). Therefore, many markers in linkage disequilibrium (LD) with QTLs contributed to targeted traits are needed to realize a relatively high prediction accuracy.

Due to the decreased cost of high-throughput genotyping methods, huge amount of genomic information of many non-model plants has been produced. Utilization of genotypic information in plant breeding has become a highly prioritized research area in recent years. Since dense genetic markers covering whole genome are available in many species, a new method for estimating breeding value, namely the genomic selection (GS) or genomic prediction (GP), showed a great potential for enhancing the accuracy of GP of BV (Meuwissen et al., 2001). It is assumed that all genes, with either large or small effects, affecting targeted traits are in LD with some markers that are distributed across the genome, paving the way to achieve a high accuracy of genomic estimated breeding value (GEBV) (Meuwissen, 2007). In a simulation study, the accuracy could be as high as 0.85 (Meuwissen et al., 2001). But this is not always the case in the real data. Several studies on GP have been done in wheat (Lado et al., 2013; Jiang et al., 2017; Sukumaran et al., 2017), maize (Riedelsheimer et al., 2012; Crossa et al., 2013; Pace et al., 2015), and other plants (Shu et al., 2013; Xu et al., 2014; Grenier et al., 2015), revealing a majority of the prediction accuracies between 0.05 and 0.8, depending on the traits, statistical methods, and experiment designs.

As mentioned above, GP can significantly improve the accuracy of estimation of breeding value. Therefore, it attracts a great interest of plant breeders worldwide. Traits being targeted in plant breeding programs are either difficult or costly to be measured. Additionally, the targeted traits (e.g., yield, phenology, and adaptation to stress) in plant breeding are mostly quantitative traits, which are controlled by multiple genes and generally sensitive to environmental variables. Phenotypic selection, neglecting the underlying biological processes and the interactions between genes and environments, cannot make a significant genetic gain in a short time frame. Considering the genetic architecture of the quantitative traits, MAS is also not the best choice. GP, following its assumption, is thus an ideal tool to be used in the plant breeding programs. Many methods have been adopted for GP or GS. Bayesian methods and GBLUP, however, are those being frequently used. Bayesian methods exhibited more advances than GBLUP in terms of prediction accuracy following a simulation study (Meuwissen et al., 2001). No matter which method is used for GP, the density of markers across the whole genome is a determining factor. Typically, two types of high throughput genotyping methods of SNP array and whole-genome re-sequencing can be employed to generate high quality genotypes of markers. For important crop species, several SNP Bead chips at different marker densities have been developed (Ganal et al., 2012). Because of the lack of SNP array, genotyping by sequencing (GBS) is therefore an alternative to alfalfa genotyping. In the current study, we investigated the impact of three Bayes statistical methods on the prediction accuracies of alfalfa agronomic and quality traits with genotypic data obtained by GBS.

MATERIALS AND METHODS

Plant Materials and Experimental Designs

The alfalfa materials used in this study were consisted of 322 genotypes representing 75 tetraploid alfalfa accessions under the experimental designs as described in Wang et al. (2016).

Phenotypic Data Collection and Analysis

A total of 25 traits (**Table 1**), including 15 quality and 10 agronomic traits, were measured for all genotypes. All the plants were harvested at early flowering stage and prepared to measure the 15 quality traits using a FOSS 5000 scanning monochromator (FOSS, Denmark). The 15 quality traits included three fiber-related traits, four digestibility-related traits, and eight nutrition component traits being measured following the procedures described in our previous studies (Wang et al., 2016; Jia et al., 2017). Before harvesting, plant height (PH) of each plot was measured as nature height on every plant. Plant branch (PB) was measured as the number of primary branches arising from the main stem. The number of main stem node (SN) for each plot was directly counted since the first node on the main stem from every plant. The first inflorescence position (FP) was measured as the position of the first inflorescence on the stem. After harvesting, BY was measured as the fresh weight by clipping all six plants in each plot at a uniform height of 5 cm. The stems and leaves

TABLE 1 | Prediction accuracies of 25 traits.

	Prediction accuracies		
Traits	**BayesA**	**BayesB**	**BayesCπ**
ADF	0.1847	0.1805	0.1824
aNDF	0.1843	0.1942	0.1958
dNDF30	0.2002	0.1933	0.1963
dNDF48	0.2538	0.2545	0.249
ADL	0.0783	0.0756	0.08
IVTDMD30h	0.0907	0.0906	0.0889
IVTDMD48h	0.0994	0.1029	0.1002
CP	0.0528	0.0524	0.0557
RUP	0.0828	0.0909	0.0746
Ash	0.087	0.0836	0.085
Ca	0.337	0.3422	0.3393
K	0.1663	0.1586	0.1572
Mg	0.2184	0.2183	0.2178
P	0.1178	0.1196	0.1203
Fat	0.1148	0.1203	0.1114
BY	0.2418	0.2511	0.2457
DM	0.1285	0.1257	0.1271
FD	0.5153	0.5139	0.5131
FH	0.6485	0.6466	0.6451
FP	0.0639	0.0596	0.0683
LS	0.3214	0.3215	0.3249
PB	0.2589	0.2598	0.2593
PH	0.1587	0.1601	0.1626
PR	0.5105	0.5111	0.5074
SN	0.0045	0.0047	0.0021

Abbreviations of traits are explained in Materials and Methods.

were separated and placed into a nylon net bag, naturally air-dried, and weighed separately to calculate the leaf to stem ratio (LS). Meanwhile, dry matter (DM) was defined as the sum of the weights of stems and leaves. Plant regrowth (PR) was measured as the PH two weeks after the first harvest. Flowering date (FD) was calculated by the date of opening of the first flower for the first two growth cycles. PH in fall (FH) was measured as the PH 21 days after the last harvest. The mean value of all six plants in each plot represents the trait value of a genotype grown in that plot. The measurements of all traits were performed on all genotypes under three consecutive years (2013, 2014, and 2015).

Linear mixed model was fitted to calculate the BLUP value and EBV for individual trait of each genotype as follows:

$$y_i = \mu + g_i + e_i + \varepsilon_i.$$

In this equation, y_i represents the phenotype of the *i*th genotype, μ is the grand mean value of the targeted trait in all environments, g_i is denoted as genetic effect, e_i is the environmental effect, and ε_i is the random error. The BLUP value was estimated for individual trait of each genotype based on the above-mentioned linear model using the lme4 model (Bates et al., 2011). The EBV of individual genotype was used as response value in GP equation to estimate marker effect.

DNA Isolation, GBS Library Construction, Sequencing, and Genotypic SNP Calling

Leaf tissues were collected from all genotypes and DNAs were extracted using the Qiagen DNeasy 96 Plant kit (Qiagen, CA, United States). DNA degradation and contamination were monitored on 1% agarose gels. DNA purity and concentration were checked using the NanoPhotometer® spectrophotometer (IMPLEN, CA, United States) and Qubit® DNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA, United States), respectively. DNA was digested by *Mse*I [New England Biolabs (NEB)] restriction enzyme. The reduced representation libraries were constructed for individual genotypes according to published GBS protocol (Elshire et al., 2011) and sequenced using Illumina HiSeq2000 platform. Raw data were submitted to the NCBI Sequence Read Archive with a reference number of SRP081825. The Tassel 3.0 Universal Network Enabled Analysis Kit (UNEAK) pipeline (Lu et al., 2013) was used for *de novo* SNP discovery and genotype calling following Li et al. (2014).

SNP Imputation

After SNP calling, NPUTE was used to impute the GBS data (Roberts et al., 2007).

Statistical Methods for GP

Three regression methods with different prior assumptions of the distribution of marker effects were used to estimate SNP effects, namely the BayesA (Meuwissen et al., 2001), BayesB (Meuwissen et al., 2001), and BayesCπ (Habier et al., 2011). A ten-fold cross validation was used to evaluate the accuracy of GP. The data were randomly split into 10 approximately equal-sized groups. For each cross validation, nine groups were used as the training population to estimate parameters and the remaining group (validation population) was used as the test sample. The linear model is denoted as follows:

$$y_i = \mu + \sum_{j=1}^{m} Z_{ij}\alpha_j + e_i$$

where, y_i is the EBV of one trait, μ is the overall mean, m is the number of markers, Z_{ij} is the *jth* SNP genotype of plant i, α_j is the average effect of allele substitution for SNP j, and e_i is the residual error with an assumed normal distribution N $(0,\sigma^2_e)$. SNP effects were estimated based on the training population using this equation. The GEBV for plant i in the validation population was predicted by summing up SNP effects over all loci. Predictive accuracy was measured as the correlation between the EBVs and GEBVs. Random sampling training and validation sets were repeated 10 times and the mean of correlations was calculated to measure the GP accuracy. All Bayes programs were run in BGLR package in R environment.[1] The number of Burn-in was 10000, thin was 20, and the total number of iteration was 30,000. Other priors of parameters were assigned following Perez and de los Campos (2014).

[1] http://www.r-project.org

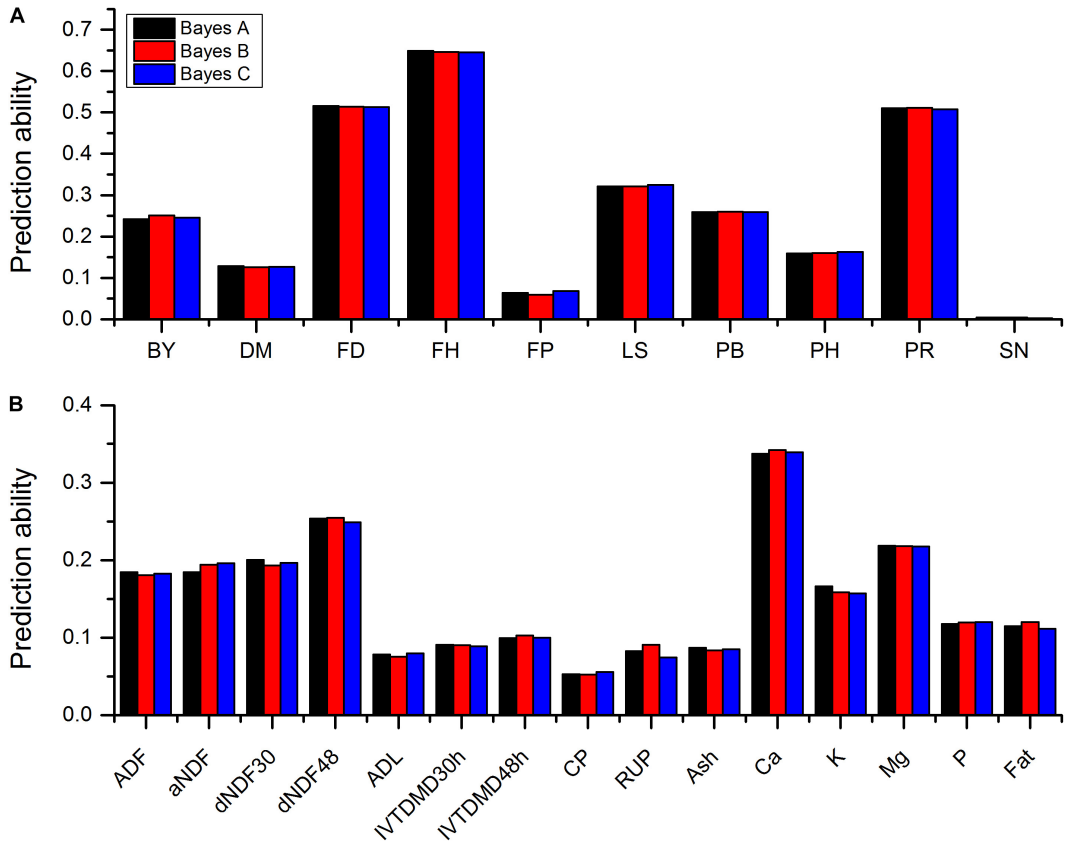

FIGURE 1 | Predictabilities of 10 agronomic traits **(A)** and 15 quality traits **(B)** plotted against three Bayesian methods. Different colors represent different methods. Abbreviations of traits are explained in Materials and Methods.

RESULTS

Phenotypic Variation

Since our previous works have described the phenotypic variations of some fiber-related traits (Wang et al., 2016) and crude protein and mineral elements (Jia et al., 2017), we will not describe them in this study. Instead, we want to represent the EBV variations of all traits incorporated in this study. The frequency distributions of EBVs for all 25 traits were symmetric as shown in **Supplementary Figure S1**.

GP Using Three Bayesian Methods

Sequencing of the GBS libraries yielded approximately 184.59 million raw reads and 178.2 million clean reads in all 322 alfalfa genotypes. After imputation, 44,757 high quality SNPs were obtained and used for GP. The results of prediction accuracies of three Bayesian methods are shown in **Table 1**. The predictabilities drawn from the ten-fold cross validation varied across different traits. SN had the lowest predictability (0.0021) but FH had the highest predictability (0.6485). Some quality traits such as crude protein (CP), RUP, and ADL exhibited relatively low prediction accuracies (< 0.1) while the remaining quality traits such as fat, K, and Ca showed low to moderate predictabilities (0.11−0.34). Agronomic traits hold similar patterns except three traits that had relatively high predictabilities with FH to be the highest (0.65), followed by FD (0.52), and PR (0.51). Other traits, such as LS, PB, and BY displayed moderate predictabilities (0.24−0.32). Similar to BayesA method, BayesB and BayesCπ methods did not reveal any significant difference from each other in terms of the predictabilities of all quality and agronomic traits (**Table 1** and **Figure 1**). The predictabilities among the three Bayesian methods are shown in **Figure 1**. From the bar-plotting, only minor differences were observed among the three methods for all 25 traits, it was therefore hard to determine which method was the best.

DISCUSSION

Since GS was proposed by Meuwissen et al. (2001), many studies have been conducted in major crop species (Heffner et al., 2011a,b; Zhao et al., 2013; Iwata et al., 2015; Spindel et al., 2015) and farm animals (Fang et al., 2017; Hay and Roberts, 2017; Tan et al., 2017). The application of GPs to alfalfa BY and forage quality breeding were also initiated recently (Annicchiarico et al., 2015b; Li et al., 2015; Biazzi et al., 2017). In alfalfa industry, BY and forage quality are the key traits for genetic improvement.

Other than the direct traits such as PH, BY, and DM that can inflect the BY of alfalfa, some phenology-related agronomic traits such as FH can also affect the BY. In this study, we therefore investigated the possibility of GP applied to alfalfa germplasm resources and GS applied to 10 important agronomic traits and 15 forage quality traits of alfalfa production.

Several methods, such as random regression BLUP, Bayesian methods and GBLUP, were employed to estimate GP and GS. Some simulation studies on different species suggested Bayesian methods to be superior than GBLUP in terms of the prediction accuracy (Meuwissen et al., 2001; Fernando et al., 2007; Clark et al., 2010; Zhang et al., 2010; Calus and Veerkamp, 2011). Compared with other methods, Bayesian methods also possessed other advantages (Gonzalez-Recio et al., 2010). In this study, we used the empirical data of 25 traits of 322 genotypes of 75 alfalfa accessions to compare the performance of GP following three statistical approaches of BayesA, BayesB, and BayesCπ. The BayesA method is based on the assumption that the prior distribution of variances of SNPs followed the scaled inverted chi-square distribution, implicating many SNPs with small effects and a small proportion of SNPs with moderate effects. BayesB assumes that many of the SNPs have no effect and the prior distribution of the variances of SNPs is a mixture of a distribution with zero variance and an inverse chi-squared distribution (Meuwissen et al., 2001). BayesCπ, however, treats the prior probability π that a SNP has zero effect as unknown (Habier et al., 2011). **Figure 1** shows that these three Bayesian methods demonstrated very similar prediction accuracies across all 25 traits, irrespective of their different assumptions. BayesA, BayesB, and BayesCπ identified six, five, and four quality traits as well as three, four, and three agronomic traits having the highest accuracies, respectively.

Besides the methods of GP discussed above, there are other factors affecting the prediction accuracies. One of them is the population composition and structure. Therefore, EBVs were directly used as the response variable to GP rather than phenotypes in the study. Since EBVs were corrected for non-genetic effects, it can be readily captured by SNPs using the Bayes methods. Methods of imputation for SNP genotypes are also important (Moghaddar et al., 2015).

Compared to previous studies, there were some differences in the accuracies of prediction for both agronomic and quality traits. For example, Biazzi et al. (2017) reported a very low accuracy (∼0.1) for LS which had nonetheless a moderate value at 0.32 in our study. DM showed a low accuracy (0.13) in our study, but Annicchiarico et al. (2015b) identified a moderate value of 0.35 in two genetically distinguished alfalfa populations. For BY, previous study showed moderate to high accuracies (0.21−0.66, Li et al., 2015) while it had an accuracy at 0.25 in the present study.

All the 15 quality traits had relatively low prediction accuracies due probably to their low heritabilities (Wang et al., 2016; Jia et al., 2017) determined by the genetic complexity of these traits. Biazzi et al. (2017) detected moderate prediction accuracy values for stem dNDF and leaf protein content (0.3–0.4) followed by leaf ADL and dNDF while the remaining traits showed low to very low accuracies. In our study, the accuracy of dNDF was almost moderate, similar to that of leaf dNDF but slightly lower than stem dNDF. These differences may be attributed to different sizes of reference populations, training populations, and number of markers. Different statistical models may lead to such discrepancies. The methods of imputation of SNP genotypes can also affect the accuracy of prediction (Moghaddar et al., 2015).

The present study was an attempt to predict alfalfa GEBVs of 25 important traits associated with BY and forage quality using three Bayesian statistical methods. Overall, they all exhibited similar predictabilities. Some traits possessed relatively high prediction accuracies (e.g., FH, FD, and PR with accuracies of 0.65, 0.52, and 0.51, respectively). Therefore, it is feasible to apply GS on these traits in alfalfa breeding programs. While GS/GP may be poorly effective for other traits such as ADL, crude protein, and RUP with low prediction accuracies.

AUTHOR CONTRIBUTIONS

ZW designed the experiments. HZ, XW, and GL phenotyped the traits. CJ and FZ analyzed the data and drafted the manuscript. ZW and JH revised the manuscript. All authors have read and approved the final manuscript.

FUNDING

This work was supported by the earmarked fund for China Agriculture Research System (CARS34), National Natural Science Foundation of China (No. 31761143013), and Agricultural Science and Technology Innovation Program (No. ASTIP-IAS-10) of China.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01220/full#supplementary-material

FIGURE S1 | Distribution of EBVs for 25 traits. Abbreviations of traits are explained in Materials and Methods.

REFERENCES

Annicchiarico, P., Barrett, B., Brummer, E. C., Julier, B., and Marshall, A. H. (2015a). Achievements and challenges in improving temperate perennial forage legumes. *Crit. Rev. Plant Sci.* 34, 327–380. doi: 10.1080/07352689.2014.898462

Annicchiarico, P., Nazzicari, N., Li, X., Wei, Y., Pecetti, L., and Brummer, E. C. (2015b). Accuracy of genomic selection for alfalfa biomass yield in different reference populations. *BMC Genomics* 16:1020. doi: 10.1186/s12864-015-2212-y

Bates, D., Mächler, M., and Dai, B. (2011). *lme4: Linear Mixed-Effects Models Using S4 Classes at*. Available at: http://lme4.r-forge.r-project.org/

Biazzi, E., Nazzicari, N., Pecetti, L., Brummer, E. C., Palmonari, A., Tava, A., et al. (2017). Genome-wide association mapping and genomic selection for alfalfa (*Medicago sativa*) forage quality traits. *PLoS One* 12:e0169234. doi: 10.1371/journal.pone.0169234

Calus, M. P. L., and Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet. Sel. Evol.* 43:26. doi: 10.1186/1297-9686-43-26

Clark, S. A., Hickey, J. M., and Van Der Werf, J. H. J. (2010). How robust are genomic selection methods? *Anim. Prod. Sci.* 50:VIII.

Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., et al. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3* 3, 1903–1926. doi: 10.1534/g3.113.008227

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. doi: 10.1371/journal.pone.0019379

Fang, L., Sahana, G., Ma, P., Su, G., Yu, Y., Zhang, S., et al. (2017). Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. *BMC Genomics* 18:604. doi: 10.1186/s12864-017-4004-z

Fernando, R. L., Habier, D., Stricker, C., Dekkers, J. C. M., and Totir, L. R. (2007). Genomic selection. *Acta Agr. Scand. A Anim. Sci.* 57, 192–195. doi: 10.1080/09064700801959395

Ganal, M. W., Polley, A., Graner, E. M., Plieske, J., Wieseke, R., Luerssen, H., et al. (2012). Large SNP arrays for genotyping in crop plants. *J. Biosci.* 37, 821–828. doi: 10.1007/s12038-012-9225-3

Gonzalez-Recio, O., Weigel, K. A., Gianola, D., Naya, H., and Rosa, G. J. M. (2010). L-2-Boosting algorithm applied to high-dimensional problems in genomic selection. *Genet. Res.* 92, 227–237. doi: 10.1017/S0016672310000261

Grenier, C., Cao, T. V., Ospina, Y., Quintero, C., Chatel, M. H., Tohme, J., et al. (2015). Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. *PLoS One* 10:e0136594. doi: 10.1371/journal.pone.0136594

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186. doi: 10.1186/1471-2105-12-186

Hay, E. H., and Roberts, A. (2017). Genomic prediction and genome-wide association analysis of female longevity in a composite beef cattle breed. *J. Anim. Sci.* 95, 1467–1471. doi: 10.2527/jas.2016.1355

Hayes, B., and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome* 53, 876–883. doi: 10.1139/G10-076

Heffner, E. L., Jannink, J. L., Iwata, H., Souza, E., and Sorrells, M. E. (2011a). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci.* 51, 2597–2606. doi: 10.2135/cropsci2011.05.0253

Heffner, E. L., Jannink, J. L., and Sorrells, M. E. (2011b). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome* 4, 65–75. doi: 10.3835/plantgenome2010.12.0029

Iwata, H., Ebana, K., Uga, Y., and Hayashi, T. (2015). Genomic prediction of biological shape: elliptic fourier analysis and kernel partial least squares (PLS) regression applied to grain shape prediction in rice (*Oryza sativa* L.). *PLoS One* 10:e0120610. doi: 10.1371/journal.pone.0120610

Jia, C., Wu, X., Chen, M., Wang, Y., Liu, X., Gong, P., et al. (2017). Identification of genetic loci associated with crude protein and mineral concentrations in alfalfa (*Medicago sativa*) using association mapping. *BMC Plant Biol.* 17:97. doi: 10.1186/s12870-017-1047-x

Jiang, Y., Schulthess, A. W., Rodemann, B., Ling, J., Plieske, J., Kollers, S., et al. (2017). Validating the prediction accuracies of marker-assisted and genomic selection of *Fusarium* head blight resistance in wheat using an independent sample. *Theor. Appl. Genet.* 130, 471–482. doi: 10.1007/s00122-016-2827-7

Lado, B., Matus, I., Rodríguez, A., Inostroza, L., Poland, J., Belzile, F., et al. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3* 3, 2105–2114. doi: 10.1534/g3.113.007807

Li, X., Wei, Y., Acharya, A., Hansen, J. L., Crawford, J. L., Viands, D. R., et al. (2015). Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. *Plant Genome* 8:90. doi: 10.3835/plantgenome2014.12.0090

Li, X., Wei, Y., Acharya, A., Jiang, Q., Kang, J., and Brummer, E. C. (2014). A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) developed using genotyping-by-sequencing is highly syntenous with the *Medicago truncatula* genome. *G3* 3, 1971–1979. doi: 10.1534/g3.114.012245

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., et al. (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* 9:e1003215. doi: 10.1371/journal.pgen.1003215

Meuwissen, T. (2007). Genomic selection : marker assisted selection on a genome wide scale. *J. Anim. Breed. Genet.* 124, 321–322. doi: 10.1111/j.1439-0388.2007.00708.x

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

Meuwissen, T. H. E., and Goddard, M. E. (1996). The use of marker haplotypes in animal breeding schemes. *Genet. Sel. Evol.* 28, 161–176. doi: 10.1186/1297-9686-28-2-161

Moghaddar, N., Gore, K. P., Daetwyler, H. D., Hayes, B. J., and Van Der Werf, J. H. J. (2015). Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. *Genet. Sel. Evol.* 47:97. doi: 10.1186/s12711-015-0175-8

Pace, J., Yu, X. Q., 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

Perez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442

Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., et al. (2012). Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.* 44, 217–220. doi: 10.1038/ng.1033

Roberts, A., McMillan, L., Wang, W., Parker, J., Rusyn, I., and Threadgill, D. (2007). Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics* 23, i401–i407. doi: 10.1093/bioinformatics/btm220

Shu, Y. J., Yu, D. S., Wang, D., Bai, X., Zhu, Y. M., and Guo, C. H. (2013). Genomic selection of seed weight based on low-density SCAR markers in soybean. *Genet. Mol. Res.* 12, 2178–2188. doi: 10.4238/2013.July.3.2

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:e1005350. doi: 10.1371/journal.pgen.1005350

Sukumaran, S., Crossa, J., Jarquin, D., Lopes, M., and Reynolds, M. P. (2017). Genomic prediction with pedigree and genotype x environment interaction in spring wheat grown in south and west Asia, North Africa, and Mexico. *G3* 3, 481–495. doi: 10.1534/g3.116.036251

Tan, C., Wu, Z., Ren, J., Huang, Z., Liu, D., He, X., et al. (2017). Genome-wide association study and accuracy of genomic prediction for teat number in duroc pigs using genotyping-by-sequencing. *Genet. Sel. Evol.* 49:35. doi: 10.1186/s12711-017-0311-8

Wang, Z., Qiang, H., Zhao, H., Xu, R., Zhang, Z., Gao, H., et al. (2016). Association mapping for fiber-related traits and digestibility in alfalfa (*Medicago sativa*). *Front. Plant Sci.* 7:331. doi: 10.3389/fpls.2016.00331

Xu, S. H., Zhu, D., and Zhang, Q. F. (2014). 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

Zhang, Z., Liu, J., Ding, X., Bijma, P., de, Koning DJ, and Zhang, Q. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One* 5:e12648. doi: 10.1371/journal.pone.0012648

Zhao, Y. S., Gowda, M., Liu, W. X., Wurschum, T., Maurer, H. P., Longin, F. H., et al. (2013). Choice of shrinkage parameter and prediction of genomic breeding values in elite maize breeding populations. *Plant Breed.* 132, 99–106. doi: 10.1111/pbr.12008

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: 15 August 2018
doi: 10.3389/fpls.2018.01165

Genomic Prediction in Tetraploid Ryegrass Using Allele Frequencies Based on Genotyping by Sequencing

Xiangyu Guo[1], Fabio Cericola[2], Dario Fè[3], Morten G. Pedersen[3], Ingo Lenk[3], Christian S. Jensen[3], Just Jensen[1] and Lucas L. Janss[1]*

[1] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark, [2] Rijk Zwaan B.V., De Lier, Netherlands, [3] Research Division, DLF Seeds A/S, Store Heddinge, Denmark

Perennial ryegrass is an outbreeding forage species and is one of the most widely used forage grasses in temperate regions. The aim of this study was to investigate the possibility of implementing genomic prediction in tetraploid perennial ryegrass, to study the effects of different sequencing depth when using genotyping-by-sequencing (GBS), and to determine optimal number of single-nucleotide polymorphism (SNP) markers and sequencing depth for GBS data when applied in tetraploids. A total of 1,515 F_2 tetraploid ryegrass families were included in the study and phenotypes and genotypes were scored on family-pools. The traits considered were dry matter yield (DM), rust resistance (RUST), and heading date (HD). The genomic information was obtained in the form of allele frequencies of pooled family samples using GBS. Different SNP filtering strategies were designed. The strategies included filtering out SNPs having low average depth (FILTLOW), having high average depth (FILTHIGH), and having both low average and high average depth (FILTBOTH). In addition, SNPs were kept randomly with different data sizes (RAN). The accuracy of genomic prediction was evaluated by using a "leave single F_2 family out" cross validation scheme, and the predictive ability and bias were assessed by correlating phenotypes corrected for fixed effects with predicted additive breeding values. Among all the filtering scenarios, the highest estimates for genomic heritability of family means were 0.45, 0.74, and 0.73 for DM, HD and RUST, respectively. The predictive ability generally increased as the number of SNPs included in the analysis increased. The highest predictive ability for DM was 0.34 (137,191 SNPs having average depth higher than 10), for HD was 0.77 (185,297 SNPs having average depth lower than 60), and for RUST was 0.55 (188,832 SNPs having average depth higher than 1). Genomic prediction can help to optimize the breeding of tetraploid ryegrass. GBS data including about 80–100 K SNPs are needed for accurate prediction of additive breeding values in tetraploid ryegrass. Using only SNPs with sequencing depth between 10 and 20 gave highest predictive ability, and showed the potential to obtain accurate prediction from medium-low coverage GBS in tetraploids.

Keywords: ryegrass, tetraploid, genomic prediction, genotyping-by-sequencing, sequencing depth

OPEN ACCESS

Edited by:
Yiwei Jiang,
Purdue University, United States

Reviewed by:
Zibei Lin,
La Trobe University, Australia
Leif Skot,
Aberystwyth University, United Kingdom

****Correspondence:***
Xiangyu Guo
xiangyu.guo@mbg.au.dk

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *13 March 2018*
Accepted: *23 July 2018*
Published: *15 August 2018*

Citation:
Guo X, Cericola F, Fè D, Pedersen MG, Lenk I, Jensen CS, Jensen J and Janss LL (2018) Genomic Prediction in Tetraploid Ryegrass Using Allele Frequencies Based on Genotyping by Sequencing. Front. Plant Sci. 9:1165. doi: 10.3389/fpls.2018.01165

INTRODUCTION

Perennial ryegrass (*Lolium perenne* L.) is one of the most widely sown forage grasses in temperate regions (Humphreys, 2005). Low production costs and the perennial character provide high agronomic value, and it is widely used for feeding ruminants (Jensen et al., 2001). The popularity of cultivating perennial ryegrass is mainly due to its re-growth capacity after defoliation and its palatability, digestibility, and nutrient content as feed for ruminants compared with other forage species (Wilkins, 1991).

Compared to diploid ryegrass, the tillers and seed heads of tetraploid ryegrass are longer and the leaves are wider. Tetraploid ryegrass is more open and more prone to wear, but is less susceptible to snow mold and has a better drought tolerance, leading to better performance under continental conditions with frequent dry periods. Palatability and digestibility are better in tetraploid varieties than in diploid varieties, and tetraploids perform better than diploids during grazing (Wilkins, 1991) and lead to a higher animal production (Lantinga and Groot, 1996; O'Donovan and Delaby, 2005).

Perennial ryegrass is an allogamous species (Cornish et al., 1979) due to a gametophytic self-incompatibility system (Cornish et al., 1979). For this reason, it is generally bred, maintained and commercialized as heterogeneous families. Evaluation of F_2 families is frequently used in breeding programs for outcrossing species such as perennial ryegrass. An F_2 family consists of the offspring from random interbreeding a full-sib F_1 family, which are the offspring from an initial bi-parental cross. F_2 families are evaluated in plot experiments over several locations and years to obtain measurements on yield, agronomic traits, and disease resistance.

Perennial ryegrass breeding has mainly relied on prediction of genetic merit using phenotypic information only (Conaghan and Casler, 2011; Hayes et al., 2013). Using this system, relevant improvements for yield and quality-related traits have been achieved (Wilkins and Humphreys, 2003; McDonagh et al., 2016). However, compared with traits such as rust resistance and spring growth, gains for yield traits like dry matter and seed yield were not as high as expected (Sampoux et al., 2011). In addition, phenotypic selection is costly and time consuming, needing up to 10 years to complete a selection cycle (Wilkins and Humphreys, 2003; Lin et al., 2016). In recent decades, the development of marker technology allowed adoption of genomic prediction (GP) strategies, which have been highly beneficial and led to a reduction of cost in practical animal and plant breeding programs (Hickey et al., 2017). In GP, dense markers distributed across the whole genome can be used simultaneously to predict breeding values (Meuwissen et al., 2001). The quantitative trait loci (QTLs) affecting the traits of interest are assumed in linkage disequilibrium (LD) with one or more single-nucleotide polymorphism (SNP) markers. Thus, a sufficiently dense and well-distributed set of markers allows all QTLs to be in LD with at least one marker, and this LD can be exploited in GP to ensure accurate prediction of breeding values as a basis for selection decisions.

The prospects for implementing GP in forage grass breeding were recently reviewed by Hayes et al. (2013). Several GP studies have been reported for crops such as maize and wheat (Crossa et al., 2010, 2014), and the first investigations in diploid perennial ryegrass also demonstrated great potential for using GP (Fè et al., 2015a, 2016). However, GP studies for tetraploid ryegrass, to our knowledge, have not yet been carried out. The implementation of GP in tetraploid ryegrass may be more challenging than in diploid ryegrass, because families of tetraploids will show a higher heterozygosity than families of diploids. This may hamper accurate estimation of genomic relationships and genomic breeding values.

Genotyping-by-sequencing (GBS) was developed by Elshire et al. (2011) as a robust genotyping approach. GBS uses methylation sensitive restriction enzymes to reduce genome complexity. GBS is a good approach to estimate genome-wide allele frequency profiles in pooled samples for outbred heterogeneous varieties (Byrne et al., 2013). Moreover, for association studies and GP studies, calling of genotypes can be avoided by directly using allele frequencies from GBS, which facilitates measurements on family pools (Ashraf et al., 2014). Use of GBS data also poses some challenges; in particular, sequencing depth needs to be optimized carefully. At low depth, genotyping errors and missing values are an issue (Poland and Rife, 2012), and result in biased estimates of allele-effect and heritability (Ashraf et al., 2014, 2016). At higher sequencing depth the accuracy of genotype estimates is improved (Sims et al., 2014), but under a fixed budget, the number of samples that can be sequenced would be reduced, which reduces power of the entire experiment (Ashraf et al., 2014). Several investigations on how sequencing depth affects association studies and estimation of genomic heritability have been conducted (Garner, 2011; Sims et al., 2014; Ashraf et al., 2016). As reviewed by Poland and Rife (2012), GBS has become a flexible and low cost tool for plant genetics and breeding. It has been demonstrated that GBS can effectively generate high-density genome wide markers in a range of species (Elshire et al., 2011; Poland and Rife, 2012; Poland et al., 2012; Beissinger et al., 2013; Crossa et al., 2013; Zhang et al., 2015; Fè et al., 2016; Cericola et al., 2018). With GBS, an accurate GP model was derived for the large, complex, and polyploid wheat genome (Poland et al., 2012). In addition, GBS also has been applied on diploid ryegrass for genomic prediction (Fè et al., 2015a, 2016). However, to our knowledge the optimization of sequencing depth for GP in tetraploid ryegrass has not been reported yet.

The aims of this study were: (1) to investigate the possibility of implementing genomic prediction in tetraploid perennial ryegrass, (2) to study the effects of different sequencing depth when using GBS, and (3) to determine the optimal number of

Abbreviations: CV, cross-validation; DM, dry matter yield; FILTBOTH, the strategy filtering out SNPs having both low average and high average depth; FILTHIGH, the strategy filtering out SNPs having high average depth; FILTLOW, the strategy filtering out SNPs having low average depth; GBS, genotyping-by-sequencing; GEBVs, genomic breeding values; GP, genomic prediction; GSLM, family × sowing year × location × management effects; GSLMF, family × sowing year × location × management × farming year effects; G×E, genotype by environment; HD, heading date; LD, linkage disequilibrium; QTLs, quantitative trait loci; RAN, the strategy keeping SNPs randomly with different data sizes; RUST, rust resistance; SNP, single-nucleotide polymorphism.

SNPs to include in genomic prediction when GBS are applied in tetraploid ryegrass.

MATERIALS AND METHODS

Plant Material

Both phenotype and genotype data were derived from 1,515 F_2 families from a commercial breeding program from DLF A/S, Denmark. F_2 families originated from a pair-cross between two parents; F_1 seeds from both parent plants were pooled; F_1 families were sown in small protected plots to cross-fertilize; and finally F_2 seeds were harvested and used for field-testing of F_2 families. A detailed description of testing procedures was provided by Fè et al. (2015b).

Phenotypic records, defined below, consisted of historical data from F_2 families, which were sown between 2004 and 2016 at 8 locations in Denmark, the Netherlands, France, and United Kingdom, and cultivated according to the local management schemes. In all locations, F_2 families were tested in trials including 12 families in a randomized experiment with two replicates for each family. Details of testing and recording procedures were the same as for diploid ryegrass as described previously (Fè et al., 2015b). The dataset analyzed included records of three traits:

(1) Dry matter yield (DM), expressed in kg/m^2 and obtained from multiple cuts over 2 years. For analyses, the total yield during the first year and the total yield during the second year were used so that each family had yield measurements from two years; to validate genomic predictions, the average yield of the two years was predicted.
(2) Heading date (HD), defined as the day on which spikes are visible over the general plots, and expressed in days since January 1st. HD was scored in plots for seed multiplication, which were farmed for one cropping season only.
(3) Rust resistance (RUST), measured during the period of maximum infection, both in regular yield plots, and in mini plots, which were cultivated only for 1 year. The level of infection was determined by visual scoring from 1 (plants completely covered by rust) to 9 (no sign of rust infection). Plots were cut between the different scoring time points to make the scores independent.

Descriptive statistics including mean value, standard deviation, minimum, maximum, number of families, number of records, number of plots, and number of sowing year × location × management levels are listed in **Table 1**.

Filtering of GBS Data and Calculation of Allele Frequencies for Each Family

Genotypic data was produced as described previously (Fè et al., 2015a). In total, 1,515 F_2 families were sequenced. A total of 51 libraries were prepared, with up to 96 families per library. Each library was sequenced on multiple lanes of an Illumina HiSeq2000 (single-end). On average, 12.9 million 100 bp single-end reads were produced per sample. A draft sequence assembly (Byrne et al., 2015) was used for the alignment of data for each family, and initially 18.6 million SNPs were identified. A first, quite liberal, filtering of the raw SNP data was performed by removing: (1) SNPs with missing rate higher than 50%; (2) SNPs with allele frequencies lower than 0.01 or higher than 0.99; (3) SNPs with average read depth smaller than 1. This left 188,832 SNPs available for our analysis, which included further, more stringent, filtering steps for the SNPs. The average read depth for the 188,832 SNPs ranged from 1 to 278, with mean of 19. The distribution of average read depth for each SNP is shown in **Figure 1**.

Differently from SNP chip data, where genotypes are explicitly called, the genotype of a SNP is obtained here in the form of an allele frequency ($\widehat{g_{ij}}$), which is estimated as the ratio between

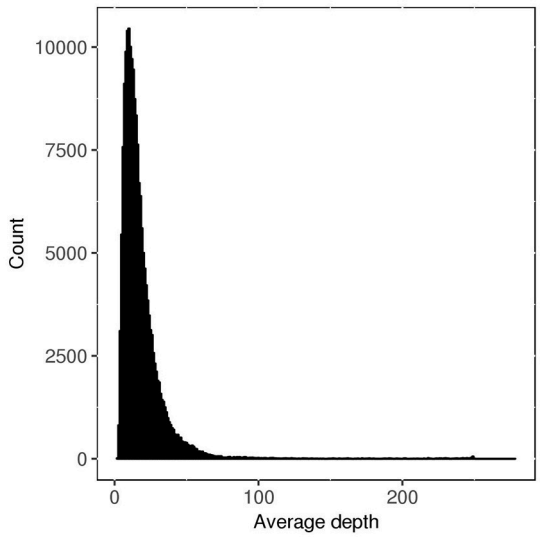

FIGURE 1 | Distribution of average read depth for each SNP before filtering by different strategies.

TABLE 1 | Descriptive statistics[a] for three traits.

Trait[b]	No. Fam	No. Rec	No. Plot	No. YLM	Mean	*SD*	Min	Max
DM	1,188	5,312	3,414	27	1.33	0.37	0.41	2.5
HD	979	1,810	1,810	7	155.64	7.51	136	178
RUST	1,506	13,545	5,368	22	5.64	1.99	1	9

[a] *No. Fam, number of families; No. Rec, number of records; No. Plot, number of plots; No. YLM, number of sowing year × location × management levels; Mean, mean value; SD, standard deviation; Min, minimum value; Max, maximum value.*
[b] *DM, dry matter yield; HD, heading date; RUST, rust resistant.*

number of reads for alternative allele ($S_{1_{ij}}$) and the total number of reads ($S_{T_{ij}}$), which is the sum of number of reads for the reference allele ($S_{2_{ij}}$) and $S_{1_{ij}}$, for each sample:

$$\widehat{g_{ij}} = \frac{S_{1_{ij}}}{S_{T_{ij}}} = \frac{S_{1_{ij}}}{S_{1_{ij}} + S_{2_{ij}}}.$$

SNP Filtering Strategies

In order to study the effect of sequencing depth of GBS data, additional SNP filtering was performed. First, SNPs having average depth lower than a certain value were filtered out in 11 levels, with minimum depth from 1 to 90 (FILTLOW1 to FILTLOW11); second, SNPs having average depth higher than a certain value were filtered out in 11 levels, with maximum depth from 100 to 5 (FILTHIGH1 to FILTHIGH11); third, SNPs having average depth outside a certain range were filtered out (equivalent to keeping SNPs with average depth within that range), using 12 different ranges (FILTBOTH1 to FILTBOTH12); finally, SNPs were kept randomly with 11 different data sizes from 5 to 180 k (RAN5 to RAN180), and repeated for 10 times. In summary, there were four filtering strategies, FILTLOW, FILTHIGH, FILTBOTH, and RAN, and the number of scenarios was, respectively, 11, 11, 12, and 11, where the latter (RAN) was repeated 10 times. A summary of SNPs used in each filtering scenario is shown in **Figure 2** and the details are shown in **Supplementary Table 1**.

Statistical Model and Methods

A single trait model was used to estimate variance components and fixed effects, and to predict breeding values as well as other random effects in the model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z_g g} + \mathbf{Z_a a} + \mathbf{Z_p p} + \mathbf{Z_{i_1} i_1} + \mathbf{Z_{i_2} i_2} + \mathbf{e},$$

in which $\mathbf{y}$ was the vector of phenotypic values of the trait DM, HD or RUST; $\mathbf{b}$ was the vector of fixed effects (sowing year × location × management × trial × farming year); $\mathbf{g}$ was the vector of additive genomic family effects; $\mathbf{a}$ was the vector of residual genetic family effects which were not explained by the genomic information; $\mathbf{p}$ was the vector of random plot effects; $\mathbf{i_1}$ and $\mathbf{i_2}$ were vectors of genotype by environment (G×E) effects [$\mathbf{i_1}$: family × sowing year × location × management (GSLM), $\mathbf{i_2}$: family × sowing year × location × management × farming year (GSLMF)]; and $\mathbf{e}$ was the vector of random residual effects. $\mathbf{X}$, $\mathbf{Z_g}$, $\mathbf{Z_a}$, $\mathbf{Z_p}$, $\mathbf{Z_{i_1}}$, and $\mathbf{Z_{i_2}}$ were incidence matrices associating $\mathbf{b}$, $\mathbf{g}$, $\mathbf{a}$, $\mathbf{p}$, $\mathbf{i_1}$, and $\mathbf{i_2}$ with $\mathbf{y}$. The random effects were assumed to be independent of each other and normally distributed, that is, $\mathbf{g} \sim N(0, \mathbf{G}^*\sigma_g^2)$, $\mathbf{a} \sim N(0, \mathbf{I}\sigma_a^2)$, $\mathbf{p} \sim N(0, \mathbf{I}\sigma_p^2)$, $\mathbf{i}_1 \sim N(0, \mathbf{I}\sigma_{i_1}^2)$, $\mathbf{i}_2 \sim N(0, \mathbf{I}\sigma_{i_2}^2)$, $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, in which $\mathbf{G^*}$ was the corrected $\mathbf{G}$ matrix of additive genomic relationships constructed based on the genomic information, $\mathbf{I}$ was the identity matrix, and σ_g^2, σ_a^2, σ_p^2, $\sigma_{i_1}^2$, $\sigma_{i_2}^2$, and σ_e^2 were the variances of additive genomic effects, residual genetic effects, random plot effects, first genetic by environment effects, second genetic by environment effects, and residuals, respectively. For DM and RUST, the general model was applied in the analysis, while for HD, the effects of $\mathbf{p}$ and $\mathbf{i_2}$ were excluded since there was only one score and only one environment in each family for HD.

The method to compute the $\mathbf{G}$ matrix was based on a modification of VanRaden (2008) to use allele frequencies

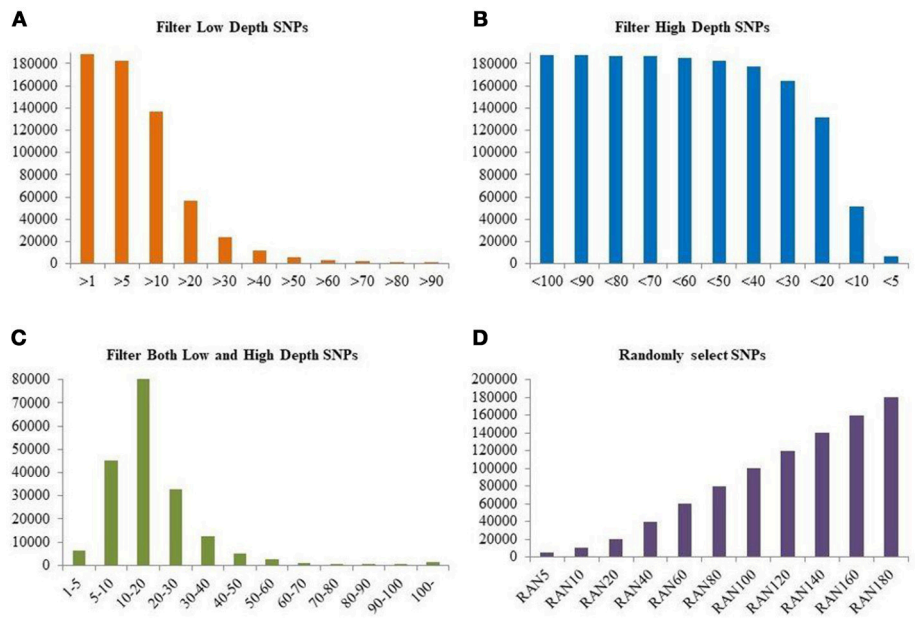

FIGURE 2 | Numbers of SNPs retained for four different filtering strategies. **(A)** Number of SNPs in each FILTLOW dataset with different lower threshold; **(B)** Number of SNPs in each FILTHIGH dataset with different upper threshold; **(C)** Number of SNPs in each FILTBOTH dataset with different lower and upper thresholds; **(D)** Number of SNPs in each RAN dataset with different size. The y-axis is the number of SNPs after filtering. In **(A–C)**, the x-axis is different subsets of SNPs with various sequencing depth, and the x-axis in **(D)** is randomly selected subset with different size.

(ranging between 0 and 1) instead of SNP genotype calls. A matrix (**F**) with allele frequencies for each sample was centered by the mean SNP frequencies to create matrix **M** ($\mathbf{M_j} = \mathbf{F_j} - \overline{\mathbf{F_j}}$). Then, the **G** matrix was obtained by computing **M** multiplied by its own transpose and scaled by the sum of expected SNP variances across genotypes ($\mathbf{G} = \mathbf{MM}'/\mathbf{K}$). The scale parameter used for tetraploid F_2 families is half that used for diploid F_2 families as computed by Ashraf et al. (2014) and as applied in the study by Fè et al. (2015a), because the number of alleles in F_2 family pools is eight for tetraploid families, which is double that of diploid families:

$$\mathbf{K} = 0.125 \sum \overline{F_j}(1 - \overline{F_j}).$$

Finally, the **G** matrix was corrected for the extra binomial variance due to limited sequencing depth. The correction was derived by Cericola et al. (2018) and simply can be done according to ploidy number and the average depth of the sample. Corrected **G** matrix (**G***) was calculated by scaling down the diagonal elements of each individual as follows:

$$D_{c_i} = D_{b_i}(1 - \frac{n-1}{\overline{S_{T_i}} + n - 1}),$$

where D_{b_i} is the *ith* element of the biased diagonal element in **G** and D_{c_i} is the corrected diagonal element in **G***, $\overline{S_{T_i}}$ is the average $S_{T_{ij}}$ for each individual across all SNPs, and n is the ploidy number, which is eight as mentioned before.

For each of the four filtering scenarios, single trait analyses were run on the subsets of SNPs, which were previously created according to different filtering strategies (**Figure 2** and **Supplementary Table 1**). Variance components and their standard errors (SE) were estimated by restricted maximum likelihood (REML) using the DMU software package (Madsen and Jensen, 2013).

The phenotypic variance of family means was calculated as the sum of weighted variance components:

$$\sigma^2_{P_f} = \overline{G^*}\sigma^2_g + \sigma^2_a + \sigma^2_p/n_p + \sigma^2_{i_1}/n_{i_1} + \sigma^2_{i_2}/n_{i_2} + \sigma^2_e/n_e,$$

where $\overline{G^*}$ is the average diagonal of the corrected genomic relationship matrix (**G*** matrix), n_p is the average number of plots for each family, n_{i_1} and n_{i_2} are the average numbers of environments for each family, and n_e is average number of replicates across all fields for each family. Accordingly, genomic family heritability based on multiple plots was calculated as $h^2_f = \frac{\overline{G^*}\sigma^2_g}{\sigma^2_{P_f}}$. To evaluate importance of each random effect in the model, phenotypic variance of a single plot was also calculated:

$$\sigma^2_{P_p} = \overline{G^*}\sigma^2_g + \sigma^2_a + \sigma^2_p + \sigma^2_{i_1} + \sigma^2_{i_2} + \sigma^2_e.$$

In the calculation of $\sigma^2_{P_f}$, $\sigma^2_{P_p}$ and h^2_f, σ^2_p, and $\sigma^2_{i_2}$ were not considered for HD due to the reduced recording strategy for this trait. This was used to compute the relative contribution of each random effect to the total phenotypic plot variance.

Cross-Validation

To estimate the accuracy of genomic breeding values (GEBVs), a leave-one-family-out cross-validation (CV) strategy was applied. In each CV round, the phenotypes of one family were masked and then all other families were used to train the prediction model and to predict the family with phenotypes masked.

Before CV, the whole dataset was used to estimate variance components and fixed effects. Corrected phenotypes (y_c) were computed by subtracting the estimates of the fixed effects. Predictive ability was measured as $cor(\overline{y_c}, \hat{g})$, which is theoretically not larger than the square root of h^2_f (Legarra et al., 2008) because breeding values predict genetic effects and not environment. $\overline{y_c}$ is the average y_c for each family. Furthermore, to assess bias of predictions, regression coefficient of $\overline{y_c}$ on $\hat{g}$ was calculated. The deviation of this regression coefficient from 1 represents the level of bias.

RESULTS

In order to interpret the results from each scenario, **Figures 2–8** were created. **Figure 2** and **Supplementary Table 1** show the numbers of SNPs retained for four different filtering strategies. **Figure 2** shows four bar charts according to the data filtering levels. **Figures 3–5** show the estimated heritability, predictive ability and bias in different SNP filtering scenarios for three traits, DM, HD and RUST, respectively. In these figures, line charts were plotted as a function of number of SNPs included in each model. **Figures 6–8** show the percentages of explained variance, i.e., each variance components over the total phenotypic variance, for three traits. In these three figures, bar charts were plotted for all scenarios.

SNP Filtering Strategies

The first filtering strategy FILTLOW used an increasing lower threshold for average SNP read depth, and the number of SNPs included decreased from 188,832 for FILTLOW1 to 1,587 for FILTLOW11. The second filtering strategy FILTHIGH used a decreasing upper threshold for average SNP read depth, and the number of SNPs included decreased from 187,516 for FILTHIGH1 to 6,389 for FILTHIGH11. In this data, a large proportion of SNPs had read depth between 10 and 20, which caused large reductions in the numbers of SNP when either the lower threshold for read depth increased to 20, or when the upper threshold for read depth decreased to 10. For instance, between FILTLOW >10 and FILTLOW >20, the number of SNPs kept dropped from 73 to 30%, and between FILTHIGH <20 and FILTHIGH <10, the number of SNPs kept dropped from 70 to 27%. The SNPs with depth lower than 10 and 5 accounted for 27 and 3% of the full data, respectively. In the third filtering strategy, SNPs were kept in a certain interval of average read depth. The percentage of SNPs kept varied from 43% for FILTBOTH3 having average read depth from 10 to 20, to 0.1% for FILTBOTH11 having average read depth from 90 to 100. In addition to three filtering strategies for average read depth, the RAN filtering strategy kept random subsets from the full dataset, ranging from 5 K (RAN5) to 180 K (RAN180) SNPs;

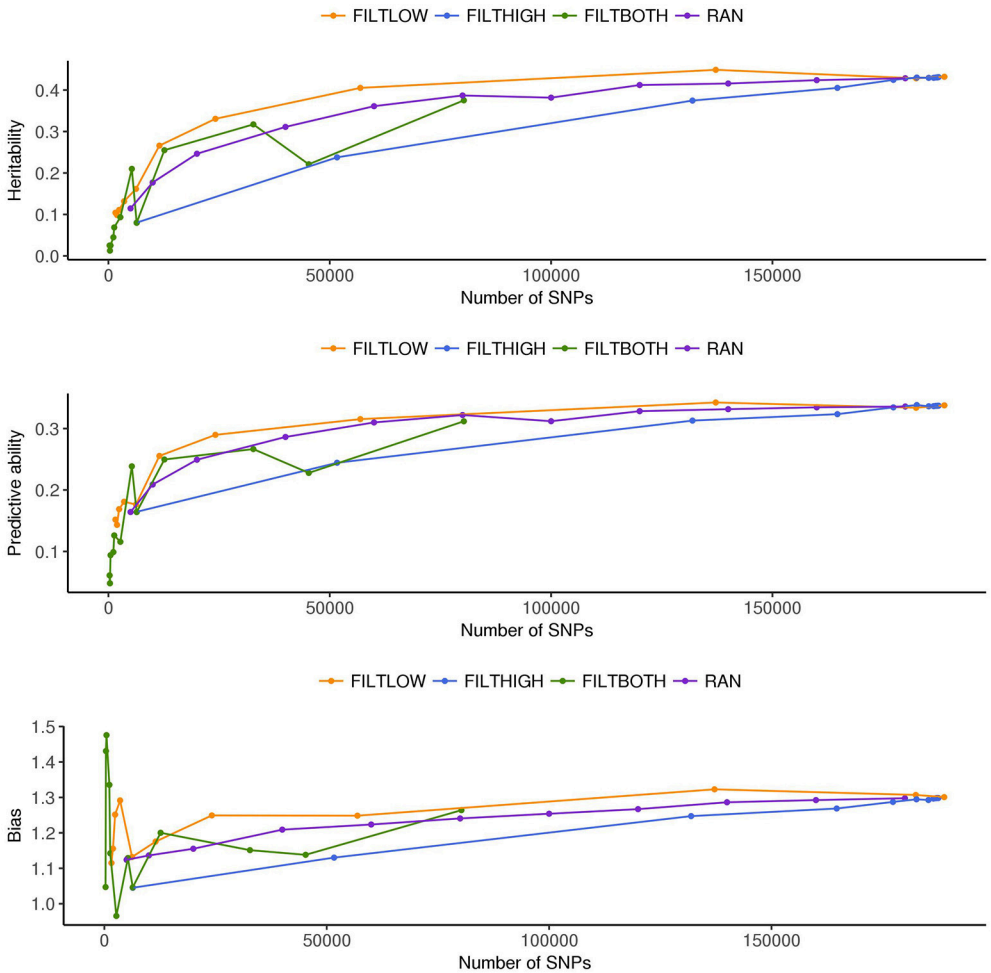

FIGURE 3 | Estimated heritability, predictive ability and bias in different SNP filtering scenarios[1] for dry matter yield. [1] FILTLOW, strategy filtering out SNPs having low average depth; FILTHIGH, strategy filtering out SNPs having high average depth; FILTBOTH, strategy filtering out SNPs having both low average and high average depth; RAN, strategy keeping SNPs randomly with different data size.

in percentages this corresponds to 3% of SNPs in RAN5 to 95% of SNPs in RAN180.

Variance Components and Heritabilities

Figures 3–5 show the effects of different filtering strategies on estimates of h_f^2 for the traits DM, DH, and RUST, respectively. For DM, the highest estimate of h_f^2 was 0.45 (FILTLOW3). In the FILTLOW scenarios, the estimated h_f^2 increased slightly from FILTLOW1 to FILTLOW3, and decreased fast afterwards. In the FILTHIGH scenarios, the estimated h_f^2 generally decreased from FILTHIGH1 to FILTHIGH11, and the decrease was more obvious from FILTHIGH8. In the FILTBOTH scenarios, the estimated h_f^2 was highest in FILTBOTH3, showed a small decrease in FILTBOTH4, and larger reductions in other FILTBOTH scenarios. In the RAN scenarios, the estimated h_f^2 increased along with the number of SNPs included in the model, i.e., increased from RAN5 to RAN180, with rate of increase gradually reducing. For HD, the trends of heritability estimates within the four filtering strategies were the same as for DM, and the highest estimate of h_f^2 was also for FILTLOW3 at 0.74. For RUST, the highest estimate of h_f^2 was 0.73 for FILTHIGH1. In the FILTLOW scenarios, the estimated h_f^2 were similar for FILTLOW1 to FILTLOW3, and also decreased fast afterwards. The trends for the other three filtering strategies (FILTHIGH, FILTBOTH and RAN) were the same as for DM and HD.

In our analysis model, we also include a variance component for residual genetic effects (σ_a^2), i.e., the part of genetic effects that cannot be explained by genomic markers. **Figures 6–8** show that the percentage of σ_a^2 over $\sigma_{P_p}^2$ changed in the different scenarios. When the number of SNPs increased, the percentage of additive genetic variance explained by markers generally increased while the percentage of residual genetic variance decreased, and the percentage of total genetic variance (sum of $\overline{\mathbf{G}^*}\sigma_g^2$ and σ_a^2) over $\sigma_{P_p}^2$ remained relatively similar for all scenarios.

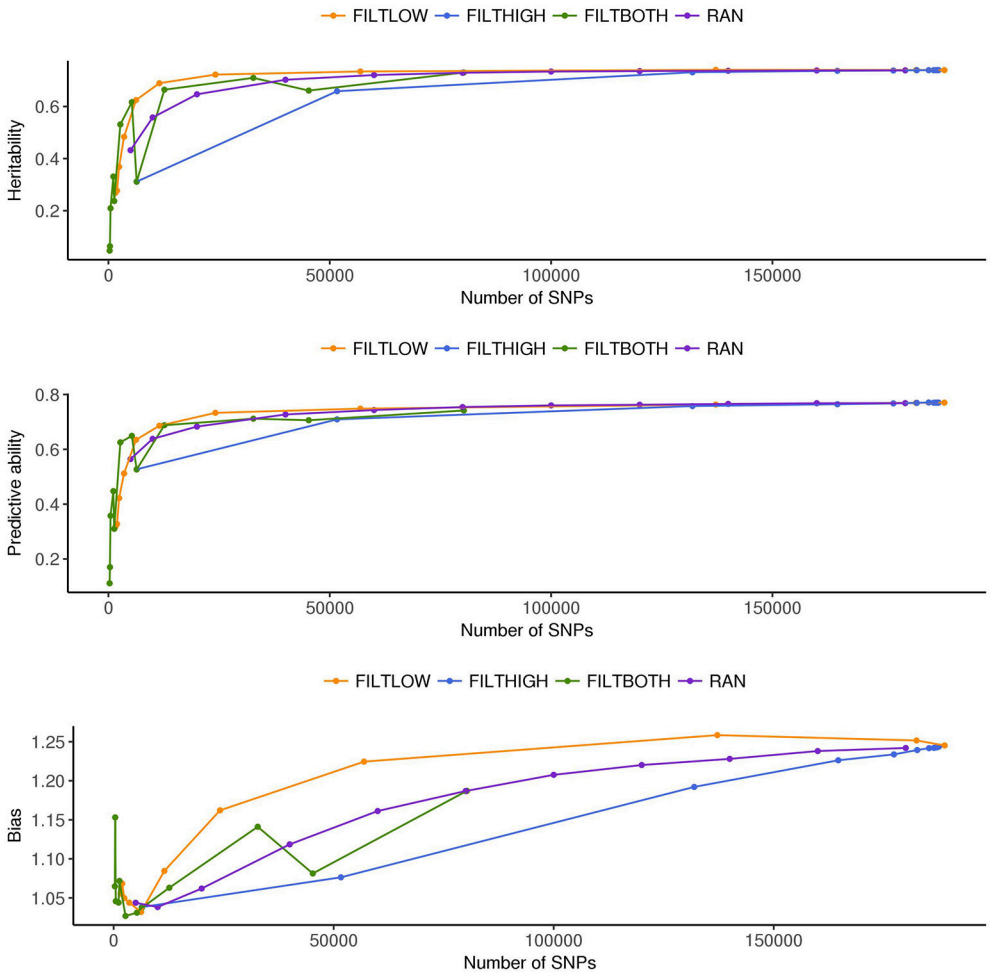

FIGURE 4 | Estimated heritability, predictive ability and bias in different SNP filtering scenarios[1] for heading date. [1] FILTLOW, strategy filtering out SNPs having low average depth; FILTHIGH, strategy filtering out SNPs having high average depth; FILTBOTH, strategy filtering out SNPs having both low average and high average depth; RAN, strategy keeping SNPs randomly with different data size.

For the variance of plot effects estimated for DM and RUST, the percentages of σ_p^2 over $\sigma_{P_p}^2$ were consistent across all the scenarios but different between DM and RUST. The percentage of variance due to plot effects in DM was about twice as large as the plot variance in RUST (**Figure 6**, **8**). For DM, σ_p^2 and total genetic variance had similar magnitude, but for RUST, σ_p^2only accounted for 29% of total genetic variance.

As shown in **Figures 6**, **8**, the estimates of variance for G×E interactions GSLM ($\sigma_{i_1}^2$) and GSLMF ($\sigma_{i_2}^2$) were similar among all the scenarios, but different in DM and RUST. For DM, estimates of $\sigma_{i_1}^2$ were not significantly different from 0. However, for RUST, estimates for both $\sigma_{i_1}^2$ and $\sigma_{i_2}^2$ were significantly different from 0, with the average percentages of $\sigma_{i_2}^2$ being slightly larger than$\sigma_{i_1}^2$. For HD, the estimates of $\sigma_{i_1}^2$ varied more between scenarios than for DM and RUST, and the percentage of $\sigma_{i_1}^2$ over $\sigma_{P_p}^2$ ranged from 11 to 28%. When the number of SNPs increased, this percentage generally decreased. For HD, less phenotypic records were available and therefore variance components were estimated with lower accuracy.

The estimation of residual variance (σ_e^2) was generally consistent among all scenarios. The average percentage of σ_e^2 was 59, 22, and 25%, for DM, HD and RUST, respectively. The largest difference between residual variance estimates for the different scenarios was 5% for HD (24% in FILTHIGH5 vs. 19% in FILTLOW11), 2% for DM (60% in FILTHIGH11 vs. 58% in FILTLOW3), and 1% in RUST (26% in FILTHIGH11 vs. 25% in FILTLOW11).

Details on estimated variance components and heritabilities, together with their standard errors (SE), for three traits in all F_2 families are available in **Supplementary Table 2** (DM), **Supplementary Table 3** (HD) and **Supplementary Table 4** (RUST).

Cross-Validation

Detailed results from CV for three traits are available in **Supplementary Table 5** (HD).

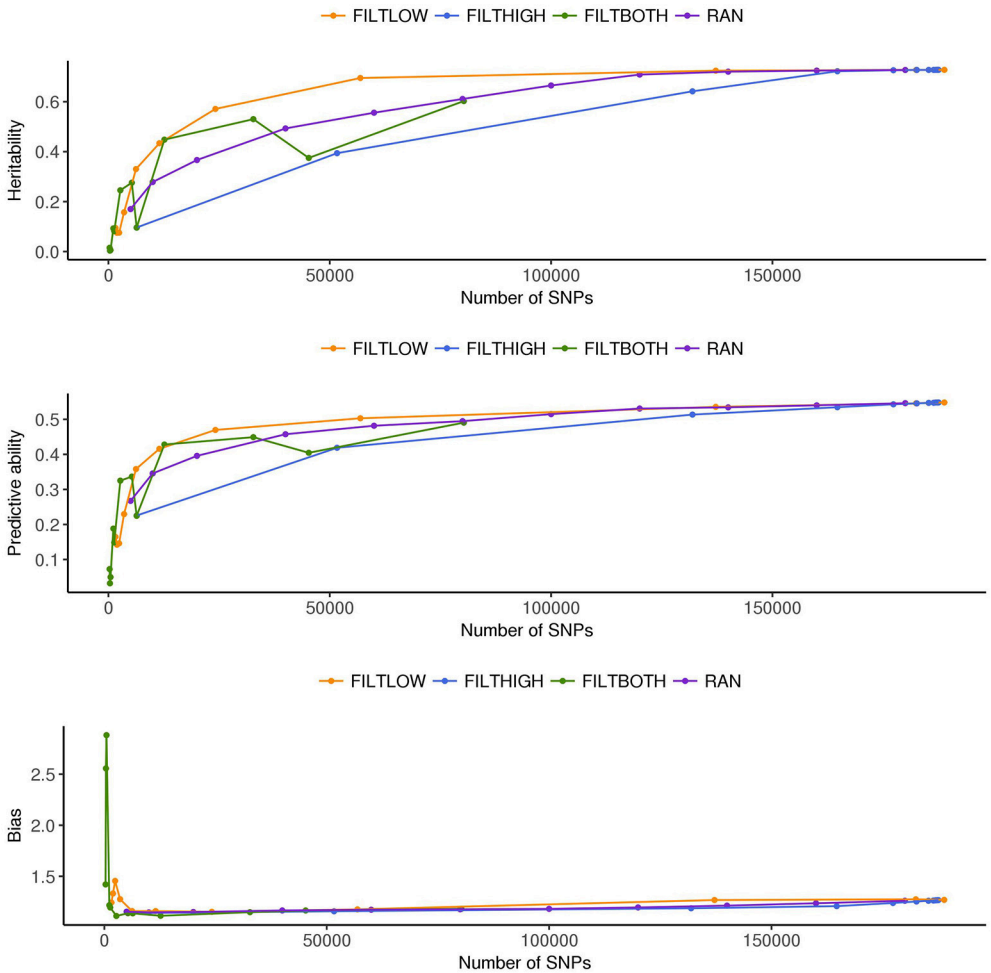

FIGURE 5 | Estimated heritability, predictive ability and bias in different SNP filtering scenarios[1] for rust resistance. [1] FILTLOW, strategy filtering out SNPs having low average depth; FILTHIGH, strategy filtering out SNPs having high average depth; FILTBOTH, strategy filtering out SNPs having both low average and high average depth; RAN, strategy keeping SNPs randomly with different data size.

Figure 3 (DM), **Figure 4** (HD), and **Figure 5** (RUST), show that the predictive ability generally increased when the number of SNPs included in the analysis increased. The highest predictive ability for DM was provided by dataset FILTLOW3 (0.34) with 137,191 SNPs having average depth higher than 10, the highest predictive ability for HD was provided by dataset FILTHIGH5 (0.77) with 185,297 SNPs having average depth lower than 60, and the highest predictive ability for RUST was provided by dataset FILTLOW1 (0.55) with 188,832 SNPs having average depth higher than 1, which was equivalent to including all markers.

Randomly filtering out SNPs and varying the number of SNPs showed that predictive ability generally increased with increasing number of SNPs included in the analysis. Above 80–100 K SNPs, effects of further increases were limited.

Overall, with an increase in the number of SNPs included in the analysis, the bias, which is the deviation from 1 for the regression of predictions on observed phenotypes, also increased. For all three traits, the FILTLOW strategy showed more biased predictions than RAN, whereas the FILTHIGH strategy showed less biased prediction than RAN. In addition, larger bias was always observed together with better predictive ability. For DM, strategy FILTLOW3 provided best predictive ability, but the bias when using this subset of SNPs was also high (regression coefficient was 1.32). For HD, the most biased prediction was also provided by dataset FILTLOW3 (regression coefficient was 1.26), though the best predictive ability was provided by the dataset FILTHIGH5, the predictive ability when using dataset FILTLOW3 was only 0.01 lower than FILTHIGH5. For RUST, FILTLOW1 provided highest predictive ability, and the bias provided by this dataset was slightly larger (regression coefficient was 1.27) than other scenarios, except for models with few SNPs included.

DISCUSSION

To investigate the potential for genomic prediction in tetraploid ryegrass we analyzed data from 1,515 F_2 families. All families were genotyped using GBS with an average sequencing depth

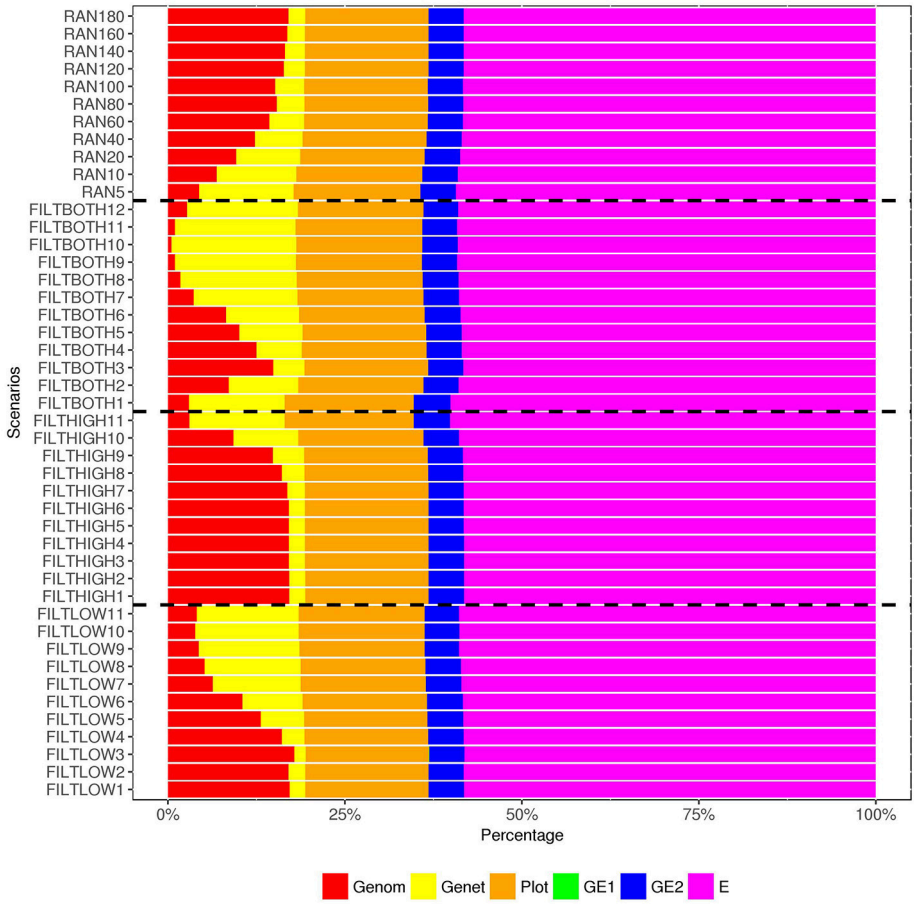

FIGURE 6 | Percentage of variance components[1, 2] over the total phenotypic variance for dry matter yield. [1] Genom, Additive genomic variance; Gene, Residual genetic variance; PLOT, random plot variance; GE1, family × sowing year × location × management variance; GE2, family × sowing year × location × management × farming year variance; E, residual environment variance. [2] GE1 is too small to be visible, so that there are only five variances can be observed in this figure.

of 19. GBS data, with various strategies for filtering SNPs, were used in GP models, and we compared heritabilities and predictive abilities to determine optimal SNP numbers and sequencing depth for genomic prediction in tetraploid ryegrass. Among all the filtering scenarios, the highest estimates for genomic heritability of family means were 0.45, 0.74 and 0.73 for DM, HD and RUST, respectively. The predictive ability generally increased as the number of SNPs included in the analysis increased. The highest predictive ability for DM was 0.34 (137,191 SNPs having average depth higher than 10), for HD was 0.77 (185,297 SNPs having average depth lower than 60), and for RUST was 0.55 (188,832 SNPs having average depth higher than 1).

Heritabilities and Variance Components

Several studies have reported heritabilities in diploid ryegrass for the same traits studied here. Fè et al. (2015b) reported analysis of total DM in two years, and heritability ranged from 0.20 to 0.25, and the estimates of heritability of total DM over two years were slightly higher than in first and second year separately. In the current study, DM was defined as the total dry matter yield in each of two farming years, and modeled as a trait with repeated records while the overall year effect was included in the fixed effects. The estimate of h_f^2 was higher than heritabilities reported by Fè et al. (2015b). Compared with the heritability for HD in diploids, where estimates ranged from 0.49 to 0.68 (Fè et al., 2015a,b) and from 0.07 to 0.22 (Ashraf et al., 2016), the estimates in the current study are higher. RUST was investigated in diploid varieties by Ravel and Charmet (1996), Waldron et al. (1998), Fè et al. (2015b), and Fè et al. (2016), and the estimates of h_f^2 in the current study were in the range reported for diploid ryegrass. In the previous study on diploid ryegrass by Fè et al. (2015b), estimated heritability for DM was similar to that for RUST, but in the current study we find a larger difference between estimated heritabilities for DM and RUST.

G×E effects accounted for about 10% of total variance for HD in diploid ryegrass (Fè et al., 2015a), which is similar to results in the current study. Although the proportion of total phenotypic variance explained by genetic marker information was much less in DM than in HD and RUST, the variances of G×E effects were also small in DM.

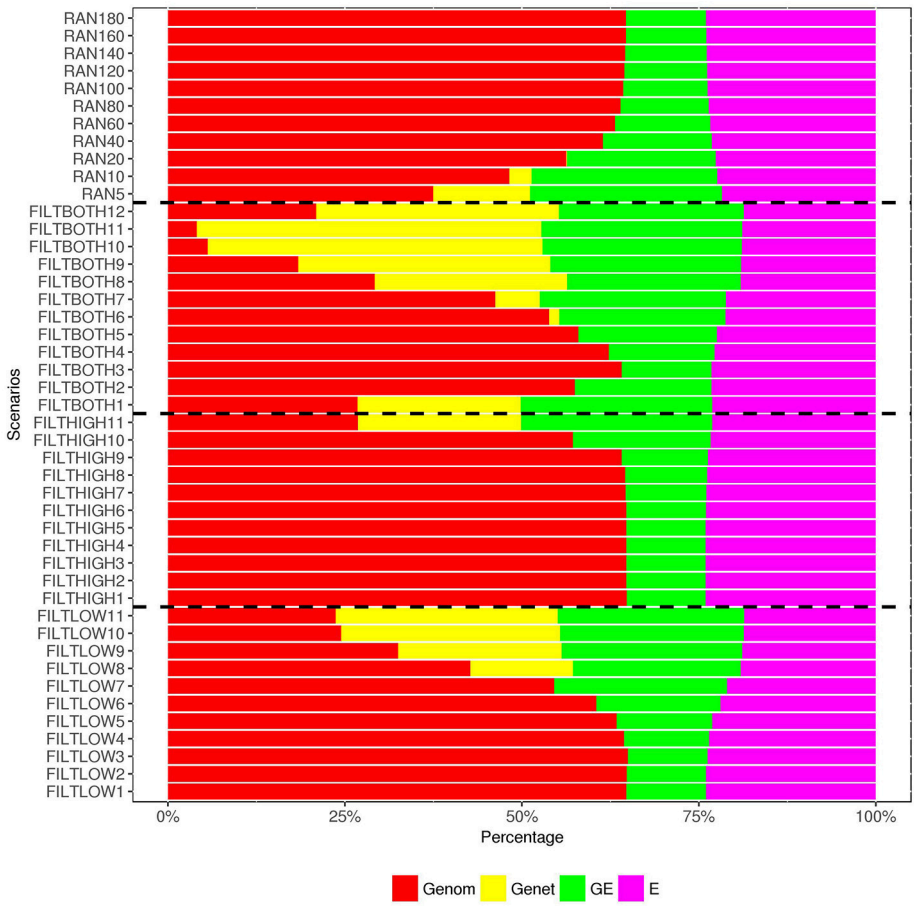

FIGURE 7 | Percentage of variance components[1] over the total phenotypic variance for heading date. [1] Genom, Additive genomic variance; Gene, Residual genetic variance; GE, family × sowing year × location × management variance; E, residual environment variance.

The proportions of variances due to the two G×E effects were different for DM and RUST. For DM, the second G×E effect (GSLMF) was important, but the first G×E effect (GSLM) explained only a small part, which indicated that growth season had a large effect on DM and can modify the ranking of different families. In diploid ryegrass, the genetic and phenotypic correlation between DM in the two years were 0.62 and 0.39 respectively (Fè et al., 2015b), which also implies large G×E effects. Variance of G×E effects in RUST was different from G×E in DM. For RUST, both the first and the second G×E effects accounted for similar amount of variances (around 12% of total phenotypic variance), which is comparable to results from the previous study on diploids (Fè et al., 2016).

The proportion of residual variance at the level of single plots was different among the three traits. The residual variance accounted for more than 50% of phenotypic variance in DM but only around 25% in both HD and RUST. The large amount of residual variance in DM indicates larger measurement errors in DM, and necessarily leaves only relatively small proportions of variance that can be attributed to the other effects.

SNP Filtering Strategies

Sequencing depth is an important factor when utilizing GBS data. An increase of sequencing depth means that the average number of times a locus been sequenced is increased, so that the accuracy of measuring the frequency of the reference allele is also increased. However, increasing sequencing depth also increases the cost of sequencing. Therefore, it is crucial to investigate the optimal sequencing depth when using GBS data. In the current study, different SNP filtering scenarios were compared with regard to parameter estimation and genomic prediction results. Four SNP filtering strategies were applied on the full GBS dataset, creating subsets of SNPs with different sequencing depth and/or different numbers of SNPs.

A previous study on diploid ryegrass (Ashraf et al., 2016) used GBS data with sequencing depth varying from 0 to 60, and divided the SNPs in 5 groups with depth interval of 10. Ashraf et al. (2016) did not correct for low accuracy of allele frequency estimates at low sequencing depth, and showed this creates a general trend of increasing genomic heritability with increasing sequencing depth. In the current study, we corrected for the effects of low accuracy at low sequencing

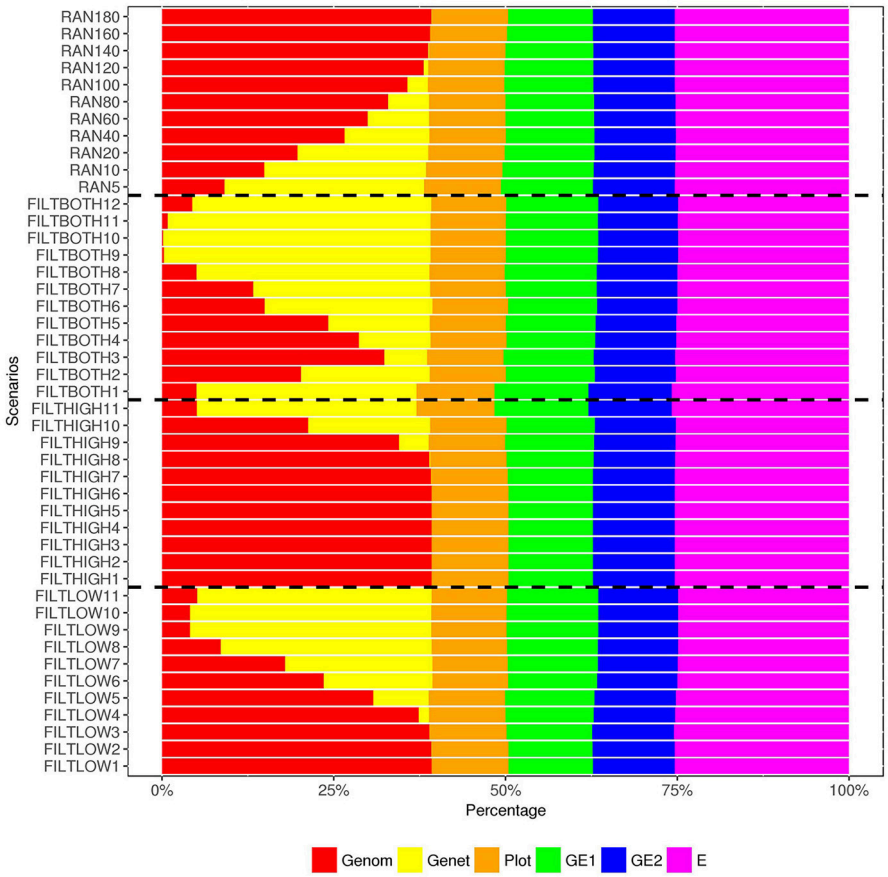

FIGURE 8 | Percentage of variance components[1] over the total phenotypic variance for rust resistance. [1] Genom, Additive genomic variance; Gene, Residual genetic variance; PLOT, random plot variance; GE1, family × sowing year × location × management variance; GE2, family × sowing year × location × management × farming year variance; E, residual environment variance.

depth, based on Cericola et al. (2018), and generally see no more clear linear correlation between sequencing depth and heritability. For instance, the FILTBOTH strategy also grouped SNPs into different depth intervals, and highest estimates of genomic heritability were found for the middle to lower levels FILTBOTH2 (depth 5–10) and FILTBOTH3 (depth 10–20). Comparable heritabilities were found between FILTBOTH 3 and RAN80, where both scenarios included similar amount of SNPs while the later one covering a larger range of sequencing depths (1–278). Hence, the corrections for bias from Cericola et al. (2018) are removing obvious trends related to sequencing depth, and seem to effectively remedy the problem of biased heritabilities at low sequencing depth reported by Ashraf et al. (2016).

For prediction accuracy, the impact from GBS sequencing depth was investigated in simulated biparental segregating populations (Gorjanc et al., 2017) as well as in outbred livestock populations (Gorjanc et al., 2015). The results from these two simulation studies showed that GBS data with low coverage (∼1X) could provide prediction accuracy comparable to SNP array data. When using field data, most of the studies were focused on settings with inbred individuals, e.g., in wheat (Poland et al., 2012) and maize (Crossa et al., 2013). The accuracy of genomic prediction using low-coverage GBS data were comparable with SNP array or diversity array technology data in inbred populations (Poland et al., 2012; Crossa et al., 2013). Different from these simulation studies or studies on inbred populations, the current study is based on the commercial tetraploid data using family-pools. In our data, we cannot confirm that GBS data with low sequencing depth of about 1X already gives accurate predictions. As expected, high heterozygosity in tetraploid ryegrass, combined with use of family-pools, requires higher sequencing depth for accurate genomic prediction. In the current study, SNPs with sequencing depth between 10 and 20 (FILTBOTH3) delivered desirable predictive ability.

In the current study, by filtering out SNPs with either low sequencing depth (FILTLOW) or with high sequencing depth (FILTHIGH), the optimal sequencing depth for practical genomic prediction in tetraploid ryegrass was investigated. In FILTLOW groups, FILTLOW1 to FILTLOW3 gave most accurate predictions. The number of SNPs included in the models with highest predictive ability was about 140–180 k. The lowest sequencing depth for SNPs in FILTLOW3 was 10.

The similar high predictive ability provided by FILTLOW1 to FILTLOW3 indicated that excluding low sequencing depth (1–10) SNPs did not affect the predictive ability significantly. In FILTHIGH groups, FILTHIGH1 to FILTHIGH9 gave similar predictive abilities, which indicated that accurate predictions can be reached even by including only SNPs with sequencing depth lower than 20. This can simply be an effect of still having sufficiently large numbers of SNPs with depth lower than 20, and removing SNPs with high depth may reduce some noise caused by repetitive sequences. Hence, filtering out SNPs with high depth can increase the proportion of useful information without reducing the prediction accuracy. Compared with the RAN strategy, filtering out SNPs with low depth provided higher predictive ability than using a similar number of randomly chosen SNPs, and when comparing the FILTHIGH strategy with the RAN strategy, filtering out SNPs with high depth provided similar predictive ability as using a similar number of randomly chosen SNPs. For the three traits investigated in the current study, the best predictive abilities were not achieved with exactly the same filtering strategy, however, differences between the best filtering strategies were small. In practical breeding, single trait evaluation can be carried out by using **G** matrices built from different sets of SNPs. It is also feasible to apply index selection on a combination of traits with different weights by using a same set of SNPs, which can provide globally accurate predictions. For example, in the current study, even though the highest predictive ability was provided by FILTLOW3, FILTHIGH5, and FILTLOW1 for DM, HD, and RUST, respectively, FILTLOW3 can be chosen as a scenario that provided accurate predictions for all the three traits analyzed. In addition, applying different sets of SNPs at the same time is also achievable by using random regression models disregarding the higher demand of computing resources.

In addition to genomic prediction accuracy, bias was also investigated in this study. In general, it was observed that predictions were biased, and with increasing number of SNPs included in the model, more biased predictions were observed. This can be due to many factors. The definition of the **G** matrix could be one of the reasons. When using GBS data, the allele frequencies can suffer from inaccuracy due to low sequencing depth, which can induce bias into the prediction. However, in the current study, biases due to low sequencing depth was corrected for using the method reported by Cericola et al. (2018). In addition, G×E interactions were modeled in a rather simple way, and bias of prediction may be reduced by better modeling of G×E effects (Fè et al., 2015b).

For diploid heterozygotic species, a minimum sequencing depth of around 10X is needed to obtain accurate calling (Chenuil, 2012). However, for tetraploid species, the requirement of sequencing depth for accurate calling of tetraploid genotypes was reported to be 60–80X (Uitdewilligen et al., 2013). For genomic prediction, however, it is not necessary to obtain accurate calling for each individual sample. The results in the current study indicate that high predictive ability can be obtained using much lower sequencing depth because only the frequency and not the individual genotypes needs to be called.

CONCLUSIONS

In the current study, phenotypic records for three traits dry matter yield (DM), rust resistance (RUST), and heading date (HD), as well as GBS data were used to obtain genomic predictions for 1,515 tetraploid F_2 ryegrass families. Different SNP filtering strategies by filtering out SNPs according to average depth and number of SNPs were compared. The estimates of genomic heritability of family means were 0.45, 0.74, and 0.73 for DM, HD and RUST, respectively. The predictive ability for DM was 0.34, for HD was 0.77, and for RUST was 0.55. The estimation of genomic heritability and the predictive ability for DM, HD and RUST clearly showed that genomic prediction can be implemented in tetraploid perennial ryegrass. Comparison of different filtering strategies showed that using only SNPs with sequencing depth between 10 and 20 would not reduce predictive ability, and showed the potential to obtain accurate prediction from medium-low coverage GBS in tetraploids. Adding SNPs with sequencing depth lower than 10 in the model also lead to accurate predictions. The predictive ability generally increased as the number of SNPs included in the analysis increased. GBS data including 80–100 K SNPs were needed for accurate prediction of additive breeding values in tetraploid ryegrass. The results clearly illustrate that genomic prediction using GBS data can help to optimize the breeding program for tetraploid ryegrass.

AUTHOR CONTRIBUTIONS

XG implemented and carried out the statistical analysis, interpreted the results and had a major role in drafting the manuscript. FC and DF contributed to the statistical analysis and the result interpretation, reviewed the manuscript. MP managed the production of the plant material and the acquisition of the phenotypic data. IL carried out part of the GBS sequencing, and subsequent bioinformatics analysis. CJ conceived the experiment and reviewed the manuscript. JJ conceived the experiment, contributed in developing the statistical models and interpreting the results, reviewed the manuscript. LJ contributed in developing the statistical models and interpreting the results, reviewed the manuscript. All authors read and approved the final manuscript.

FUNDING

This project was funded by The Danish Council for Strategic Research (http://www.fivu.dk/en/dsf/), Center for Genomic Selection in Animals and Plants (GenSAP), grant number 12-132452.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01165/full#supplementary-material

REFERENCES

Ashraf, B. H., Byrne, S., Fe, D., Czaban, A., Asp, T., Pedersen, M. G., et al. (2016). Estimating genomic heritabilities at the level of family-pool samples of perennial ryegrass using genotyping-by-sequencing. *Theor. Appl. Genet.* 129, 45–52. doi: 10.1007/s00122-015-2607-9

Ashraf, B. H., Jensen, J., Asp, T., and Janss, L. L. (2014). Association studies using family pools of outcrossing crops based on allele-frequency estimates from DNA sequencing. *Theor. Appl. Genet. Theoretische Und Angewandte Genetik* 127, 1331–1341. doi: 10.1007/s00122-014-2300-4

Beissinger, T. M., Hirsch, C. N., Sekhon, R. S., Foerster, J. M., Johnson, J. M., Muttoni, G., et al. (2013). Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics* 193, 1073–1081. doi: 10.1534/genetics.112.147710

Byrne, S., Czaban, A., Studer, B., Panitz, F., Bendixen, C., and Asp, T. (2013). Genome wide allele frequency fingerprints (GWAFFs) of populations via genotyping by sequencing. *PLoS ONE* 8:e57438. doi: 10.1371/journal.pone.0057438

Byrne, S. L., Nagy, I., Pfeifer, M., Armstead, I., Swain, S., Studer, B., et al. (2015). A synteny-based draft genome sequence of the forage grass *Lolium perenne*. *Plant J.* 84, 816–826. doi: 10.1111/tpj.13037

Cericola, F., Lenk, I., Fè, D., Byrne, S., Jensen, C. S., Pedersen, M. G., et al. (2018). Optimized use of low-depth genotyping-by-sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (*Lolium perenne* L.). *Front. Plant Sci.* 9:369. doi: 10.3389/fpls.2018.00369

Chenuil, A. (2012). How to infer reliable diploid genotypes from NGS or traditional sequence data: from basic probability to experimental optimization. *J. Evol. Biol.* 25, 949–960. doi: 10.1111/j.1420-9101.2012.02488.x

Conaghan, P., and Casler, M. D. (2011). A theoretical and practical analysis of the optimum breeding system for perennial ryegrass. *Irish J. Agric. Food Res.* 50, 47–63. Available online at: http://www.jstor.org/stable/41348155

Cornish, M. A., Hayward, M. D., and Lawrence, M. J. (1979). Self-incompatibility in ryegrass I. Genetic control in diploid *Lolium perenne L. Heridity* 43, 95–106. doi: 10.1038/hdy.1979.63

Crossa, J., Beyene, Y., Kassa, S., Perez, P., Hickey, J. M., Chen, C., et al. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3 (Bethesda)* 3, 1903–1926. doi: 10.1534/g3.113.008227

Crossa, J., Campos Gde, L., Perez, P., Gianola, D., Burgueno, J., Araus, J. L., et al. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186, 713–724. doi: 10.1534/genetics.110.118521

Crossa, J., Perez, P., Hickey, J., Burgueno, J., Ornella, L., Ceron-Rojas, J., et al. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity (Edinb)* 112, 48–60. doi: 10.1038/hdy.2013.16

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

Fè, D., Ashraf, B. H., Pedersen, M. G., Janss, L., Byrne, S., Roulund, N., et al. (2016). Accuracy of genomic prediction in a commercial perennial ryegrass breeding program. *Plant Genome* 9. doi: 10.3835/plantgenome2015.11.0110

Fè, D., Cericola, F., Byrne, S., Lenk, I., Ashraf, B. H., Pedersen, M. G., et al. (2015a). Genomic dissection and prediction of heading date in perennial ryegrass. *BMC Genomics* 16:921. doi: 10.1186/s12864-015-2163-3

Fè, D., Pedersen, M. G., Jensen, C. S., and Jensen, J. (2015b). Genetic and environmental variation in a commercial breeding program of perennial ryegrass. *Crop Sci.* 55, 631–640. doi: 10.2135/cropsci2014.06.0441

Garner, C. (2011). Confounded by sequencing depth in association studies of rare alleles. *Genet. Epidemiol.* 35, 261–268. doi: 10.1002/gepi.20574

Gorjanc, G., Cleveland, M. A., Houston, R. D., and Hickey, J. M. (2015). Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genet. Sel. Evol.* 47:12. doi: 10.1186/s12711-015-0102-z

Gorjanc, G., Dumasy, J.-F., Gonen, S., Gaynor, R. C., Antolin, R., and Hickey, J. M. (2017). Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. *Crop Sci.* 57, 1404–1420. doi: 10.2135/cropsci2016.08.0675

Hayes, B. J., Cogan, N. O. I., Pembleton, L. W., Goddard, M. E., Wang, J., Spangenberg, G. C., et al. (2013). Prospects for genomic selection in forage plant species. *Plant Breed.* 132, 133–143. doi: 10.1111/pbr.12037

Hickey, J. M., Chiurugwi, T., Mackay, I., Powell, W., and Implementing Genomic Selection in CGIAR Breeding Programs Workshop Participants (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* 49, 1297–1303. doi: 10.1038/ng.3920

Humphreys, M. O. (2005). Genetic improvement of forage crops–past, present and future. *J. Agric. Sci.* 143, 441–448. doi: 10.1017/s0021859605005599

Jensen, C. S., Salchert, K., and Nielsen, K. K. (2001). A Terminal flower1-like gene from perennial ryegrass involved in floral transition and axillary meristem identity. *Plant Physiol.* 125, 1517–1528. doi: 10.1104/pp.125.3.1517

Lantinga, E. A., and Groot, J. C. J. (1996). Optimization of grassland production and herbage feed quality in an ecological context. *EAAP Publication* 84, 58–67.

Legarra, A., Robert-Granié, C., Manfredi, E., and Elsen, J.-M. (2008). Performance of genomic selection in mice. *Genetics* 180, 611–618. doi: 10.1534/genetics.108.088575

Lin, Z., Cogan, N. O. I., Pembleton, L. W., Spangenberg, G. C., Forster, J. W., Hayes, B. J., et al. (2016). Genetic gain and inbreeding from genomic selection in a simulated commercial breeding program for perennial ryegrass. *Plant Genome* 9. doi: 10.3835/plantgenome2015.06.0046

Madsen, P., and Jensen, J. (2013). *A User's Guide to DMU*. Tjele: University of Aarhus, Faculty Agricultural Sciences (DJF), Department of Genetics and Biotechnology Research Centre Foulum. Available online at: http://dmu.agrsci.dk/DMU/Doc/Current/

McDonagh, J., O'Donovan, M., McEvoy, M., and Gilliland, T. J. (2016). Genetic gain in perennial ryegrass (*Lolium perenne*) varieties 1973 to 2013. *Euphytica* 212, 187–199. doi: 10.1007/s10681-016-1754-7

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829.

O'Donovan, M., and Delaby, L. (2005). A comparison of perennial ryegrass cultivars differing in heading date and grass ploidy with spring calving dairy cows grazed at two different stocking rates. *Anim. Res.* 54, 337–350. doi: 10.1051/animres:2005027

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome J.* 5, 103–113. doi: 10.3835/plantgenome2012.06.0006

Poland, J. A., and Rife, T. W. (2012). Genotyping-by-Sequencing for Plant Breeding and Genetics. *Plant Genome J.* 5, 92–102. doi: 10.3835/plantgenome2012.05.0005

Ravel, C., and Charmet, G. (1996). A comprehensive multisite recurrent selection strategy in perennial ryegrass. *Euphytica* 88, 215–226. doi: 10.1007/BF00023893

Sampoux, J.-P., Baudouin, P., Bayle, B., Béguier, V., Bourdon, P., Chosson, J.-F., et al. (2011). Breeding perennial grasses for forage usage: an experimental assessment of trait changes in diploid perennial ryegrass (*Lolium perenne* L.) cultivars released in the last four decades. *Field Crops Res.* 123, 117–129. doi: 10.1016/j.fcr.2011.05.007

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15, 121–132. doi: 10.1038/nrg3642

Uitdewilligen, J. G., Wolters, A. M., D'Hoop, B. B., Borm, T. J., Visser, R. G., and van Eck, H. J. (2013). A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS ONE* 8:e62355. doi: 10.1371/journal.pone.0062355

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Waldron, B. L., Ehlke, N. J., Wyse, D. L., and Vellekson, D. J. (1998). Genetic variation and predicted gain from selection for winterhardiness

and turf quality in a perennial ryegrass topcross population. *Crop Sci.* 38, 817–822.

Wilkins, P. W. (1991). Breeding perennial ryegrass for agriculture. *Euphytica* 52, 201–214. doi: 10.1007/BF00029397

Wilkins, P. W., and Humphreys, M. O. (2003). Progress in breeding perennial forage grasses for temperate agriculture. *J. Agric. Sci.* 140, 129–150. doi: 10.1017/s0021859603003058

Zhang, X., Perez-Rodriguez, P., Semagn, K., Beyene, Y., Babu, R., Lopez-Cruz, M. A., et al. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity (Edinb)* 114, 291–299. doi: 10.1038/hdy.2014.99

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: 07 August 2018
doi: 10.3389/fpls.2018.01149

The Capacity to Buffer and Sustain Imbalanced D-Subgenome Chromosomes by the BBAA Component of Hexaploid Wheat Is an Evolved Dominant Trait

Xin Deng[1], Yan Sha[1], Zhenling Lv[1], Ying Wu[1], Ai Zhang[1], Fang Wang[2] and Bao Liu[1]**

[1] Key Laboratory of Molecular Epigenetics of the Ministry of Education (MOE), Northeast Normal University, Changchun, China, [2] College of Oceanology and Food Science, Quanzhou Normal University, Quanzhou, China

OPEN ACCESS

Edited by:
Yiwei Jiang,
Purdue University, United States

Reviewed by:
Lifeng Zhu,
Nanjing Normal University, China
Ekaterina D. Badaeva,
Vavilov Institute of General Genetics (RAS), Russia
Fangpu Han,
Institute of Genetics and Developmental Biology (CAS), China

***Correspondence:**
Fang Wang
dwf320@163.com
Bao Liu
baoliu@nenu.edu.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *22 February 2018*
Accepted: *18 July 2018*
Published: *07 August 2018*

Citation:
Deng X, Sha Y, Lv Z, Wu Y, Zhang A, Wang F and Liu B (2018) The Capacity to Buffer and Sustain Imbalanced D-Subgenome Chromosomes by the BBAA Component of Hexaploid Wheat Is an Evolved Dominant Trait. Front. Plant Sci. 9:1149. doi: 10.3389/fpls.2018.01149

Successful generation of pentaploid wheat (genome, BBAAD) via interspecific hybridization between tetraploid wheat (BBAA) and hexaploid wheat (BBAADD) holds great promise to mutually exchange desirable traits between the two cultivated wheat species, as well as providing a novel facet for evolutionary studies of polyploid wheat. Taking advantage of the viable and fertile nature of an extracted tetraploid wheat (ETW) with a BBAA genome that is virtually identical with the BBAA component of a hexaploid common wheat, and a synthetic hexaploid wheat, we constructed four pentaploid wheats with several distinct yet complementary features, of which harboring homozygous BBAA subgenomes is a common feature. By using a combined FISH/GISH method that enables diagnosing all individual wheat chromosomes, we precisely karyotyped a larger number of cohorts from the immediate progenies of each of the four pentaploid wheats. We found that the BBAA component of hexaploid common wheat possesses a significantly stronger capacity to buffer and sustain imbalanced D genome chromosomes and appears to harbor more structural chromosome variations than the BBAA genome of tetraploid wheat. We also document that this stronger capacity of the hexaploid BBAA subgenomes behaves as a genetically controlled dominant trait. Our findings bear implications to the known greater than expected level of genetic diversity in, and the remarkable adaptability of, hexaploid common wheat as a staple crop of global significance, as well as in using pentaploidy as intermediates for reciprocal introgression of useful traits between tetraploid and hexaploid wheat cultivars.

Keywords: pentaploidy, aneuploidy, polyploidy, dosage imbalance, buffering capacity, genetic diversity, evolution, *Triticum*

INTRODUCTION

Aneuploidy, with losses and/or gains of individual chromosomes, and hence deviating from the default balanced chromosome complement(s) of any species, is a large-effect genetic variant with profound biological consequences. Aneuploidy generally causes compromised fitness and is the causation of many important human diseases, e.g., the Down Syndrome due to trisomy

of chromosome 21 (Megarbane et al., 2009; Letourneau et al., 2014). However, aneuploidy is not always associated with reduced cellular or organismal fitness. Recent studies have revealed aneuploidy as a major mechanism underlying adaption in unicellular microbes especially under strong selection (Rancati et al., 2008; Pavelka et al., 2010; Kaya et al., 2015; Millet et al., 2015). Also, certain unbalanced karyotypes may promote cellular growth and are principle drivers for the evolution of many types of cancers (Torres et al., 2008; Rutledge and Cimini, 2016; Sansregret and Swanton, 2017). Furthermore, recent studies have documented that aneuploidy is common and often persistent (rather than transit as commonly thought) in newly formed plant polyploids (Xiong et al., 2011; Chester et al., 2012; Zhang et al., 2013), which has led to the suggestion that numerical chromosome changes (aneuploidy) may have played a protracted role at the initial stages of polyploid establishment, adaptation, and evolution (Soltis et al., 2015). This possibility is bolstered by the observation that many types of aneuploidy are reversible in the sense that euploidy progenitors can be readily generated from aneuploid progenitors, which however may retain some of the desirable properties of the aneuploid progenitors (Henry et al., 2010). This is consistent with our recent observation that aneuploidy-induced epigenetic modifications in the form of altered DNA methylation were imparted to the aneuploidy-derived euploid progenies at appreciable frequencies (Gao et al., 2016).

The capacity to harbor additional chromosome(s) in the form of aneuploidy is a property intrinsically different between species. For example, plants in general are more tolerant to an imbalanced genome composition than animals (Matzke et al., 2003; Henry et al., 2005). Within a given species, a polyploid genome is more permissive to unbalanced chromosomes (especially chromosome loss) than its diploid or haploid counterparts (Ramsey and Schemske, 1998; Birchler et al., 2001; Wu et al., 2018). Indeed, in several polyploid crops, such as common wheat, complete sets of aneuploidies can be generated and maintained (Sears, 1944).

A recent study documented that there exists a fundamental difference between laboratory strains and wild collections in budding yeast (*Saccharomyces cerevisiae*) with respect to their capacities to harbor additional chromosomes, i.e., being aneuploidy (Hose et al., 2015). Specifically, it was found that laboratory strains of *S. cerevisiae* were poorly tolerant to numerical chromosome variation (NCV), while their wild counterparts showed little detrimental impacts when extra chromosome were present (Hose et al., 2015). A major mechanism underlying this disparity in harboring an additional chromosome between the laboratory and wild yeast strains is due to their difference in dosage compensation whereby expression level of genes encoded by the extra chromosome can be attenuated to that of the euploidy in wild but not in laboratory strains (Gasch et al., 2016; but see Torres et al., 2016). This finding in yeast is reminiscent of earlier studies in *Arabidopsis thaliana*, which already documented that a locus named *SDI* (sensitive to dosage imbalance) located on chromosome 1 affects ploidy-dependent transmission distortion, has a role in aneuploid viability, and hence impacts chromosome composition of triploid-derived progeny cohorts (Henry et al., 2005, 2007). Together, it is clear that the capacity to buffer and retain extra chromosome(s), i.e., differences in sustaining the severity and diversity of aneuploidies, is a genetically controlled trait in a given organism, and the phenotypic penetrance of which is likely ploidy level-dependent. However, the issue remains understudied in any organism.

Hexaploid common wheat (*Triticum aestivum* L., genome BBAADD) is a very young (*ca.* 8,500 year-old) species formed by allopolyploidization [hybridization and whole genome duplication (WGD)] between tetraploid emmer wheat (*Triticum turgidum*, genome BBAA) and diploid *Aegilops tauschii* (genome DD) (Kihara, 1944; McFadden and Sears, 1946; Feldman et al., 1995; Huang et al., 2002). Nearly a century ago, the pioneering works by Sax and Kihara have independently demonstrated that hexaploid common wheat and tetraploid emmer wheat could be hybridized to produce fertile pentaploid hybrids (genome BBAAD) (Sax, 1922; Kihara, 1925). Recent years have witnessed a renewed interest in the generation of pentaploid wheat for the purpose of reciprocally introgressing desirable traits from one species to the other (reviewed in Padmanaban et al., 2017b, 2018). Apart from the practical success (Wang et al., 2005; Eberhard et al., 2010; Padmanaban et al., 2017a,b), an interesting observation is that, in the progenies of pentaploid wheat, there exists a significant positive correlation between proportions of the hexaploid wheat BBAA genomic content and the retention of unbalanced D chromosome (Martin et al., 2011; Padmanaban et al., 2017a). This finding suggests that the BBAA components of hexaploid wheat have a stronger capacity to retain unbalanced D chromosomes than that of the tetraploid wheat BBAA genome. However, the experimental design in these studies, being breeding-oriented, all involved heterozygous BBAA genomes due to combining different genotypes of hexaploid and tetraploid wheat, and hence, does not allow a direct comparison to reach a confirmative conclusion.

With largely intact subgenomes, as well as having both progenitor species still being in extant, it is possible to extract the BBAA component from a given hexaploid wheat cultivar, by hybridization with a *durum* wheat and repeated backcrossing with the hexaploid wheat as a recurrent parent, to reconstitute an "extracted" tetraploid wheat (Kerber, 1964). The extracted tetraploid wheat (ETW) is viable and partially fertile although with severe pleiotropic growth and development abnormalities (Kerber, 1964; Zhang et al., 2014; Liu et al., 2015). Thus, crossing a hexaploid common wheat genotype, from which the ETW was extracted, with ETW will generate a pentaploid wheat with homozygous BBAA subgenomes representing that of the hexaploid wheat. Accordingly, crossing a synthetic hexaploid with the same tetraploid wheat cultivar whereby the hexaploid wheat was produced will generate pentaploid wheat with homozygous BBAA genomes representing that of the tetraploid wheat genome. The present study was designed according to this rational, along with additional considerations, to construct four pentaploid wheat lines with distinct features that are suitable to specifically address the question, i.e., whether the capacity to buffer and sustain imbalanced D-genome

chromosomes by the BBAA component of hexaploid wheat is an evolved trait. Our results, based on high resolution FISH/GISH karyotyping of large numbers of immediate progeny cohorts of each of the four pentaploid wheat lines, have confirmed the previously only implied possibility (Martin et al., 2011; Padmanaban et al., 2017a). Our results also provide additional insights into the extent and trend of numerical and structural chromosome instabilities in the pentaploid wheat-derived progenies.

MATERIALS AND METHODS

Plant Materials

We used two genotypes of tetraploid wheat and three genotypes of hexaploid wheat to construct four pentaploid wheat lines (genome BBAAD, 2*n* = 35). TTR13 (*T. turgidum,* ssp. *durum*, BBAA, 2*n* = 28) represents the natural *durum* wheat. The ETW (BBAA, 2*n* = 28) represents the BBAA component of a natural hexaploid bread wheat (*T. aestivum* L., BBAADD, 2*n* = 42) cultivar, TAA10, as detailed previously (Kerber, 1964; Zhang et al., 2014). By hybridization with a *durum* wheat, and then recurrent backcrossing with TAA10 for 9-times (Zhang et al., 2014), ETW could be regarded as BBAA component of TAA10, as they are >98% identical based on recombination and Mendelian inheritance (Zhang et al., 2014). XX329 (2*n* = 42, BBAADD) is a resynthesized hexaploid wheat line by hybridizing and doubling hybrid of ETW and TQ18 (*A. tauschii*, 2*n* = 14, DD) (Zhang et al., 2014). Synthetic hexaploid wheat (SHW) (BBAADD, 2*n* = 42) represents a newly synthesized hexaploid wheat lines by crossing TTR13 with TQ18 followed by spontaneous genome doubling of the F1 hybrid.

Four independent pentaploid wheat lines were separately generated from the hexaploid × tetraploid combinations. These include pentaploid XE (XX329 × ETW), pentaploid TE (TAA10 × ETW), pentaploid TT (TAA10 × TTR13) and pentaploid ST (SHW × TTR13), and in all four lines, the hexaploid wheat was used as the maternal parent. All pentaploid wheat individuals were obtained by embryo rescue by inoculating the inter-ploidy F1 hybrid immature embryos on Murashige and Skoog (MS) medium (PhytoTechnology, M519) at a constant temperature of 25°C. The germinated seedlings were transplanted to soil when they were 6–7 cm in height.

Seeds were harvested from several individuals of each pentaploid wheat line. All grown seedlings from the germinated seeds of each progeny individual of the pentaploid wheats were referred to as immediate selfed progenies. They were transferred to pots containing nutrient-sufficient soil under normal greenhouse conditions of constant 25°C with 16/8 h light/dark photoperiod. Root-tips were sampled for cytological analysis, while the plants were grown to maturity for phenotyping.

Chromosome Preparation and Fluorescence *in situ* Hybridization

Roots about 1.5∼2.0 cm long were taken from the seedlings and treated in N_2O gas for 2 h. These roots were then fixed in 90% (v/v) acetic acid and stored in 75% (v/v) alcohol at −20°C until use. Mitotic spread chromosome slides were prepared from the root-tips according to Zhang et al. (2013). For each plant, chromosomes were counted for at least five well-spread metaphase cells.

Sequential florescence *in situ* hybridization (FISH) and genomic *in situ* hybridization (GISH) were performed according to a protocol reported previously (Zhang et al., 2013) with minor modifications. Briefly, Texas red-5-dCTP (PerkinElmer, NEL426001EA) was used for labeling the repeated sequence clone pAs1 (Rayburn and Gill, 1986) and genomic DNA isolated from *A. tauschii*, respectively. ChromaTide™ Alexa Fluor™ 488-5-dUTP (Thermo Fisher, C11397) was used for labeling the rye repeated sequence clone pSc119.2 and genomic DNA from *T. urartu*, respectively. In GISH, genomic DNA of *A. bicornis* was used as blocker. In both FISH and GISH, DAPI (Vector, H-1200) was also used to counterstain the chromosomes.

Both the FISH and GISH slides were examined under an Olympus BX63 fluorescence microscope and captured by Q-capture imaging software (QImaging, Version 2.90.1). Brightness, contrast and background were adjusted as an entirety in Adobe Photoshop CC.

Phenotyping

Nine morphological traits, including plant height, tiller number, stem diameter, flag-leaf width, spike length, spike density, seed length, seed width, and seed set, were phenotyped for at least three individuals from each of the pentaploid wheat lines. Plant height, tiller number, stem diameter, and flag-leaf width were measured at the mature stage. Stem diameter was represented by the maximum value of stems at 1.5 cm below the last node. Spike length, seed length, and seed width were measured after harvest. The longest spike was used to measure the spike length and spikelet number. Spikelet density was calculated through spikelet number divided by the corresponding spike length. We also used the maximum value of seed set per spike to represent seed set of a given plant. Seed set was the total seed number of two base florets relative to the total number of two base florets. Seed length and width were represented by the average values of the same 10 seeds.

Data Analysis and Statistical Test

Statistical test of each of the data comparisons was performed in R software (version 3.4.0). Vioplots of chromosome number distribution was depicted using R packages of vioplot (Hintze and Nelson, 1998). Karyotypes of each plant were depicted by R packages of pheatmap[1]. The Student's *t*-test and F-test were used to interrogate whether the capacities to buffer and sustain unbalanced D chromosomes by the BBAA component were significantly different among the four pentaploid wheat lines. A prop.test was performed to determine possible differences in composition of the D chromosomes within a given pentaploid wheat line via pairwise comparisons between any two D chromosomes. A Wilcoxon test was used to determine differences in each of the nine morphological traits among the four pentaploid wheat lines.

[1]https://CRAN.R-project.org/package=pheatmap

RESULTS

Features of the Four Constructed Pentaploid Wheat Lines

We produced four lines of pentaploid wheat (genome BBAAD) using three hexaploid wheat genotypes (genome BBAADD) and two tetraploid wheat genotypes (genome BBAA) (**Figure 1A**). As such, three pentaploid wheat lines (designated XE, TE, and ST) harbor *homozygous* BBAA genomes plus a single D genome (**Figure 1A**). This is because the maternal hexaploid wheat genotype and paternal tetraploid genotype for each of these three pentaploid lines contained the same BBAA component (**Figure 1A**). The fourth pentaploid line (designated TT) harbors heterozygous BBAA genome with the BBAA components being a F1 hybrid between the maternal hexaploid genotype and the paternal tetraploid genotype (**Figure 1A**). Two additional features characterize the three pentaploid wheat lines with homozygous BBAA genomes. First, the D genome in two pentaploid lines, XE and TE, are from different *A. tauschii* accessions; second, pentaploid lines XE and ST share the same D genome while their homozygous BBAA genomes are different (**Figure 1A**). Together, characteristics of these four purposely constructed pentaploid lines render them suitable for addressing the primary objective of this study, i.e., weather the capacity to buffer and sustain imbalanced D-genome chromosomes by the BBAA component of hexaploid wheat is an evolved trait.

Similarity and Difference in Overall Chromosome Number Distribution by the Immediate Selfed Progenies of the Four Pentaploid Wheat Lines

As described above, we selfed each of the four pentaploid lines (XE, TE, TT, and ST) (**Figure 1A**) to produce their immediate progeny swarms (**Figure 1B**). We karyotyped 234, 196, 181, and 169 progeny individuals of XE, TE, TT, and ST, respectively, by the sequential FISH/GISH karyotyping protocol, which enables reliable diagnosis of each of the wheat chromosomes (Zhang et al., 2013). The overall chromosome number distribution in the four pentaploid progeny swarms, each as a population, was tabulated. Data showed that the chromosome distributions in progeny swarms from three (XE, TE, and TT) of the four pentaploid lines are similar, but that of the fourth pentaploids (ST) is strikingly different (**Figure 1B**). Specifically, while chromosome numbers in the progeny swarms of XE, TE, and TT showed more or less similar quantities of individuals with chromosome numbers fewer or more than $2n = 35$, those of ST substantially biased toward individuals (93.5%) with chromosome numbers fewer than $2n = 35$, moreover, 33.7% of the plants have reverted to tetraploid wheat, i.e., with all D chromosomes being eliminated (**Figure 1B**). Expectedly, in the progeny swarms of all four pentaploid lines, chromosomes of the A and B subgenomes were found to be stable (i.e., they all contained seven pairs of chromosomes for each subgenome) in great majority (>92.8%) of the plants; the small proportions of plants with one or more BBAA chromosomes being in an aneuploid state and/or with structural variations in each population are described separately in later sections and not included here. Thus, only the D subgenome chromosomes are variable and contributing to differences in the overall chromosome number distribution. Taken together, these observations suggest that while the BBAA components of three pentaploids, XE, TE, and TT, showed similar and strong capacities to buffer and sustain unbalanced D chromosomes, this capacity by the BBAA component of ST is significantly weaker (*T*-test, both *P*-values < 0.05 in the comparisons of XE versus ST, TE versus ST, and TT versus ST).

In theory, assuming random assortment of the unpaired D chromosomes during formation of male and female gametes in the pentaploid lines, the chromosome numbers of their progeny swarms should range from 28 (reverting to tetraploid wheat) to 42 (converting to hexaploid wheat) (**Figure 1C**). Compared to this theoretical chromosome number distribution, the observed chromosome numbers in the progeny swarms of three pentaploid lines (XE, TT, and ST) all showed significant deviation ($\chi^2 = 50.94$, d.f. = 14, *P*-value = 4.23e-06 for progenies of XE; $\chi^2 = 46.22$, d.f. = 14, *P*-value = 2.58e-05 for progenies of TT; $\chi^2 = 1524.90$, d.f. = 14, *P*-value < 2.2e-16 for progenies of ST). Interestingly, chromosome number distribution in progeny swarms of the pentaploid line TE was not significantly deviating from the theoretically expected distribution ($\chi^2 = 20.13$, d.f. = 14, *P*-value = 0.13), suggesting a strong parental combination difference. In any case, although the three pentaploid lines (XE, TE, and TT) are different with respect to their conformities to the theoretical chromosome number distribution in their progeny swarms, their BBAA components manifested similar capacities to buffer and sustain the unbalanced D chromosomes, and which was significantly different from that of ST (**Figure 1D**).

Similarity and Difference in Chromosome Composition of the Immediate Selfed Progenies of the Four Pentaploid Wheat Lines

The foregoing results concern the overall chromosome number distributions in progeny swarms of the four pentaploid lines. We further analyzed the exact chromosome compositions in each of these progeny plants. Because all plants for this analysis contained the complete A- and B-subgenome chromosomes, described above, the only variable is the number of D-subgenome chromosomes. Data showed that almost each D chromosome-containing progeny individual of all four pentaploid lines contained a different "D chromosome composition" configured by both individuality, and copy number thereof, of the harbored D chromosome(s) (**Figure 2A**). Consequently, vast karyotypic diversity was seen in these progeny plants due to NCV of the D chromosome(s), with those of pentaploids XE, TE, and TT being markedly more diverse than those of ST (**Figure 2A**), which is consistent with the distributions of overall chromosome numbers (**Figure 1D**). An additional feature characteristic of the progeny plants of pentaploids XE, TE, and TT is that their substantial proportions contained two copies of the D chromosomes while this situation was rare in progeny plants of ST (**Figure 2A**).

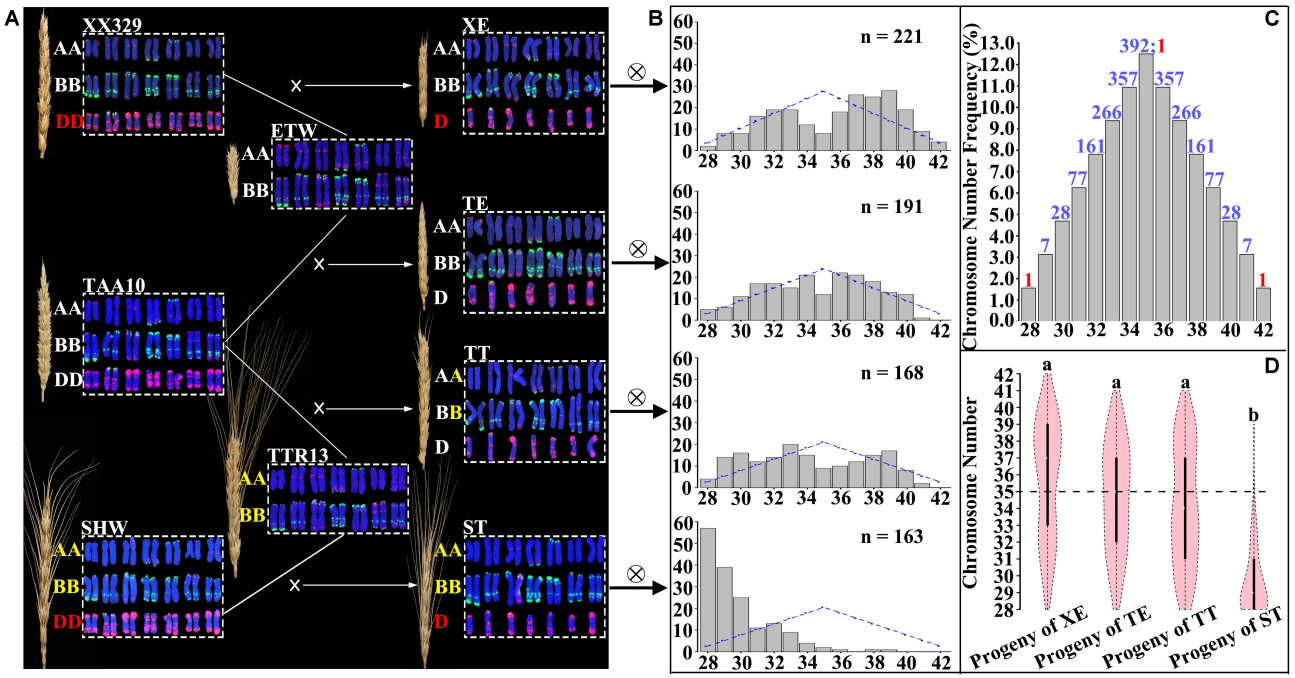

FIGURE 1 | Four different pentaploid wheat lines generated by crossing three hexaploid wheat genotypes with two tetraploid wheat genotypes, and the overall chromosome number distributions in the immediate progeny swarms of the pentaploids. **(A)** FISH-based karyotypes and spike morphology of hexaploid wheat, tetraploid wheat and pentaploid wheat. PSc119.2 (green coloration) and pAs1 (red coloration) were used as FISH probes. Subgenome designations **(A,B,D)** in the same and different colored fonts denote that they were from the same source or difference sources. **(B)** Theoretically expected (blue lines) and observed (gray bar) plant numbers in each of the immediate progenies of the four pentaploid wheat lines. Expected values were obtained by transforming the total karyotyped plant number according to the expected frequency in **(C)**. The numbers of observed (karyotyped) individuals in each F2 population are given. The x-axis is the chromosome number. **(C)** Theoretical (expected) chromosome number distribution in the immediate progeny swarms of pentaploid wheat. The numbers on the top of the gray bars refer to the expected values in each of the chromosome number categories belonging to either aneuploidy (colored blue) or euploidy (colored red). **(D)** Comparison of the overall chromosome number distributions among the immediate progeny swarms of the four pentaploid wheat lines. Letters above each violplots denote statistical significance of pairwise comparisons (Var. test, P-values < 0.05).

We found that occurrence of the D-subgenome chromosomes was not equal in the progeny plants of a given pentaploid line (**Figure 2B**). For progeny plants of XE, chromosome 5D was significantly overrepresented relative to the rest five D chromosomes, i.e., 1D, 2D, 3D, 4D, and 6D, among which the difference was not statistically significant; for progeny plants of TE, chromosomes 1D and 5D were significantly overrepresented than chromosomes 2D, 3D, and 4D, while no difference was detected in pairwise comparisons concerning the rest D chromosomes; for progeny plants of TT, chromosomes 1D and 5D were significantly overrepresented than chromosome 3D, while no difference was detected in the rest pairwise comparisons; for progeny plants of ST, chromosomes 2D was significantly overrepresented than chromosomes 1D, 3D, 4D, and 6D, while no difference was detected in the pairwise comparisons for the rest D chromosomes. However, all D chromosomes occurred at low frequencies in the progeny plants of ST relative to those of the other three pentaploid lines (**Figure 2B**). Notably, for still unknown reasons, chromosome 5D was significantly overrepresented in progenies of the three pentaploid wheats (XE, TE, and TT). Another striking feature is, in the progeny plants of all four pentaploid lines, the 1 copy versus 2 copies for a given retained D chromosome appeared proportional (**Figure 2B**), suggesting that if selection has been in action at the gametophytic stage, then it has been unbiased between the female and male gametes, an issue warrants further investigations.

Difference in the Capacity to Buffer and Sustain Imbalanced D-Genome Chromosomes by the BBAA Components of Hexaploid Wheat With Different Evolutionary Histories

The above results have shown that both the overall chromosome number distributions and chromosome compositions are similar among the immediate progeny swarms of the three pentaploids, XE, TE, and TT, but which are strikingly different from those of ST (**Figures 1A,B**). Naturally, we analyzed these differences in relation to the BBAA components constituting the pentaploid lines, which have different evolutionary histories. Specifically, of the three pentaploid wheat lines (XE, TE, and ST) harboring homozygous BBAA genomes (**Figure 1A**), two (XE and TE) contained the BBAA component of a common wheat, cv. TAA10; this component was extracted from TAA10 by hybridization and repeated backcrossing with TAA10 (Kerber, 1964;

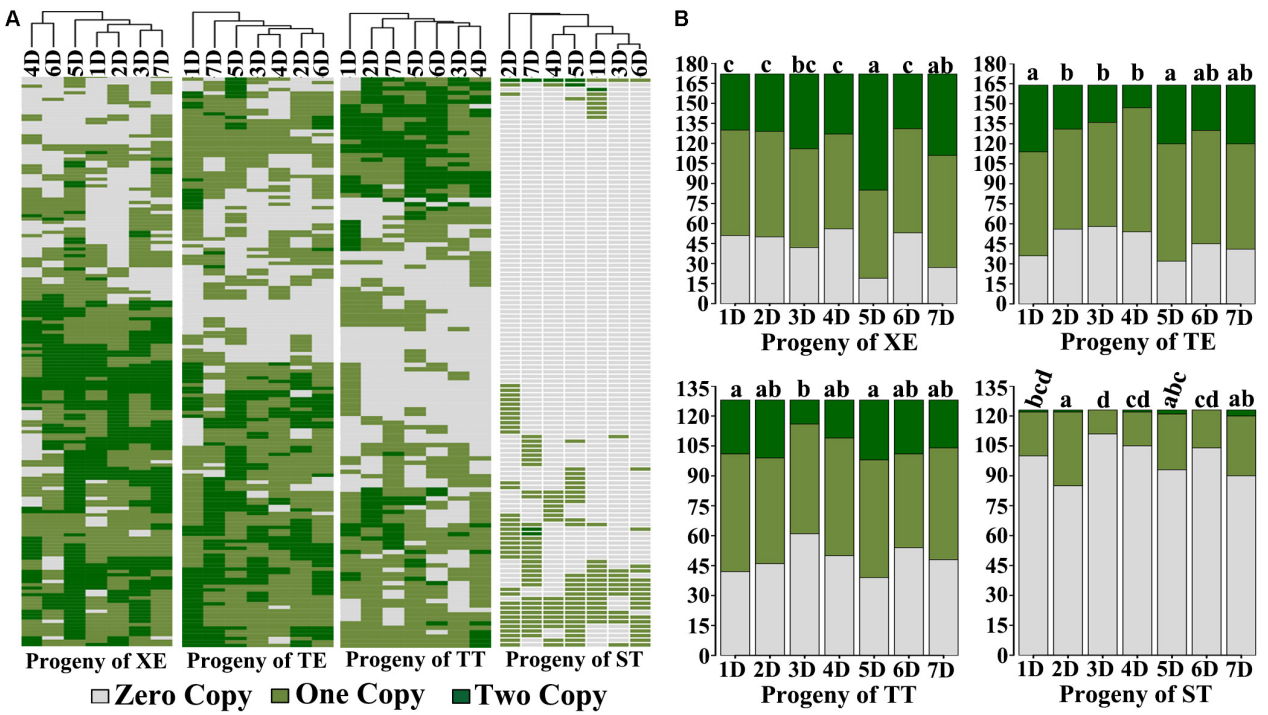

FIGURE 2 | Exact D-subgenome chromosome composition in the immediate progeny swarms of the four pentaploid wheat lines. **(A)** Heatmaps illustrating the D chromosome composition in each karyotyped cohort of the immediate selfed progeny swarms of the four pentaploid wheat lines. **(B)** Bar-plots showing the summarized copy number for each of the seven D chromosomes (1D–7D) in the immediate selfed progeny swarms of the four pentaploid wheat lines. The small letter(s) above each bar plot denotes for statistical difference of pairwise comparisons (Prop.test, *P*-values < 0.05). The y-axis in both **(A,B)** represents the number of karyotyped cohorts, while the x-axis is the seven D chromosomes.

Zhang et al., 2014), and hence was termed ETW (Zhang et al., 2014). Because the hexaploid wheat XX329 is a resynthesized line via allohexaploidization (hybridization coupled with WGD) between ETW and *A. tauschii* (accession TQ18) (Zhang et al., 2014), the D genome in the derived pentaploid wheat XE is of diploid *A. tauschii* origin (accession TQ18). In contrast, the D genome in TE was that of the original common wheat TAA10 (**Figure 1A**). The third pentaploid line harboring homozygous BBAA genomes is ST (**Figure 1A**). The hexaploid parent SHW (synthetic hexaploid wheat) was a synthetic line constructed via allohexaploidization between *durum* wheat cv. TTR13 and *A. tauschii* (accession TQ18), hence the BBAA component in ST was the same as TTR13 while the D genome was that of TQ18, i.e., the same as that in XE (**Figure 1A**). Thus, while the BBAA components of pentaploid lines XE and TE are of domesticated common wheat (cv. TAA10), that of ST is the tetraploid *durum* wheat (cv. TTR13). A fundamental difference between the BBAA component of hexaploid common wheat and the BBAA genomes of *durum* wheat is that the former has been co-evolved with the D subgenome for *ca.* 8,500 years since speciation of common wheat, *T. aestivum* L. (Kihara, 1944; McFadden and Sears, 1946; Feldman et al., 1995; Huang et al., 2002), while the later has never been co-exiting with the D genome. Therefore, our observation that the BBAA component in pentaploid lines XE and TE has a strong capacity to buffer and sustain unbalanced D chromosomes, while that in pentaploid line ST has a much weaker capacity can be most parsimoniously explained by their different properties, i.e., the former has co-evolved with the D subgenome, while the later has not. The BBAA component of pentaploid TT was a F1 hybrid between that of TAA10 and TTR13. The F1 hybrid BBAA component of TT showed a similar strong capacity to buffer and sustains the unbalanced D chromosomes as shown by that of XE and TE. This suggests that the capacity to buffer and sustain imbalanced D-chromosomes by the BBAA components of hexaploid wheat is a dominant trait.

The above said, an alternative possibility concerns whether nature of the D genome chromosomes *per se* may also play a role in determining their retention versus elimination. Intuitively, like the BBAA genomes, the presumably co-evolved D chromosomes may also have adapted to become more compatible with the BBAA subgenome that those of *A. tauschii* that have never been coexisting in the same nucleus/cytoplasm with the BBAA genomes. As described above, the fact that there was no overt difference between XE that contained the D genome of *A. tauschii* (TQ18) from the two pentaploid lines (TE and TT) that harbored the co-evolved D genome with respect to the particular phenotypic manifestation (the capacity to buffer and sustain imbalanced D-chromosomes) clearly ruled out this possibility, and hence further reaffirms our scenario that the evolved dominant trait was encoded by the BBAA genomes.

Numerical Chromosome Variations Also Occurred in the A- and B-Subgenome Chromosomes in Progenies of the Pentaploid Wheat Lines

We also detected NCVs, i.e., aneuploidy, concerning the A and B chromosomes in small proportions of the progeny swarms of all four pentaploid wheat lines (**Supplementary Table S1**). The NCVs may involve gain or loss of one or more chromosomes of either the A or B subgenome, or simultaneous gain and loss of chromosomes involving both A and B subgenomes (**Supplementary Table S1**). Notably, sometimes, aneuploidies of the A and/or B chromosomes were accompanied with gain of extra D chromosomes (i.e., one D chromosomes being at three copies). Collectively, the A and B chromosome aneuploidies accounted for 5.1% (12 of 234), 2.0% (4 of 196), 7.2% (13 of 181), and 3.0% (5 of 169) of the karyotyped progeny plants derived from the XE, TE, TT, and ST pentaploid lines (**Supplementary Table S1**).

Structural Chromosome Variations Involving All Three Subgenomes Occurred in Progenies of the Pentaploid Wheat Lines

Variable types of structural chromosome variations (SCVs) were detected in certain proportions of the karyotyped progeny individuals from all four pentaploid lines. These were found to include telocentrics, translocations, isochromosomes, truncations, as well as complex structural variations containing more than one type of SCVs involving multiple chromosomes (**Figure 3**). Collectively, 16.2% (38 of 234), 13.3% (26 of 196), 19.9% (36 of 181), and 21.9% (37 of 169) of the karyotyped progeny plants from the XE, TE, TT, and ST pentaploid lines, respectively, were found to contain one or more SCVs (**Figure 4A**). Telocentrics were the most abundant type of SCVs, followed by translocations (**Figure 4A**); the two type of SCVs accounted for greater than 72.22% (26 of 36 in TT) and were significantly more abundant than the other SCV types in all four pentaploid wheat progeny populations (Prop.test, all P-values < 0.05).

Because more than one SCV may occur in a given individual, we further tabulated numbers of the various types of SCVs in the progeny swarms of each of the four pentaploid wheat lines. In each SCV types, we divided the number of SCVs belonging to intra or inter subgenome by the total number of SCVs to get the frequency of SCVs. We found that great majority of telocentrics involved the D chromosomes in progenies of all four pentaploid lines (**Figure 4B**), as expected given the propensity of unpaired univalents to undergo mis-division at anaphase of meiosis II. Actually, except for 4DL, telocentrics were found for both the long- and short-arm of all seven D chromosomes concerning all four pentaploid lines (**Supplementary Table S2**). Also as expected, isochromosomes were detected for many of the D chromosome arms, including 2DS, 2DL, 3DS, 3DL, 5DL, 6DS, and 7DS (e.g., **Figure 4B** and **Supplementary Table S2**). A less expected type of SCVs concerning the D chromosomes was intra-subgenome translocations, which were detected between chromosomes 1D and 2D, 1D and 3D, 1D and 4D, 1D and 7D, 2D and 3D, 2D and 4D, 2D and 6D, 2D and 7D, 4D and 7D, 5D and 6D, and 6D and 7D (e.g., **Figure 3A** and **Supplementary Table S2**).

Structural chromosome variation were not confined to the D chromosomes. For example, telocentrics of chromosomes 1AS, 4AS, and 5BS were also observed in progeny individuals of TE, XE, and ST, respectively. Moreover, inter-subgenome translocations involving A-D and/or B-D chromosomes were also observed in progeny plants of all the four pentaploid wheat lines. A striking feature of SCVs associated with progenies of ST relative to the other three pentaploid lines was that most of the translocations (six out of seven) were between chromosomes within the D subgenome (**Figure 4B** and **Supplementary Table S2**). By contrast, majority of the translocations in the other three pentaploid lines (XE, TE, and TT) was inter-subgenomic (**Figure 4B** and **Supplementary Table S2**).

Phenotypes of the Pentaploid Wheat Lines

We measured nine phenotypic traits for each of the four pentaploid wheat lines. We found significant differences for some, but not all, of the phenotyped traits among the lines (**Figure 5**). Also, the phenotypic differences were not consistent across the four lines, that is, a given line was not inferior to others in all traits (**Figure 5**). In particular, we did not find statistical differences among the four lines in seed-setting (**Figure 5**), a major reflection of reproductive fitness. Together, the phenotypic data suggest that the different capacities to buffer and sustain the unbalanced D chromosomes by the BBAA components of the pentaploid wheat lines are probably not related to the performance (fitness) of the pentaploids themselves.

DISCUSSION

We have shown in this study that all three manifestations, i.e., overall chromosome number distribution/exact chromosome composition (concerning the D subgenome chromosomes), NCVs (concerning the A and B subgenomes), and SCVs (concerning all three subgenomes), are strikingly different between the immediate progenies of a pentaploid wheat (ST) with the BBAA genome of a tetraploid *durum* wheat (cv. TTR13) and two pentaploid wheats (XE and TE) with the BBAA component of a hexaploid common wheat (cv. TAA10). Given the common feature of these three pentaploid wheat lines, i.e., all harboring *homozygous* BBAA subgenomes (**Figure 1A**), our results have unequivocally shown that the BBAA component of a hexaploid wheat possess a greater capacity than the BBAA genome of a *durum* wheat to buffer and sustain unbalanced D chromosomes, and hence, to constitute more kinds, more complex, greater severity of aneuploidies, as well as higher levels of SCVs in the progeny cohorts of the corresponding pentaploid wheats constituted by the former than that by the later. The observation that the immediate progenies of the pentaploid wheat (TT) with its BBAA subgenomes as a F1 hybrid

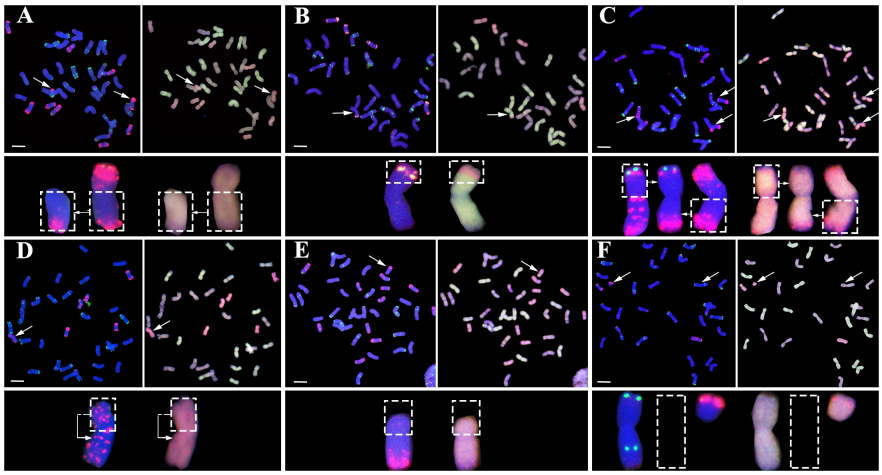

FIGURE 3 | Representative structural chromosome variations (SCVs) in the immediate progenies of the pentaploid wheat lines. **(A)** Telocentrics, **(B)** Inter-subgenome translocations, **(C)** Intra-subgenome translocations, **(D)** Isochromosomes, **(E)** Truncations, and **(F)** Telocentrics coupled with monosomic 2B (loss of one 2B chromosome). In **(A)** through **(F)**, the FISH images (left of each panel pair) are signals of pSc119.2 (green) and pAs1 (red), while the GISH images (right of each panel pair) are the signals of A subgenome (green), D subgenome (red), and B subgenome (blue), respectively. Structural variant chromosomes (marked by arrows) are zoomed in and shown below the images accordingly; while the dashed-lined empty frames in **(F)** refer to loss of one 2B chromosome (monosomic 2B). In all images, the bars = 10 μm.

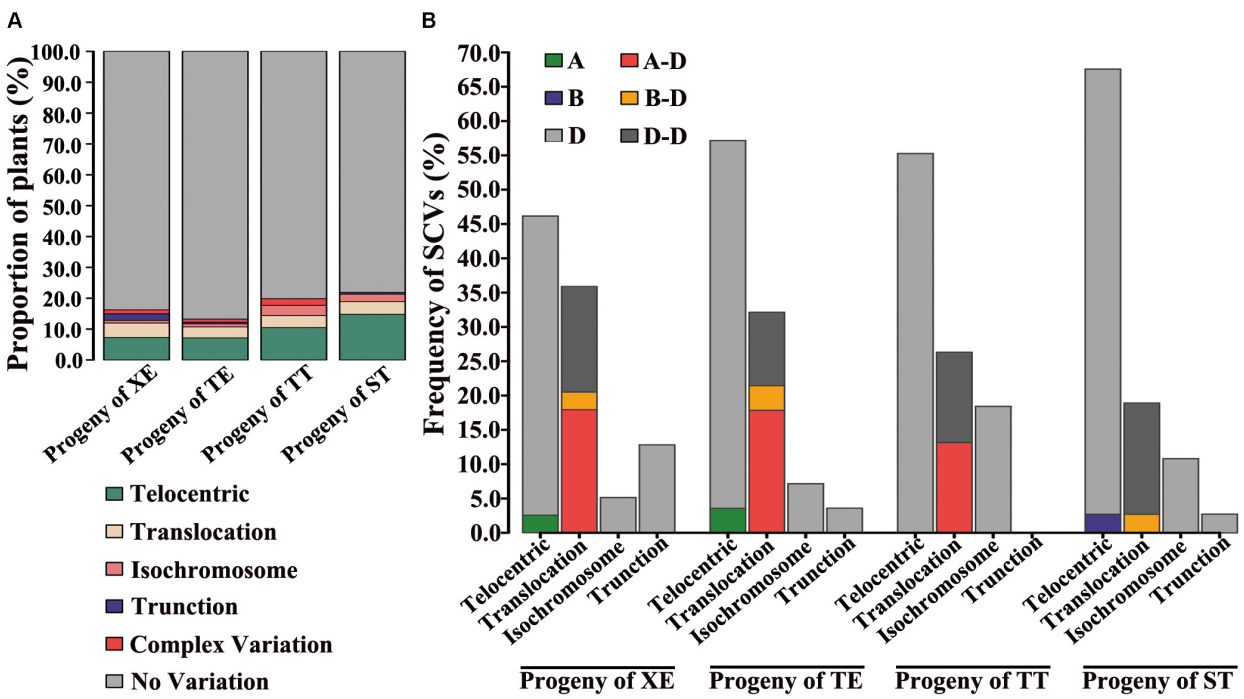

FIGURE 4 | SCVs in the immediate progeny swarms of each of the four pentaploid wheat lines. **(A)** Relative proportions of plants with the variable types of SCVs in the four immediate progeny swarms of the pentaploid wheat lines. **(B)** Relative proportions of chromosomes of each subgenome with SCVs in the four immediate progeny swarms of the pentaploid wheat lines.

between that of the hexaploid wheat (TAA10) and the tetraploid *durum* wheat (TTR13) does not differ from the two pentaploid wheats (XE and TE) with homozygous BBAA component of the hexaploid wheat indicates that the trait (capacity to buffer and sustain unbalanced D chromosomes) is genetically dominant. The gene(s) responsible for this trait has to be encoded by the A and/or B subgenomes, because no discernible difference was observed with respect to the three manifestations (above) between pentaploid wheats XE and TE which possess the same BBAA subgenomes (both are the BBAA component of hexaploid

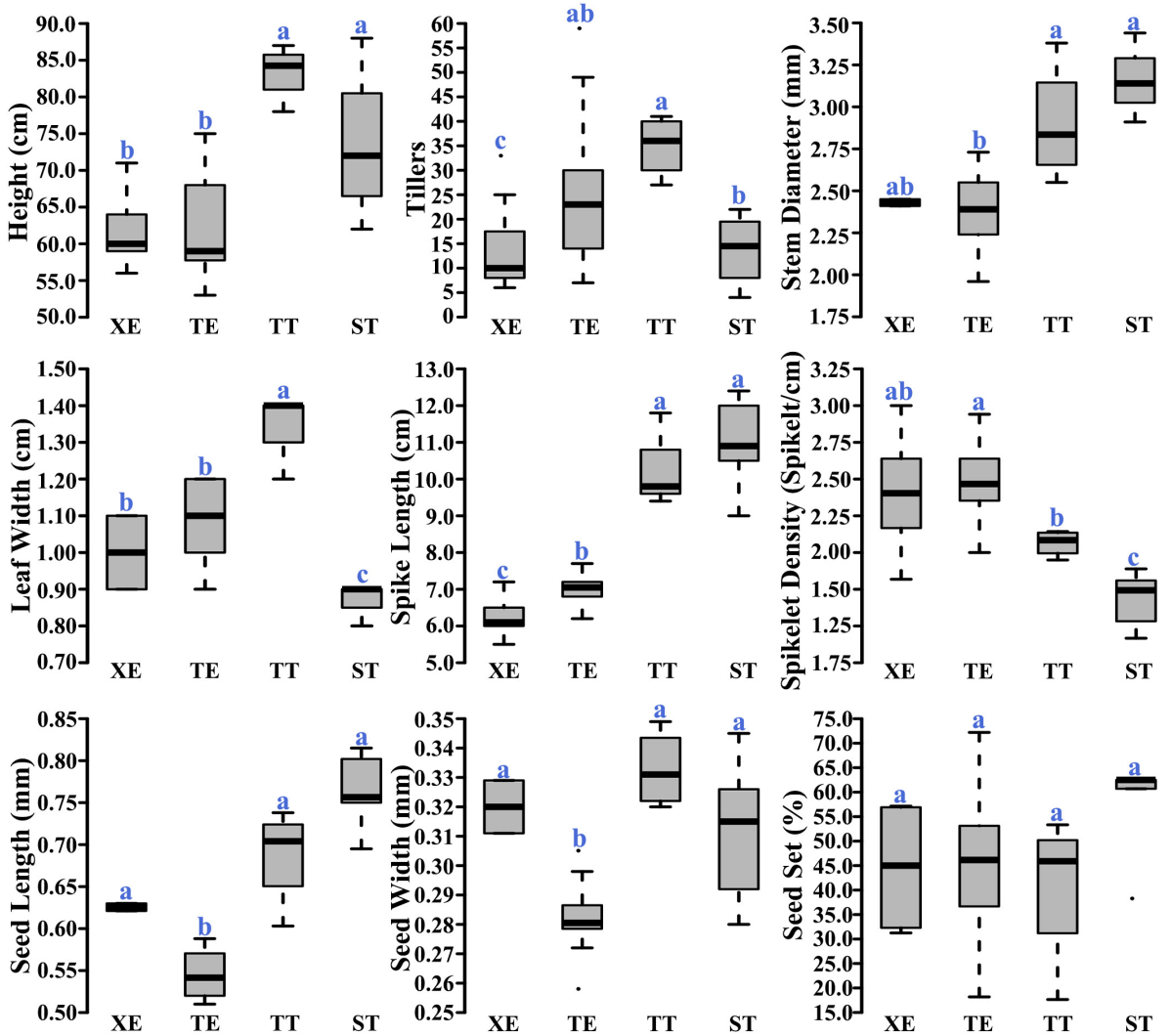

FIGURE 5 | Comparisons of the nine phenotypic traits among the four immediate progeny swarms of the pentaploid wheat lines. Letters above the boxplots denote statistical significance of the pairwise comparisons (Wilcox. test, *P*-value < 0.05).

wheat TAA10) but with different sources of the D genomes, one being the original of TAA10 while the other being of *A. tauschii*, accession TQ18 (**Figure 1A**). Our results are consistent with an earlier observation that in segregating progenies of given pentaploid wheat, those individuals with more BBAA alleles from the hexaploid parent were trended to contain more D chromosomes than those with more alleles from the *durum* wheat (Martin et al., 2011; Padmanaban et al., 2017a, 2018). However, because the pentaploid wheats constructed by these prior studies, being breeding oriented, are all with *heterozygous* BBAA genomes (hybrids between different genotypes of hexaploid wheat and *durum* wheat), other confounding factors, e.g., heterosis in the concerned trait, cannot be ruled out, and hence, an affirmative conclusion cannot be reached. Therefore, our empirical results have confirmed the previously only implicated possibility. Our results are also reminiscent of the earlier studies concerning chromosome compositions of selfed progenies derived from *Arabidopsis thaliana* intraspecific triploids (genome designations are CCC and CWW, respectively) constructed by crossing the same diploid line (Col-0, genome CC) with two different tetraploids, a natural ecotype (Warschau-1, genome WWWW), and an induced line (4x-Col, genome CCCC); it was found that progenies of the CWW triploid showed substantially more complex and extreme aneuploidies than those of CCC (Henry et al., 2005, 2007). However, there is a distinct difference between the *A. thaliana* triploids and the wheat pentaploids, because genomes in the former are from difference ecotypes (i.e., intraspecific) rather than different species, and therefore they are autotriploids.

What might be the mechanistic basis underlying this dominant trait evolved in the BBAA subgenomes of hexaploid common wheat? As proposed in the *A. thaliana* study (Henry et al., 2007), it has to do with buffering against dosage imbalance of genes and their products. Conceivably, this can

be accomplished by either or both of dosage compensation and dosage insensitivity. In case of wheat, it is intuitive, as well as experimentally documented, that the BBAA component of hexaploid common wheat possesses an improved (enhanced) version for either or both of these mechanisms. First, given its evolutionary history that the BBAA component has been coexisting and presumably co-evolving with the D subgenome for *ca.* 8,500 years (Kihara, 1944; McFadden and Sears, 1946; Feldman et al., 1995; Huang et al., 2002), naturally, the three subgenomes have been compatible with each other, and therefore incompatibility at each level (structural, epigenetic, or transcriptomic) has been resolved, for example, by rapid genetic and epigenetic changes and rewired gene expression (Kashkush et al., 2002, 2003; Zhao et al., 2011; Feldman and Levy, 2012; Zhang et al., 2016). Second, being at a higher ploidy level, a stronger compensatory capacity at least for WGD has conceivably been reinforced. Third, a strong compensation for dosage imbalance was found to exist in hexaploid common wheat, at least at the RNA transcript level, evidenced in various types of whole-chromosome aneuploidies (Zhang et al., 2017). Forth, it was found that hexaploid wheat formation (synthetics) at the initial stages is not different from neopolyploids of other studied plant species (e.g., Xiong et al., 2011; Chester et al., 2012) in that it is also associated with persistent whole chromosome aneuploidies (Zhang et al., 2013). However, it should be cautioned that properties of the original founder stand(s) of a primitively domesticated form of the tetraploid emmer wheat, *T. turgidum*, which served as the tetraploid maternal parent to give rise to the hexaploid wheat (*T. aestivum*) may differ from the current tetraploid wheats used to reconstruct the synthetic hexaploid wheats (Li et al., 2015). Nevertheless, assuming they have shared the same property, i.e., being associated with persistent aneuploidies, then it is conceivable that the BBAA component of hexaploid common wheat would have not only experienced WGD but also aneuploidy, therefore naturally, it has evolved the capacity to better buffer for not only the potential negative effects of balanced dosage change (due to WGD) but also imbalanced dosage (aneuploidy). Taken together, it is convincing that the BBAA component of hexaploid common wheat has evolved a stronger capacity to buffer and sustain imbalanced D chromosomes and SCVs than the BBAA genome of tetraploid wheat, as shown in this study.

What evolutionary advantages might have been bestowed to hexaploid common wheat by evolving a stronger capacity to buffer and sustain imbalanced D chromosomes and be more inclusive to numerical and SCVs in general? We consider this trait indeed matters. An intrinsic problem associated with polyploid speciation is genetic bottleneck, because very limited progenitor founder stands (if not only one) should have been involved in the initial polyploidization event under most natural settings. Thus, it is surprising that common wheat (*T. aestivum*) as a very young allohexaploid species was found to harbor a level of genetic diversity that is much greater than expected (Dubcovsky and Dvorak, 2007). Among others, one plausible scenario proposed is that both within the geographic region of its origin, known as the Fertile Crescent in the Near East (Salamini et al., 2002), and during its human-mediated global dispersal, hybridizations with wild and/or domesticated tetraploid emmer wheat (*T. turgidum*) might have been frequent when the latter was within pollinating adjacency (Dubcovsky and Dvorak, 2007). For these inter-ploidal hybridizations to occur and being evolutionarily consequential, fitness, fecundity as well as karyotype heterogeneity (chromosome composition) of the successive progenies descended from the pentaploid intermediates would have been critical to enable an eventual return to euploid hexaploidy. Expectedly, a stronger capacity to buffer and sustain unbalanced D genome chromosomes by the BBAA subgenomes would be in favor of these properties and facilitate the persistence of pentaploid-derived lineages for long enough to ensure successful reversion to euploid hexaploid wheat with incorporated genetic variations from the tetraploid wheats. Analogically, a stronger inclusiveness to whole chromosome aneuploidies for protracted period may generate additional genomic variations as documented in yeast (Sheltzer et al., 2011). Together, repeated hybridizations with diverse accessions of its tetraploid emmer wheat progenitor would apparently contribute to the higher than expected level of genetic diversities seen in hexaploid common wheat. These introgressed or *de novo* generated genetic diversities (due to protracted states of aneuploidy) might have served as raw materials for evolving the remarkable adaptability of hexaploid common wheat to a wide spectrum of climatic and environmental conditions around the world.

By using the pentaploids as intermediates, several agriculturally important traits have been reciprocally introgressed between tetraploid *durum* wheat and hexaploid common wheat (Xu et al., 2013; Kalous et al., 2015; Han et al., 2016). However, it is recognized that the efficacy of these endeavors is cross (genotype) dependent. Further understanding of the genetic basis underlying the trait of capacitating imbalanced D chromosomes by the BBAA component will undoubtedly enable more judicious designing of the crosses and increase the breeding efficiency.

AUTHOR CONTRIBUTIONS

BL and XD conceived and designed the project. XD and YS performed the experiments. FW and BL contributed reagents, materials, and analysis tools. XD, FW, and BL wrote the paper. ZL, YW, and AZ contributed with essential suggestions and data analyses in the course of the project. All authors contributed to manuscript revision, proofreading, and approval of the final manuscript.

FUNDING

This work was supported by the National Key Research and Development Program of China (2016YFD0102003), the National Natural Science Foundation of China (31290210), and the Program for Introducing Talents to Universities (B07017).

ACKNOWLEDGMENTS

We thank the section editor, Dr. Yiwei Jiang of Purdue University, for the kind invitation, and helpful comments by our colleagues at Northeast Normal University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01149/full#supplementary-material

REFERENCES

Birchler, J. A., Bhadra, U., Bhadra, M. P., and Auger, D. L. (2001). Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev. Biol.* 234, 275–288. doi: 10.1006/dbio.2001.0262

Chester, M., Gallagher, J. P., Symonds, V. V., da Silva, A. V. C., Mavrodiev, E. V., Leitch, A. R., et al. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). *Proc. Natl. Acad. Sci. U.S.A.* 109, 1176–1181. doi: 10.1073/pnas.1112041109

Dubcovsky, J., and Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316, 1862–1866. doi: 10.1126/science.1143986

Eberhard, F. S., Zhang, P., Lehmensiek, A., Hare, R. A., Simpfendorfer, S., and Sutherland, M. W. (2010). Chromosome composition of an F2 *Triticum aestivum*× *T. turgidum* spp. durum cross analysed by DArT markers and MCFISH. *Crop Pasture Sci.* 61, 619–624. doi: 10.1071/CP10131

Feldman, M., and Levy, A. A. (2012). Genome evolution due to allopolyploidization in wheat. *Genetics* 192, 763–774. doi: 10.1534/genetics.112.146316

Feldman, M., Lupton, F. G. H., and Miller, T. E. (1995). "Evolution of crop plants," in *Wheats*, eds J. Smartt and N. W. Simmonds (London: Longman Scientific & Technical), 184–192.

Gao, L., Diarso, M., Zhang, A., Zhang, H., Dong, Y., Liu, L., et al. (2016). Heritable alteration of DNA methylation induced by whole-chromosome aneuploidy in wheat. *New Phytol.* 209, 364–375. doi: 10.1111/nph.13595

Gasch, A. P., Hose, J., Newton, M. A., Sardi, M., Yong, M., and Wang, Z. (2016). Further support for aneuploidy tolerance in wild yeast and effects of dosage compensation on gene copy-number evolution. *eLife* 5:e14409. doi: 10.7554/eLife.14409

Han, C., Zhang, P., Ryan, P. R., Rathjen, T. M., Yan, Z., and Delhaize, E. (2016). Introgression of genes from bread wheat enhances the aluminium tolerance of durum wheat. *Theor. Appl. Genet.* 129, 729–739. doi: 10.1007/s00122-015-2661-3

Henry, I. M., Dilkes, B. P., and Comai, L. (2007). Genetic basis for dosage sensitivity in *Arabidopsis thaliana*. *PLoS Genet.* 3:e70. doi: 10.1371/journal.pgen.0030070

Henry, I. M., Dilkes, B. P., Miller, E. S., Burkart-Waco, D., and Comai, L. (2010). Phenotypic consequences of aneuploidy in *Arabidopsis thaliana*. *Genetics* 186, 1231–1245. doi: 10.1534/genetics.110.121079

Henry, I. M., Dilkes, B. P., Young, K., Watson, B., Wu, H., and Comai, L. (2005). Aneuploidy and genetic variation in the *Arabidopsis thaliana* triploid response. *Genetics* 170, 1979–1988. doi: 10.1534/genetics.104.037788

Hintze, J. L., and Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *Am. Stat.* 52, 181–184.

Hose, J., Yong, C. M., Sardi, M., Wang, Z., Newton, M. A., and Gasch, A. P. (2015). Dosage compensation can buffer copy-number variation in wild yeast. *eLife* 4:e05462. doi: 10.7554/eLife.05462.001

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., et al. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* 99, 8133–8138. doi: 10.1073/pnas.072223799

Kalous, J. R., Martin, J. M., Sherman, J. D., Heo, H.-Y., Blake, N. K., Lanning, S. P., et al. (2015). Impact of the D genome and quantitative trait loci on quantitative traits in a spring durum by spring bread wheat cross. *Theor. Appl. Genet.* 128, 1799–1811. doi: 10.1007/s00122-015-2548-3

Kashkush, K., Feldman, M., and Levy, A. A. (2002). Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics* 160, 1651–1659.

Kashkush, K., Feldman, M., and Levy, A. A. (2003). Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* 33, 102–106. doi: 10.1038/ng1063

Kaya, A., Gerashchenko, M. V., Seim, I., Labarre, J., Toledano, M. B., and Gladyshev, V. N. (2015). Adaptive aneuploidy protects against thiol peroxidase deficiency by increasing respiration via key mitochondrial proteins. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10685–10690. doi: 10.1073/pnas.1505315112

Kerber, E. R. (1964). Wheat: reconstitution of the tetraploid component (AABB) of hexaploids. *Science* 143, 253–255. doi: 10.1126/science.143.3603.253

Kihara, H. (1925). Weitere über die pentaploiden *Triticum* - Bastarde. I. *Jpn. J. Bot.* 2, 299–305.

Kihara, H. (1944). Discovery of the DD-analyser, one of the ancestors of *Triticum vulgare*. *Agric. Hortic.* 19, 13–14. doi: 10.1016/j.jgg.2011.07.002

Letourneau, A., Santoni, F. A., Bonilla, X., Sailani, M. R., Gonzalez, D., Kind, J., et al. (2014). Domains of genome-wide gene expression dysregulation in down's syndrome. *Nature* 508, 345–350. doi: 10.1038/nature13200

Li, A. L., Geng, S. F., Zhang, L. Q., Liu, D. C., and Mao, L. (2015). Making the bread: insights from newly synthesized allohexaploid wheat. *Mol. Plant* 8, 847–859. doi: 10.1016/j.molp.2015.02.016

Liu, C., Yang, X., Zhang, H., Wang, X., Zhang, Z., Bian, Y., et al. (2015). Genetic and epigenetic modifications to the BBAA component of common wheat during its evolutionary history at the hexaploid level. *Plant Mol. Biol.* 88, 53–64. doi: 10.1007/s11103-015-0307-0

Martin, A., Simpfendorfer, S., Hare, R. A., Eberhard, F. S., and Sutherland, M. W. (2011). Retention of D genome chromosomes in pentaploid wheat crosses. *Heredity* 107, 315–319. doi: 10.1038/hdy.2011.17

Matzke, M. A., Florian Mette, M., Kanno, T., and Matzke, A. J. M. (2003). Does the intrinsic instability of aneuploid genomes have a causal role in cancer? *Trends Genet.* 19, 253–256.

McFadden, E. S., and Sears, E. R. (1946). The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J. Hered.* 37, 81–89. doi: 10.1093/oxfordjournals.jhered.a105590

Megarbane, A., Ravel, A., Mircher, C., Sturtz, F., Grattau, Y., Rethore, M.-O., et al. (2009). The 50th anniversary of the discovery of trisomy 21: the past, present, and future of research and treatment of Down syndrome. *Genet. Med.* 11, 611–616. doi: 10.1097/GIM.0b013e3181b2e34c

Millet, C., Ausiannikava, D., Le Bihan, T., Granneman, S., and Makovets, S. (2015). Cell populations can use aneuploidy to survive telomerase insufficiency. *Nat. Commun.* 6:8664. doi: 10.1038/ncomms9664

Padmanaban, S., Sutherland, M. W., Knight, N. L., and Martin, A. (2017a). Genome inheritance in populations derived from hexaploid/tetraploid and tetraploid/hexaploid wheat crosses. *Mol. Breed.* 37:48. doi: 10.1007/s11032-017-0647-3

Padmanaban, S., Zhang, P., Hare, R. A., Sutherland, M. W., and Martin, A. (2017b). Pentaploid wheat hybrids: applications, characterisation, and challenges. *Front. Plant Sci.* 8:358. doi: 10.3389/fpls.2017.00358

Padmanaban, S., Zhang, P., Sutherland, M. W., Knight, N. L., and Martin, A. (2018). A cytological and molecular analysis of D-genome chromosome retention following F2–F6 generations of hexaploid × tetraploid wheat crosses. *Crop Pasture Sci.* 69, 121–130. doi: 10.1071/CP17240

Pavelka, N., Rancati, G., Zhu, J., Bradford, W. D., Saraf, A., Florens, L., et al. (2010). Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature* 468, 321–325. doi: 10.1038/nature09529

Ramsey, J., and Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.* 29, 467–501. doi: 10.1104/pp.16.01768

Rancati, G., Pavelka, N., Fleharty, B., Noll, A., Allen, R., Walton, K., et al. (2008). Aneuploidy and polyploidy underlie adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. *Cell* 135, 879–893. doi: 10.1016/j.cell.2008.09.039

Rayburn, A. L., and Gill, B. S. (1986). Isolation of a genome-specific repeated DNA sequence from *Aegilops squarrosa*. *Plant Mol. Biol. Rep.* 4, 102–109. doi: 10.1007/BF02732107

Rutledge, S. D., and Cimini, D. (2016). Consequences of aneuploidy in sickness and in health. *Curr. Opin. Cell Biol.* 40, 41–46. doi: 10.1016/j.ceb.2016.02.003

Salamini, F., Ozkan, H., Brandolini, A., Schafer-Pregl, R., and Martin, W. (2002). Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.* 3, 429–441. doi: 10.1038/nrg817

Sansregret, L., and Swanton, C. (2017). The role of aneuploidy in cancer evolution. *Cold Spring Harb. Perspect. Med.* 7:a028373. doi: 10.1101/cshperspect.a028373

Sax, K. (1922). Sterility in wheat hybrids. III. Endosperm development and F2 sterility. *Genetics* 7, 553–558.

Sears, E. R. (1944). Cytogenetic studies with polyploid species of wheat. II. Additional chromosomal aberrations in *Triticum vulgare*. *Genetics* 29, 232–246.

Sheltzer, J. M., Blank, H. M., Pfau, S. J., Tange, Y., George, B. M., Humpton, T. J., et al. (2011). Aneuploidy drives genomic instability in yeast. *Science* 333, 1026–1030. doi: 10.1126/science.1206412

Soltis, P. S., Marchant, D. B., Van de Peer, Y., and Soltis, D. E. (2015). Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.* 35, 119–125. doi: 10.1016/j.gde.2015.11.003

Torres, E. M., Springer, M., and Amon, A. (2016). No current evidence for widespread dosage compensation in *S. cerevisiae*. *eLife* 5:e10996. doi: 10.7554/eLife.10996

Torres, E. M., Williams, B. R., and Amon, A. (2008). Aneuploidy: cells losing their balance. *Genetics* 179, 737–746. doi: 10.1534/genetics.108.090878

Wang, H. Y., Liu, D. C., Yan, Z. H., Wei, Y. M., and Zheng, Y. L. (2005). Cytological characteristics of F2 hybrids between *Triticum aestivum* L. and *T. durum* Desf. with reference to wheat breeding. *J. Appl. Genet.* 46, 365–369.

Wu, Y., Sun, Y., Sun, S., Li, G., Wang, J., Wang, B., et al. (2018). Aneuploidization under segmental allotetraploidy in rice and its phenotypic manifestation. *Theor. Appl. Genet.* 131, 1273–1285. doi: 10.1007/s00122-018-3077-7

Xiong, Z., Gaeta, R. T., and Pires, J. C. (2011). Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proc. Natl. Acad. Sci. U.S.A.* 108, 7908–7913. doi: 10.1073/pnas.1014138108

Xu, L. S., Wang, M. N., Cheng, P., Kang, Z. S., Hulbert, S. H., and Chen, X. M. (2013). Molecular mapping of Yr53, a new gene for stripe rust resistance in durum wheat accession PI 480148 and its transfer to common wheat. *Theor. Appl. Genet.* 126, 523–533. doi: 10.1007/s00122-012-1998-0

Zhang, A., Li, N., Gong, L., Gou, X., Wang, B., Deng, X., et al. (2017). Global analysis of gene expression in response to whole-chromosome aneuploidy in hexaploid wheat. *Plant Physiol.* 175, 828–847. doi: 10.1104/pp.17.00819

Zhang, H., Bian, Y., Gou, X., Zhu, B., Xu, C., Qi, B., et al. (2013). Persistent whole-chromosome aneuploidy is generally associated with nascent allohexaploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* 110, 3447–3452. doi: 10.1073/pnas.1300153110

Zhang, H., Gou, X., Zhang, A., Wang, X., Zhao, N., Dong, Y., et al. (2016). Transcriptome shock invokes disruption of parental expression-conserved genes in tetraploid wheat. *Sci. Rep.* 6:26363. doi: 10.1038/srep26363

Zhang, H., Zhu, B., Qi, B., Gou, X., Dong, Y., Xu, C., et al. (2014). Evolution of the BBAA component of bread wheat during its history at the allohexaploid level. *Plant Cell* 26, 2761–2776. doi: 10.1105/tpc.114.128439

Zhao, N., Xu, L., Zhu, B., Li, M., Zhang, H., Qi, B., et al. (2011). Chromosomal and genome-wide molecular changes associated with initial stages of allohexaploidization in wheat can be transit and incidental. *Genome* 54, 692–699. doi: 10.1139/g11-028

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: 03 August 2018
doi: 10.3389/fpls.2018.01045

Genome-Wide Association Studies to Identify Loci and Candidate Genes Controlling Kernel Weight and Length in a Historical United States Wheat Population

*Sintayehu D. Daba[1], Priyanka Tyagi[2], Gina Brown-Guedira[2,3] and Mohsen Mohammadi[1]**

[1] Department of Agronomy, Purdue University, West Lafayette, IN, United States, [2] Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC, United States, [3] Small Grains Genotyping Laboratory, United States Department of Agriculture, Agricultural Research Services, Raleigh, NC, United States

OPEN ACCESS

Edited by:
Shuizhang Fei,
Iowa State University, United States

Reviewed by:
Lifeng Zhu,
Nanjing Normal University, China
Jacqueline Campbell,
Iowa State University, United States
Fei Lu,
Institute of Genetics and Developmental Biology (CAS), China

***Correspondence:**
Mohsen Mohammadi
mohamm20@purdue.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *08 February 2018*
Accepted: *27 June 2018*
Published: *03 August 2018*

Citation:
Daba SD, Tyagi P, Brown-Guedira G and Mohammadi M (2018) Genome-Wide Association Studies to Identify Loci and Candidate Genes Controlling Kernel Weight and Length in a Historical United States Wheat Population. Front. Plant Sci. 9:1045. doi: 10.3389/fpls.2018.01045

Although kernel weight (KW) is a major component of grain yield, its contribution to yield genetic gain during breeding history has been minimal. This highlights an untapped potential for further increases in yield via improving KW. We investigated variation and genetics of KW and kernel length (KL) via genome-wide association studies (GWAS) using a historical and contemporary soft red winter wheat population representing 200 years of selection and breeding history in the United States. The observed changes of KW and KL over time did not show any conclusive trend. The population showed a structure, which was mainly explained by the time and location of germplasm development. Cluster sharing by germplasm from more than one breeding population was suggestive of episodes of germplasm exchange. Using 2 years of field-based phenotyping, we detected 26 quantitative trait loci (QTL) for KW and 27 QTL for KL with $-log_{10}(p) > 3.5$. The search for candidate genes near the QTL on the wheat genome version IWGSCv1.0 has resulted in over 500 genes. The predicted functions of several of these genes are related to kernel development, photosynthesis, sucrose and starch synthesis, and assimilate remobilization and transport. We also evaluated the effect of allelic polymorphism of genes previously reported for KW and KL by using Kompetitive Allele Specific PCR (KASP) markers. Only *TaGW2* showed significant association with KW. Two genes, i.e., *TaSus2-2B* and *TaGS-D1* showed significant association with KL. Further physiological studies are needed to decipher the involvement of these genes in KW and KL development.

Keywords: kernel weight, kernel length, QTL, GWAS, candidate gene, historical germplasm

INTRODUCTION

Yield genetic gains in wheat slowed down over the last two decades (Brisson et al., 2010; Lin and Huybers, 2012; Ray et al., 2012), threatening world food security. Simmonds et al. (2014) highlighted that grain number (GN) per unit area and kernel weight (KW) are main determinants of grain yield (GY). These two traits, i.e., GN and KW together represent total sink-strength in

wheat. Over the course of the breeding history of cereals, the per unit area GN has considerably increased, while KW showed no significant increase or even decreased slightly (Brancourt-Hulmel et al., 2003; Carver, 2009). KW is determined by kernel size, which is a function of kernel width, length, and thickness, and degree of grain filling (Simmonds et al., 2014; Su et al., 2016). Though complex, KW is the most heritable trait among yield components (Su et al., 2016), with heritability reaching as high as 0.87 (Bergman et al., 2000; Wiersma et al., 2001). Kernel development in wheat involves cell division, water uptake, accumulation of starch and protein, maturation, and desiccation (Altenbach and Kothari, 2004). While grain expansion enforced by endosperm cell division and water uptake are components of sink-strength, assimilate (e.g., starch) supply (Emes et al., 2003) through current photosynthesis or remobilization of reserves from vegetative tissues (Bidinger et al., 1977; Schnyder, 1993; Gebbing and Schnyder, 1999) are components of source-strength.

Several QTL for KW and kernel dimension traits have been localized across the 21 wheat chromosomes (Zhang et al., 2012; Jaiswal et al., 2015; Chen et al., 2016). Only a few loci were functionally validated in wheat, compared to other cereals such as rice for KW and dimension traits, due to the lack of reference genome sequence and ploidy complexity (allohexaploid, $2n = 6X = 42$) of the wheat (Simmonds et al., 2014; Su et al., 2016). To this end, several genes identified in other cereals were postulated to be involved in kernel trait determination in wheat. *Sucrose transporter TaSUT* was shown to regulate the translocation of assimilates from source to sink tissues (Aoki et al., 2004; Deol et al., 2013). *Sucrose synthase TaSus* catalyzes the first step in the conversion of sucrose to starch, particularly the conversion of sucrose to fructose and *UDP-glucose* (Jiang et al., 2011; Hou et al., 2014). Cytokinin oxidase *TaCKX* which inactivates *cytokinin* reversibly was shown to have an effect on KW (Zhang et al., 2010; Lu et al., 2015; Chang et al., 2016). *Cytokinin* oxidase such as TaCKX1 highly expressed during early seed development (Song et al., 2012). *Cell wall invertase TaCWI* exerts a role in kernel size control by sink tissue development and carbon partitioning (Ma et al., 2012). Several other grain size related genes include *TaGS-D1* which codes for *Glutamine synthase* with effect on KW and kernel length (KL) (Zhang et al., 2014); *TaGW6,* which encodes for *indole-3-acetic acid (IAA)-glucose hydrolase* (Hu et al., 2016); and *TaGW2* (Pflieger et al., 2001; Su et al., 2011; Bednarek et al., 2012) encodes for a *RING-type protein with E3 ubiquitin ligase* activity that controls KW and interestingly, positively regulates grain size as opposed to the rice *GW2* gene which has negative effect on grain size (Bednarek et al., 2012). Deployment and transferability of these genes in populations and environments beyond the discovery populations and environments is a valuable applied research question.

Genome-wide association studies (GWAS) that dissect the genetic basis of traits and propose candidate genes (Pflieger et al., 2001), could be an important step for trait improvement. The scope of genes and alleles that are identified in GWAS pipelines depends, to a large extent, on the variation that is in the germplasm. In most cases, discovery panels consist of elite lines from multiple breeding programs (Mohammadi et al., 2015), which usually demonstrate high familial relatedness; or GenBank accessions (Zhao et al., 2011), which are often genetically structured by the geography of origin. The third type of diversity panel could be accessions sampled from adapted breeding materials in a spectrum of time, i.e., from the past to present time, which can identify alleles that became extinct or are recently introduced. Analyses of genetic gain in wheat have not postulated significant improvement in KW parallel to what observed in GN. The quest for increases in KW parallel to GN will depend on the genetic nature of KW that may be gained from a past-to-present perspective of an allelic composition of wheat accessions. Crossing schemes and selections from among segregating progeny, which is a landmark of the breeding process, can be thought as accelerated evolutionary forces that either rapidly fix or purge alleles. Therefore, current elite germplasm is likely unable to depict alleles that contributed in the past or are now fixed. Mapping using in-time diversity panels allows understanding of the realized trends and gain or loss of beneficial alleles, both very important factors for strategizing breeding programs.

Development of molecular markers for KW will greatly facilitate the selection process. In this study, we utilize a unique wheat population composed of historical and contemporary germplasm, representing breeding history and selection of over 200 years in the United States wheat industry. The panel has a considerable variation for several traits including KW, allowing a high power of QTL detection. The objectives of this study include, (1) to identify quantitative trait loci (QTL) for KW and KL in a historical and contemporary set of soft red winter wheat (SRWW) in the United States, (2) to search the recently published wheat reference genome IWGSC RefSeq v1.0 annotation v1.0 to mine candidate genes that are putatively responsible for determination of KW and KL in wheat.

MATERIALS AND METHODS

Plant Materials and Field Trials

Historical and contemporary SRWW cultivars and breeding lines, representing 200 years (1814–2015) of selection and breeding history in diverse geographical regions in the United States were phenotyped. The seed for most of the entries was provided by the National Small Grains Collection (NSGC), United States Department of Agriculture (USDA) in Aberdeen, Idaho. Accessions were field grown to maturity at the Agronomy Center for Research and Education (ACRE), Purdue University, West Lafayette in the cropping seasons of 2015–2016 and 2016–2017. We grow each entry in a 1-m long single row plot with 25 cm row spacing. The crop received 106 kg N ha^{-1} in both years just after the winter dormancy break. As old accessions with no height reducing (*Rht*) genes were at the risk of lodging and disruption of grain filling process, we assembled guards and ropes around row plots to prevent lodging.

Phenotyping

Each single-row was hand harvested and processed at ACRE. Two kernel characteristics were measured, i.e.; KW and KL. We

hand-harvested multiple heads from each entry, oven-dried, and measured the weight of two replicates of 100 kernels. The average KW was then expressed in milligram (mg). The experiments in 2015–2016 and 2016–2017 seasons did not include the same number of entries. In the 2015–2016 season, 265 entries were phenotyped. In the 2016–2017 season, 214 entries were phenotyped. Only 160 entries were in common between 2015–2016 and 2016–2017. Altogether, in both years KW from 325 entries were measured. The KW data of 2015–2016 and 2016–2017 are referred to as KW16 and KW17. The Best Unbiased Linear Predictor (BLUP) of KW across both years with 325 entries is referred to as KW1617. For KL, 265 entries were measured in 2015–2016 and 217 entries were measured in 2016–2017. The common entries between both years were 160. Altogether, in both years KL from 323 entries were measured. For measuring KL, we aligned 10 kernels to the side of a ruler. The resulting measurements were divided by 10 and expressed in millimeter (mm) for a single kernel. Similar to KW, KL data are referred to as KL16, KL17, and KL1617 for 2016, 2017, and combined BLUP estimates, respectively.

Analysis of Traits and Trends

The relationship between the datasets generated in different environments was used as a measure of repeatability of the phenotypic measurements. Correlations among the different datasets can be indicative of technical heritability and repeatability of KW and KL in diverse environments. We also estimated the broad-sense heritability values for both traits using the variance components. The trend of traits over time was visualized by using boxplots of KW1617 and KL1617 datasets of the four year-groups ($YG \leq 1920$, $1920 < YG \leq 1960$, $1960 < YG < 2000$, and $YG \geq 2000$). The total number of entries in each YG and the number of entries phenotyped for KW and KL in the 2 years are shown in **Table 1**.

TABLE 1 | Distribution of the lines in each of the four year-groups and years.

Year-group	Total over 2016 and 2017	Phenotyped	
		2016	2017
Before 1920	35	33	15
1920–1960	64	57	28
1960–2000	168	152	121
After 2000	57	23	50

Genotyping

For genotyping, we extracted DNA from 15-day-old leaf samples and followed a sequencing-based genotyping procedure explained by Poland et al. (2012). The genomic libraries were created using *Pst1-Msp1* restriction enzyme combinations. The samples were pooled together in 96-plex to create libraries and each library was sequenced on a single lane of Illumina Hi-Seq 2500. SNP calling was performed using the TASSEL5 GBSv2 pipeline[1] using 64 base kmer length and minimum kmer count of 5. Reads were aligned to the wheat reference "IWGSC_WGAv1.0"[2] using aln method of Burrows–Wheeler Aligner (BWA) version 0.7.10 (Li and Durbin, 2009). We used the default parameters of BWA. This resulted in 309,711 unfiltered SNP loci. The SNPs not assigned to any chromosome were removed. The remaining markers were filtered for minor allele frequency (MAF) $\geq$5% and missing values $\leq$30%, which resulted in 60,132 SNP. Missing data were imputed using the Linkage Disequilibrium K-number neighbor imputation (LDKNNi) (Money et al., 2015) algorithm implemented in Tassel 5.0 (Bradbury et al., 2007). We also estimated the error rates of LDKNNi imputation for the different level of masking and the results are given in **Supplementary Table S1**.

[1]https://bitbucket.org/tasseladmin/tassel-5-source/wiki/Tassel5GBSv2Pipeline

[2]https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies

Population Structure

Population structure was evaluated using principal component analysis (PCA) of 60,132 SNP markers, implemented in TASSEL5.0 (Bradbury et al., 2007). Population structure was then visualized using a three-dimensional plot of the first three principal components (PCs) using the R package "Scatterplot3d" (Ligges and Maechler, 2003). We also conducted model-based Bayesian clustering analysis using Structure 2.3.4 (Pritchard et al., 2000). Total of 16,313 tag SNPs were used for this analysis, which were selected using tagger function in Haploview (Barrett et al., 2005). The parameters in the tagger function set to "pairwise tagging only" with $R^2 = 0.8$. To infer population structure for 325 wheat genotypes, we ran structure analysis for K-values from 2 to 10. Both the length of burn-in period and the number of iterations were set at 10,000. The K-value reached a plateau when the minimal number of groups that best described the population sub-structure has been attained (Pritchard et al., 2000). The average K-values were plotted against their respective logarithm of the probability of likelihood, i.e., LnP(D). The rate of change in the log probability of data between successive K-values (Evanno et al., 2005) was used to predict the most appropriate number of subpopulations. We described the differentiation among the four clusters using fixation index (F_{ST}) method (Wright, 1951, 1978).

Genome-Wide Association Studies

Association mapping was performed for the two kernel traits using the 60,132 SNP markers in GAPIT package (Lipka et al., 2012). We used mixed linear model (MLM), applying P3D (Population Parameters Previously Determined) described as Mixed-Model Association on eXpedited (EMMAX) algorithm (Kang et al., 2010). Our model included markers and the first three PCs of the population structure as fixed effects. Kinship as familial relatedness matrix and residual terms were considered as random effects. Manhattan plots were produced using the negative logarithm at base 10 of the p-values, shortened as $-log10(p)$ using "qqman" package of R (Turner, 2014) across the physical map. The markers with $-log10(p) > 3.5$ were identified for further characterization. We constructed LD block for significant SNP markers within a chromosome using HAPLOVIEW v4.2 (Barrett et al., 2005) to assign markers to short blocks. Changes in favorable alleles over time was evaluated using the same four year-groups that were used for trend analysis.

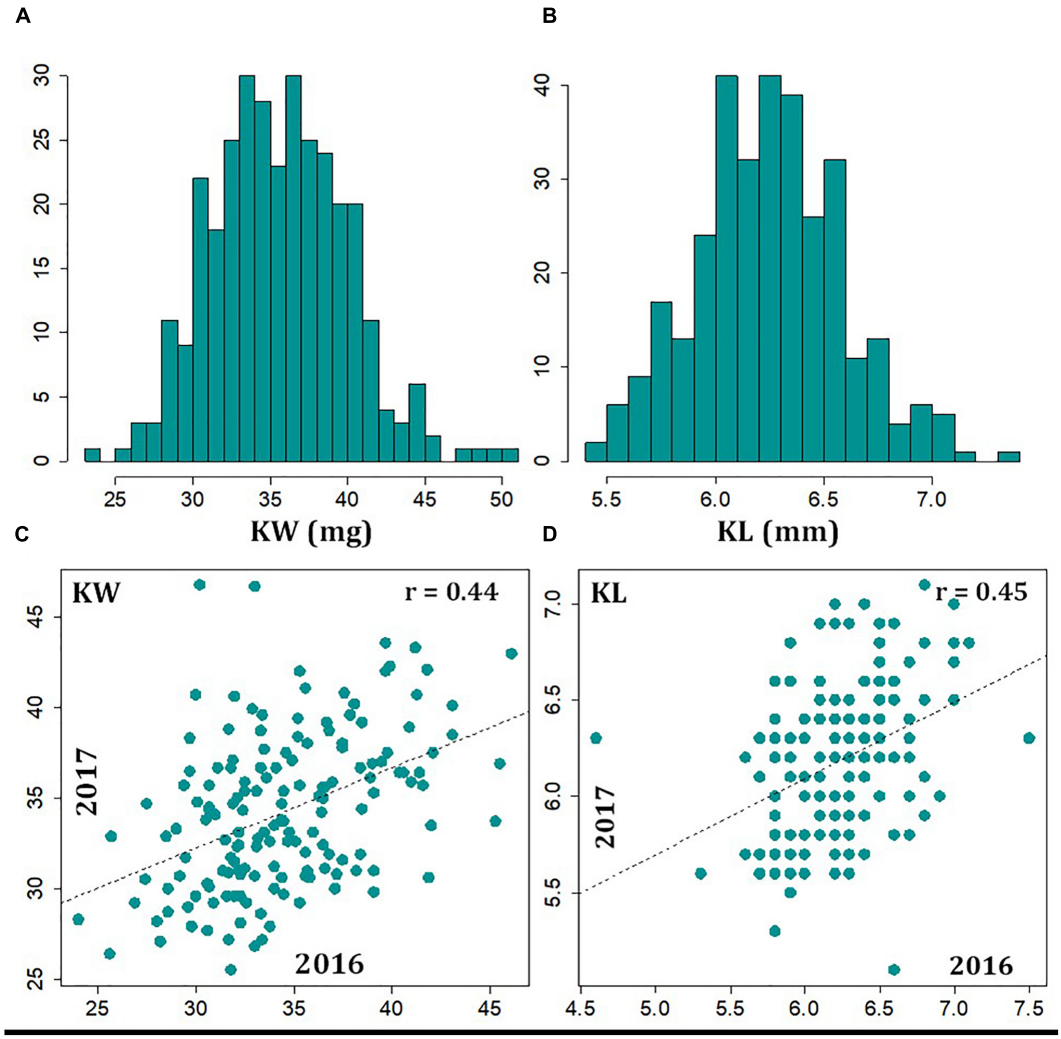

FIGURE 1 | Distribution of 2-year BLUP values kernel weight **(A)** and kernel length **(B)**, and correlations, as evidence of technical repeatability, observed between 2016 and 2017 data for kernel weight **(C)** and kernel length **(D)**.

The cumulative effect of identified favorable alleles on the kernel traits was also evaluated.

Effect of Known Loci/Genes on Kernel Traits

Allelic composition of previously reported loci/genes implicated in kernel traits, i.e., *TaSus2-2B*, *TaCWI-4A*, *TaCWI-5D*, *TGW6*, *TaTGW6-A1*, *TaGS-D1*, *TaGW2*, *Rht-1B*, and *Rht-1D* were evaluated using KASP markers described in Rasheed et al. (2016). These polymorphisms were used in a Student's *t*-test to statistically assess the effect of each known locus/gene on the variation of kernel traits.

Candidate Gene Identification

We retrieved high confidence wheat genes surrounding (within ±250 kb) representative SNPs for the genomic regions identified both for KW and KL. For gene search purpose, we used IWGSC RefSeq v1.0 annotation v1.0, iwgsc_refseqv1.0_HighConf_2017Mar13.gff3.zip[3].

RESULTS

Phenotypic Variation

We evaluated the variation of KW and KL in a historical and contemporary collection of cultivars and experimental breeding lines, representing 200 years of breeding and selection history. Across the 2 years of study, the Best Linear Unbiased Estimate (BLUE) values (i.e., KW1617) showed a mean of 35.6 mg with a range from 23.5 to 50.6 mg (**Figure 1A**). The 20 greatest KW entries showed an average of 44.8 ± 2.5 mg and the 20 smallest KW entries showed an average of 27.7 ± 1.4 mg. The mean

[3]https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/

phenotype value for KW16 and KW17 were 35.2 mg (with a range of 23.3–50.7 mg) and 35.2 mg (with a range of 25.5 – 49.8 mg), respectively (**Supplementary Figure S1**). The mean of KL1617 BLUE values was 6.3 mm, with a range of 5.3–7.4 mm (**Figure 1B**). The 20 longest kernel entries showed an average of 7.0 ± 0.15 mm and the 20 shortest kernel entries showed an average of 5.6 ± 0.07 mm. The mean phenotype value for KL16 and KL17 were 6.3 mm (with a range of 4.6–7.5 mm) and 6.2 mm (with a range of 5.1–7.1 mm), respectively (**Supplementary Figure S1**).

The correlation of traits among the different environments can be used as a measure of repeatability. Using the common entries between the 2 years, a moderate correlation (r = 0.44, p-value < 0.01) was observed for KW between the 2 years (**Figure 1C**). Similarly, we observed moderate correlation (r = 0.45, p-value < 0.01) between KL measurements from the 2 years (**Figure 1D**). The broad-sense heritability for both KW and KL, based on measurements in the 2 years, turned out to be 0.61 and 0.55, respectively. The correlation of data between the 2 years and measures of heritability suggests that both KW and KL are reasonably stable traits across years. The correlation of KW and KL BLUP values across 323 lines over 2 years was r = 0.20 (**Supplementary Figure S2**).

One of the claims about GY and KW in wheat breeding and selection history is that KW showed no significant increase or even decreased slightly while GY increased (Brancourt-Hulmel et al., 2003; Carver, 2009). Thus, one of our objectives was to investigate whether selection and breeding have increased or decreased kernel traits over the course of breeding history. Overall, the trend for KW was not consistent for the years across the four year-groups (**Supplementary Figure S3**). Though non-significant, for example, KW16 showed a slightly decreasing trend, with a mean of 36.1 mg across the entries developed before 1920 while 34.6 mg for entries developed after 2000. On the contrary, KW17 showed an increasing trend, with a mean of 33.4 mg before 1920 and a mean of 38.0 mg after 2000. The discrepancy of the trend between 2016 and 2017 could be due to an overrepresentation of Purdue-bred lines in the 2017 trial. The added Purdue lines (N = 35) exhibited greater KW (mean of 40.5 g), causing an increasing trend. KL16 remained unchanged over time while KL17 increased until 1960 then dropped afterward (**Supplementary Figure S3**).

Population Structure

We used all the 60,132 SNP markers in the analysis of population structure using PCA. The A, B, and D sub-genomes were represented by 35%, 44%, and 21% of SNPs, respectively. The first three PCs of marker data, altogether, explained 15.0% of the total variation and were used to draw a 3D-plot of the population structure. PC1 clearly grouped the germplasm based on the era of development, i.e., after or before 2000 (**Figure 2**). We also make the grouping for 3D-plot based on 2B.2G translocation form *T. timopheevii* represented by *TaSus2-2B* (**Figure 2**). The result revealed that the panel of 324 genotypes was clustered clearly into two groups, i.e., possessing or not possessing the 2B.2G translocation. The variation in this translocation is also reflected in the values of the PC1.

We performed model-based clustering using 16,313 tag SNPs, selected using the tagger function of Haploview (Barrett et al., 2005) with the parameters of "pairwise tagging only" and R^2 = 0.8. The result from this analysis revealed four sub-populations (**Figure 3**). The number of the entries assigned to each cluster ranged from 28 in Cluster3 to 177 in Cluster2. The detail descriptions of cluster membership is given in **Supplementary Table S2**. In total, 42.9% of the entries were developed by the Purdue Small Grains Breeding Program and therefore, membership of Purdue lines in all clusters is expected. Year of release and geographical region explained group membership partially. For example, Cluster1 was predominantly represented by germplasm developed before 1960 (91.4%) and Cluster2 was predominantly represented by germplasm developed before 2000 (93.8%). A majority (82.1%) of the entries in Cluster3 were developed after 2000. Cluster4 was mainly comprised of genotypes developed between 1920 and 2000. Cluster-sharing among entries originated in the different breeding programs could be an evidence of historical and recent germplasm exchange among breeding programs.

The differentiation among the four clusters and the four year-groups was assessed using the F_{ST}. The F_{ST} estimates for pairwise clusters revealed varied levels of allelic differentiation among the four clusters (**Supplementary Figure S4**). The Cluster3 was differentiated more from the other three clusters, with several of the SNP loci showing F_{ST} > 0.15 (Wright, 1978). Among the four clusters generated by the model-based analysis, a total of 457 SNP loci out of 60,132 showed significant F_{ST} (>0.15), indicating allelic differentiation. The majority of significant differentiations were observed between Cluster3 and Cluster4 (224 SNPs), followed by the comparison between Cluster1 and Cluster3, which yielded 215 significant (F_{ST} > 0.15) SNPs. The comparison between Cluster2 and Cluster3 yielded 102 significant SNPs. The least differentiated clusters were Cluster1 and Cluster2 with all the SNP loci showing a F_{ST} below 0.15.

GWAS and Allele Frequency Changes Over Time for KW

Any QTL in an individual year or combined 2-year analysis with *–log10(p)* > 3.5 was considered for further discussion. GWAS has resulted in 77 QTL for KW (**Figure 4B**, **Supplementary Figure S5**, and **Supplementary Table S3**), of which, 30 QTL were stacked in seemingly one genomic location on chromosome 3B. A pair-wise LD criterion of $R^2 \geq 0.75$ resolved all 30 QTL on 3B clustered into six LD block regions, with a minimum of one SNP to a maximum of 12 SNP markers per LD block (**Figure 4C**). A similar short-range LD block characterization for all the chromosomes, following $R^2 \geq 0.75$, enabled us to assign the 67 QTL to 26 genomic regions (**Supplementary Table S4**) distributed on chromosomes 1B, 2A, 2B, 3B, 4A, 4B, 5A, 6B, 7A, and 7B. Each of these regions was represented with a single SNP with the highest *−log10(p)*.

The highest *–log10(p)* value for KW was for a marker on chromosome 7B, designated as *QKWpur-7B.1* with *–log10(p)* of 5.4 and 4.5 in KW16 and KW1617, respectively. This marker

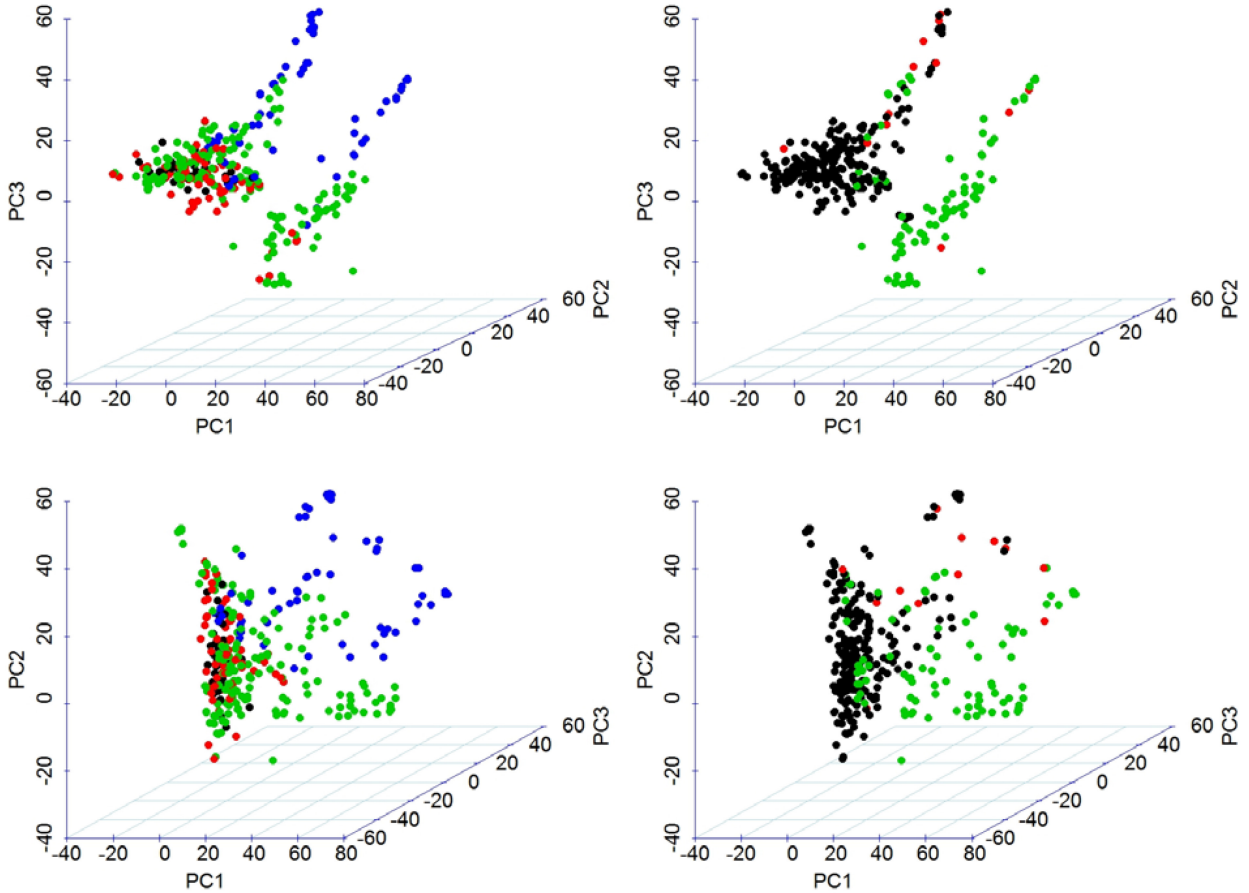

FIGURE 2 | Visualization of population stratification using three-dimensional scatter plot of the first three PCs estimated from 60,134 SNP markers. Top panel shows PC1, PC2, and PC3 in *x*-, *y*-, and *z*-axis configuration. Bottom panel shows PC1, PC3, and PC2 in *x*-, *y*-, and *z*-axis configuration. This allows a more-informed visualization. The left panel shows grouping by year-group and the right panel shows grouping by TaSus2-2B representing 2B.2G translocation from *T. timopheevii*.

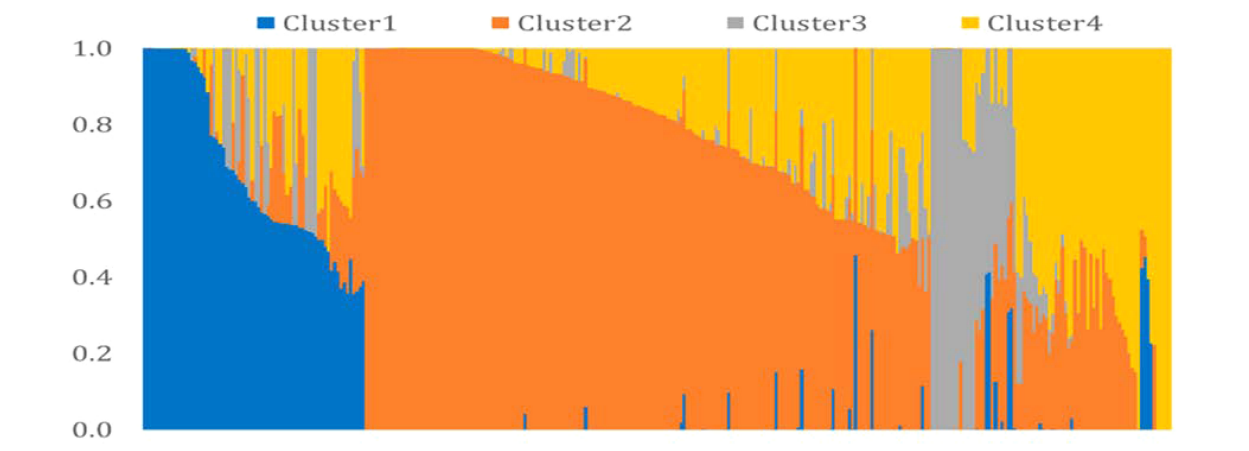

FIGURE 3 | Model-based cluster analysis using 16,313 tag SNPs selected by Haploview software.

explained 8.3% of phenotypic variation in 2016 with a marker effect of 0.9 mg. Out of 26 QTL identified for KW, 13 represented signals detected in 2016 (four of them also detected in the combined 2-year analysis). These 13 QTL detected for KW16 individually explained a low of 5.1% to a high of 8.3% of the variation in KW16. For KW17, eight genomic regions were

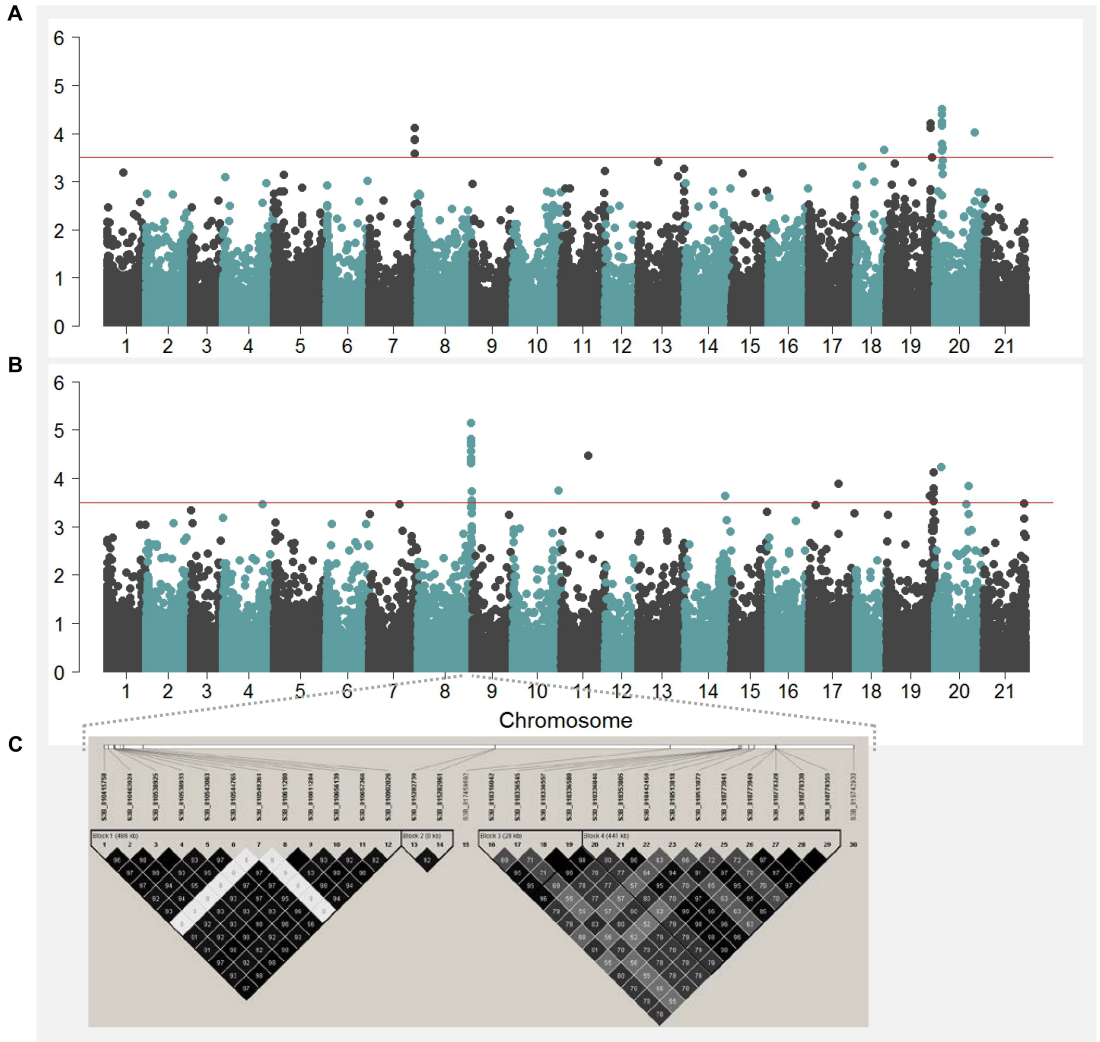

FIGURE 4 | Manhattan plots showing negative log *p*-values of SNPs tested across the 21 chromosomes (i.e., 1 = 1A, 2 = 1B, 3 = 1D, . . ., 20 = 7B, and 21 = 7D) obtained from GWAS by using 2-year BLUP values for kernel length **(A)** and kernel weight **(B)**. The haplotype blocks estimated for the 30 SNP markers located on the region significantly associated with kernel weight on chromosome 3B is shown in **(C)**.

identified (one of them also detected in the combined 2-year analysis). Individually, these eight QTL explained from a low of 5.5% to a high of 9.0% of the variation in KW17. Combined 2-year analysis revealed five unique QTL in addition to the four overlapping QTL of KW16 and one overlapping QTL of KW17. These 10 QTL for the combined 2-year data accounted for a low of 3.7% to a high of 5.0% to the phenotypic variation in KW1617.

We were interested in evaluating the frequency of favorable alleles in the identified loci. Out of 26 loci, 13 showed lower than 50% and 13 showed higher than 50% frequency for the favorable alleles. The trend of these allele frequency changes was given only for a subset of loci across year-groups in **Supplementary Figure S6**. When evaluated over the four year-groups, the frequency of favorable alleles decreased in 18 out of 26 of identified loci. The frequency of favorable alleles increased in four loci. For the remaining loci, it did not show a clear trend.

GWAS and Allele Frequency Changes Over Time for KL

We considered any QTL in an individual year or combined 2-year analysis with $-log10(p) > 3.5$ as significant and discussed further. GWAS has resulted in 45 QTL for KL (**Figure 4A**, **Supplementary Figure S5**, and **Supplementary Table S5**). With short-range LD block characterization for all the chromosomes, with criteria of considering SNPs with $R^2 \geq 0.75$ in one LD block, we assigned the 45 QTL to 27 genomic regions (**Supplementary Table S6**) distributed on chromosomes 1A, 1B, 2A, 2B, 2D, 3A, 3B, 4D, 6A, 6B, 7A, 7B, and 7D. Each genomic region was represented with a single SNP with the highest $-log10(p)$. The highest $-log10(p)$ value for KL was for a marker on chromosome 7B, designated as *QKLpur-7B.3* with $-log10(p)$ of 4.5 in KL1617. This genomic region explained 4.8% of phenotypic variation in KL1617 with a marker effect of 0.05 mm. Eleven of the genomic

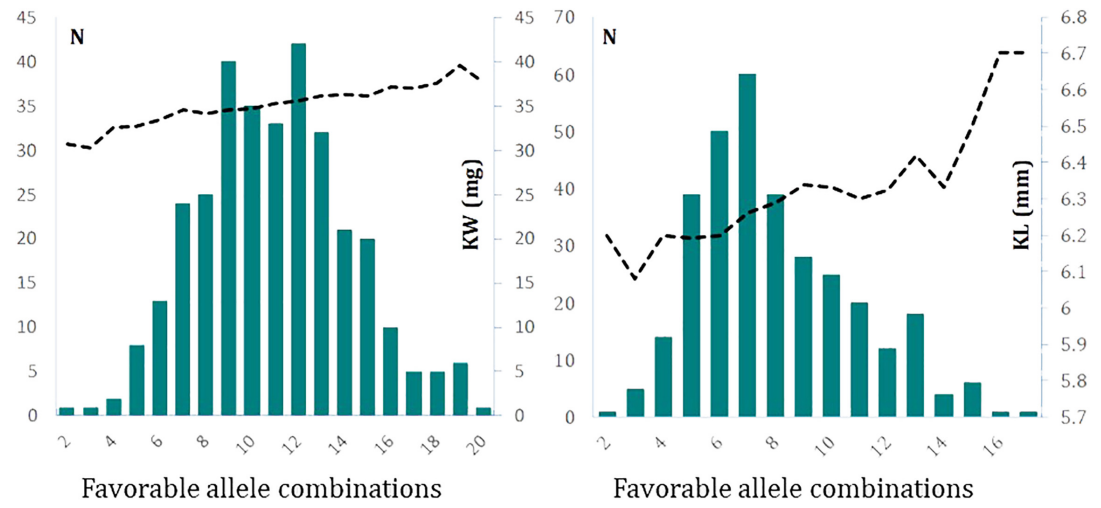

FIGURE 5 | The frequency distribution of combinations of favorable alleles with their cumulative effects on determination of kernel weight **(Left)** and kernel length **(Right)**. The data used to produce these graphs are based on 2-year BLUP values. The left *y*-axis of each graph represents the frequency of lines and the right *y*-axis of graphs represent kernel weight in mg **(Left)** and kernel length in mm **(Right)**.

TABLE 2 | Effects of allelic variation of previously reported agronomic loci/genes on kernel weight and kernel length in the current mapping panel.

KASP assay	Frequency (variant)	Kernel weight			Kernel length		
		Mean AA	Mean BB	*p*-value	Mean AA	Mean BB	*p*-value
Rht-B1	231 (Rht-B1a)/88 (Rht-B1b)	35.36	36.38	0.1015	6.28	6.19	0.0364
Rht_B1a_160IND	205 (Rht-B1a)/116 (Rht-B1a+160)	35.59	35.56	0.9421	6.24	6.28	0.3506
Rht_B1_197IND	315 (Rht-B1a)/8 (Rht-B1a+197)	35.66	34.36	0.3451	6.25	6.46	0.1101
Rht-D1	276 (Rht-D1a)/45 (Rht-D1b)	35.80	34.79	0.1211	6.26	6.25	0.9481
Ppd-A1	256 (Ppd-A1a)/53 (Ppd-A1a.1_insens)	35.42	36.18	0.2588	6.26	6.27	0.8326
Ppd-D1-Ciano67	271 (Ppd-D1a)/37 (Ppd-D1a_Ciano67_insens)	35.51	36.02	0.5506	6.27	6.26	0.9127
Ppd-D1-Mercia	269 (Ppd-D1a)/47 (Mercia_type_insertion)	35.75	34.38	0.0188	6.25	6.29	0.4281
Ppd-D1-Norstar	130 (Ppd-D1a)/189 (Norstar_type_deletion)	35.99	35.40	0.2437	6.24	6.27	0.5413
TaSus2-2B	85 (TaSus2-2B)/226 (no TaSus2-2B)	35.32	35.66	0.5634	6.15	6.31	0.0003
TaCWI-4A	221 (Hap-4A-C)/88 (Hap-4A-T)	35.49	35.80	0.5750	6.25	6.30	0.2532
TaTGW6-A1	171 (TaTGW6-A1b)/143 (TaTGW6-A1b)	35.52	35.56	0.9400	6.25	6.27	0.5466
TaGS-D1	108 (TaGS-D1a)/199 (TaGS-D1b)	35.43	35.41	0.9661	6.33	6.23	0.0200
TaGW2	305 (TaGW2)/16 (TaGW2_SS-MPV57)	35.78	32.14	0.0006	6.25	6.37	0.3133

regions were detected in 2016, with three of them also detected in the combined 2-year analysis. These 11 QTL detected for KL16 individually explained from a low of 4.7% to a high of 6.1% of the variation in KL16. For KL17, eight genomic regions were identified, with individual QTL explaining a low of 5.5% to a high of 6.1% of the variation in KW17. The combined analysis revealed eight unique QTL in addition to the three overlapping QTL of KL16. These 11 genomic regions identified for KL1617 accounted from a low of 3.6% to a high of 4.8% to the phenotypic variation in KW1617.

The trend of these allele frequency changes was given only for a subset of loci across the YG in **Supplementary Figure S7**. Of the 27 loci, seven were higher than 50% in favorable allele frequency while the remaining loci were lower than 50% for favorable allele frequency (data not shown). Fourteen loci showed a decrease in frequency of favorable alleles across the four year-groups. Six loci exhibited an increasing trend of favorable allele across the four year-groups. The remaining seven loci did not show a clear trend across the four year-groups.

Cumulative Effect of Identified Loci on KW

We were also interested to see up to how many favorable alleles are naturally present in a given germplasm. To do this, we counted the number of germplasm that accumulated from the lowest to the highest number of favorable alleles in the association panel. The frequency distribution of number of favorable alleles identified for KW in the germplasm followed a normal distribution (**Figure 5**). For the 26 identified loci for KW, we found lines with a minimum of two favorable alleles and lines with a maximum of 20 favorable alleles. Majority of

TABLE 3 | Candidate genes within the identified regions controlling kernel weight and their putative physiological roles.

QTL loci	Gene	Protein	Function	Reference
QKWpur-2B.1	TraesCS2B01G034100	Glycosyltransferase	Role in the biosynthesis of oligosaccharides, polysaccharides, and glycoconjugates	Breton et al., 2006; Lairson et al., 2008
QKWpur-2D.1	TraesCS2D01G020800	Photosystem II reaction center protein K	Photosynthesis	Vinyard et al., 2013; Caffarri et al., 2014
QKWpur-2D.1	TraesCS2D01G020900	Photosystem II reaction center protein I	Photosynthesis	Vinyard et al., 2013; Caffarri et al., 2014
QKWpur-2D.1	TraesCS2D01G021000	Photosystem II D2 protein	Photosynthesis	Vinyard et al., 2013; Caffarri et al., 2014
QKWpur-2D.1	TraesCS2D01G020200	Apyrase	Role in regulating growth and development	Riewe et al., 2008
QKWpur-2D.2	TraesCS2D01G141100	E3 Ubiquitin ligase family protein	Role in ubiquitin pathway	Li and Li, 2014
QKWpur-3B.1	TraesCS3B01G582000	Histone-lysine *N*-methyltransferase	Epigenetic regulation of expression (changes in DNA methylation or histone modification states)	Pontvianne et al., 2010
QKWpur-3B.4	TraesCS3B01G598100	Pectinesterase	Cellular adhesion and stem elongation	Micheli, 2001
QKWpur-3B.4	TraesCS3B01G597100	Phosphoenolpyruvate carboxykinase (ATP)	photosynthetic CO_2-concentrating mechanisms of C4 photosynthesis [9] and crassulacean acid metabolism	Leegood and Walker, 2003
QKWpur-3B.4	TraesCS3B01G598200	Glycosyltransferase	Role in the biosynthesis of oligosaccharides, polysaccharides, and glycoconjugates	Breton et al., 2006; Lairson et al., 2008
QKWpur-3B.4	TraesCS3B01G595200	RING/U-box superfamily protein	Role in ubiquitin pathway	Yee and Goring, 2009
QKWpur-3B.4	TraesCS3B01G595400	Embryogenesis transmembrane protein-like	Involve in hormone transport system active during embryogenesis	Jahrmann et al., 2005
QKWpur-4A.2	TraesCS4A01G028000	Pectinesterase	Cellular adhesion and stem elongation	Micheli, 2001
QKWpur-4A.3	TraesCS4A01G440500	Protein nrt1 ptr family 1.2	Nitrate transporters in plants: structure, function and regulation	Forde, 2000
QKWpur-4A.3	TraesCS4A01G440600	Protein nrt1 ptr family 1.2	Nitrate transporters in plants: structure, function and regulation	Forde, 2000
QKWpur-4A.3	TraesCS4A01G440700	Protein nrt1 ptr family 1.2	Nitrate transporters in plants: structure, function and regulation	Forde, 2000
QKWpur-4B	TraesCS4B01G193000	6-phosphofructo-2-kinase/fructose-2, 6-bisphosphatase	Sucrose biosynthesis	Lunn, 2016
QKWpur-5A	TraesCS5A01G024700	Protein FANTASTIC FOUR 3	Potential to regulate shoot meristem size	Wahl et al., 2010
QKWpur-7A.1	TraesCS7A01G468200	SAUR-like auxin-responsive protein family	Role in auxin-mediated cell elongation	Jain et al., 2006
QKWpur-7B.1	*TraesCS7B01G082500*	*O-fucosyltransferase family protein*	Role in cell-to-cell adhesion	Verger et al., 2016

entries (91.0%) possessed 6–16 favorable alleles. KW increased clearly with the increase in the number of favorable alleles. Using KW1617 BLUP values, the mean KW of entries with up to five favorable alleles combined (n = 12) was 32.3 g while the mean KW1617 for entries with ≥16 favorable alleles combined (n = 27) was 37.8 g, a difference of about 5.5 mg.

Commutative Effect of Identified Loci on KL

Similar to the procedure performed for KW, considering the 27 identified loci for KL, we found lines with a minimum of two favorable alleles combined to lines with a maximum of 17 favorable alleles combined. The majority of entries (94.4%) possessed 4–13 favorable alleles combined. Increases in the number of the combinations of favorable alleles clearly increased KL (**Figure 5**). Using KL1617 BLUP values, the mean KL of entries with up to five favorable alleles combined (n = 59) was 6.17 mm while the mean KL for entries with ≥12 favorable alleles combined (n = 42) was 6.41 mm, a difference of about 0.23 mm.

Effect of Previously Known Loci/Genes

The t-test results of comparing KW and KL of lines homozygous for alternate alleles of KASP markers is shown in **Table 2**. Most of loci/genes tested did not show a significant effect on KW and KL of this specific population. Of the six grain-related KASP markers tested, *TaGW2* has shown to be significantly associated with KW (p-value < 0.001) while *TaSus2-2B* and *TaGS-D1* were significantly associated with KL, with p-values < 0.001 and 0.02, respectively. The plant height loci *Rht-B1* was significant (p-value < 0.05) for KL, where the wild-type tall allele was associated with longer KL. The Mercia allele at the *Ppd-D1* locus has been shown to be significant for KW (p-value < 0.05).

TABLE 4 | Candidate genes within the identified regions controlling kernel length and their putative physiological roles.

QTL loci	SNP	Gene	Protein	Function	Reference
QKLpur-1D	S1D_445262848	TraesCS1D01G363700	Beta-galactosidase	Regulate cytokinins	Song et al., 2010
QKLpur-2A.2	S2A_719213280	TraesCS2A01G483000	Glycosyltransferase	Role in the biosynthesis of oligosaccharides, polysaccharides, and glycoconjugates	Breton et al., 2006; Lairson et al., 2008
QKLpur-3A.1	S3A_593313534	TraesCS3A01G343800	Photosystem I reaction center subunit VIII	Photosynthesis	Vinyard et al., 2013; Caffarri et al., 2014
QKLpur-3A.2	S3A_700575251	TraesCS3A01G467300	E3 ubiquitin-protein ligase BRE1-like 2	Role in ubiquitin pathway	Li and Li, 2014
QKLpur-3A.2	S3A_700575251	TraesCS3A01G467000	Late embryogenesis abundant (LEA) protein	Role in desiccation tolerance	
QKLpur-3A.3	S3A_700575251	TraesCS3A01G469200	Late embryogenesis abundant (LEA) protein	Role in desiccation tolerance	
QKLpur-6A	S6A_131449965	TraesCS6A01G149200	Ubiquitin-conjugating enzyme	Role in ubiquitin pathway	Li and Li, 2014
QKLpur-6D	S6D_436639209	TraesCS6D01G334300	Protein pelota homolog	Role in meiotic cell division	Eberhart and Wasserman, 1995; Caryl et al., 2000
QKLpur-7A.3	S7A_691163936	TraesCS7A01G501600	RING/U-box superfamily protein	Role in ubiquitin pathway	Yee and Goring, 2009

Candidate Gene Identification

The annotated wheat reference genome was used to pull out high confidence protein-coding genes that are in the vicinity (±250 kb) of the polymorphic sites. This gene search has resulted in a total of 258 genes for KW (**Supplementary Table S7**) and 235 genes for KL (**Supplementary Table S8**). A short list of identified genes is categorized into functional groups of (1) cell cycle related genes, (2) carbohydrate metabolism and transport, (3) nitrogen metabolism and transport, (4) cell wall, (5) plant hormones, (6) post-translation modifications such as ubiquitination, and (7) seed maturation and biological events that resemble stress responses (**Tables 3**, **4**).

DISCUSSION

Much of the genetic gains for GY has been attributed to the increases in GN, while KW generally remained unchanged if not decreased (Sayre et al., 1997; Brancourt-Hulmel et al., 2003; Carver, 2009; Hawkesford et al., 2013). We could not conclude a definitive trend for KW and KL over the breeding history. Though a long-standing belief that correlation of GN and KW is negative, Miralles and Slafer (1995) and Acreche and Slafer (2006) argued that this negativity is not due to competition between grains. That means, it is possible to develop progeny with high KW and GN concurrently by carefully selecting parents, as was evidenced by the work of Bustos et al. (2013). Therefore, there may exist an untapped potential in KW to improve GY if given due consideration in the variety development process. While further increases in GY can be dependent on maintaining, if not increasing, KW, an alternative breeding strategy could be to increase KW while maintaining GN or increasing KW and GN simultaneously. Careful recycling of high KW accessions including those developed before 1920 could improve kernel traits and ultimately result in gains in GY.

In this study, we detected 26 regions for KW and 27 regions for KL on most of the chromosomes, indicating that these traits are controlled by a complex genetic system. Previously, a large number of QTL for KW and dimension traits (kernel length, width, and thickness) have been reported across all 21 chromosomes of wheat (McCartney et al., 2005; Röder et al., 2008; Jiang et al., 2011; Bednarek et al., 2012; Deol et al., 2013; Simmonds et al., 2014; Hanif et al., 2015; Jiang et al., 2015; Su et al., 2016). Our evaluation of some of the previously reported genes and related functional markers like Kompetitive Allele Specific PCR (KASP) markers for kernel-related traits revealed that most of them had no significant effect of KW and KL in this panel. The exceptions were *TaGW2* for KW; and *TaSus2-2B* and *TaGS-D1* for KL. The non-significant effect for most of the loci may be that these genes are background dependent, inviting further evaluation of the effect of these genes in the different genetic background.

Kernel weight, as one of the main GY determinant (Simmonds et al., 2014), holds a very high heritability, reaching to h^2 = 87% (Bergman et al., 2000). In the current study, we also reported high heritability estimates of 61% for KW and 55% for KL. In allele enrichment schemes, breeders usually work to increase the frequency of favorable alleles. Our data suggest that favorable alleles at QKW_{pur}*-3B.1*, QKW_{pur}*-4A.1*, QKW_{pur}*-4A.2*, and QKW_{pur}*-5B.1* having low frequencies (3–9%) in germplasm released after 2000 and are prospect targets of selection for KW improvement. Similarly, loci QKL_{pur}*-2A.1*, QKL_{pur}*-2D*, QKL_{r}*-3A.2*, $QKL_{pur\_}$*3A.3*, QKL_{pur}*-3A.4*, QKL_{pur}*-4D and* $QKL_{pur\_}$*6B* could be potential targets for breeding via enriching the favorable allele frequency in the current breeding populations.

Wheat lags diploid model plants such as rice and Arabidopsis for the availability of genome-wide resources and tools. Recently, mutant resources in tetraploid and hexaploid wheat have become available[4]. In addition, the wheat reference genome

[4]http://www.wheat-tilling.com/

IWGSC RefSeq v1.0 annotation v1.0[5] (see footnote 2) made it possible to connect next-generation sequencing-based markers to candidate gene identification in GWAS studies using a position-dependent strategy. In our study, we assessed the genes within 250 kb of the QTL loci and listed potential candidate genes.

Kernels that have the potential for growth and are well filled during grain-fill period weigh more (Jenner et al., 1991; Altenbach and Kothari, 2004). A fine component of sink-strength is grain enlargement, which is enforced by endosperm cell division followed by water uptake (Jenner et al., 1991; Emes et al., 2003; Altenbach and Kothari, 2004). Source-strength, on the other hand, is an expression of supply of assimilates, i.e., starch and storage protein through current photosynthesis or remobilization of reserves from vegetative tissues (Bidinger et al., 1977; Schnyder, 1993; Gebbing and Schnyder, 1999). The conceptual framework for grain development may involve processes such as cell division, enlargement, and embryogenesis; photosynthesis, carbohydrate metabolism, and nitrogen metabolism; and post-translational modifications. Thus, our discussion for candidate genes for KW and KL concentrate on genes involved in the above-mentioned processes.

Grain enlargement commences with fertilization, wrapped-up within about 20 days after fertilization, and it also coincides with the period of mitotic activity (Jenner et al., 1991), as was observed in this study. The association with the largest signal [$-log_{10}(p)$ = 5.4] was QKW_{pur}-*7B.1* and this locus was found within 107 kb from *TraesCS7B01G082500*, which codes for *O-fucosyltransferase family protein* (**Table 3**). This protein was reported to have a function in cell-to-cell adhesion during plant growth and development (Verger et al., 2016). The gene *TraesCS3B01G595400* was in proximity of QKW_{pur}-*3B.4* [$-log_{10}(p)$ = 3.8] and encodes an embryogenesis transmembrane protein-like (**Table 3**). Jahrmann et al. (2005) highlighted that an embryogenesis transmembrane protein involved in hormone transport during embryogenesis. The *TraesCS5A01G024700* encoding for a *FANTASTIC FOUR 3* was associated with QKW_{pur}-*5A* [$-log_{10}(p)$ = 3.6], is potentially involved in regulating shoot meristem size (Wahl et al., 2010). A *SAUR-like auxin-responsive* protein family (*TraesCS7A01G468200*) that we show it to be associated with QKW_{pur}-*7A.1* [$-log_{10}(p)$ = 3.6], may have a role in auxin-mediated cell elongation (Jain et al., 2006). The *QKLpur-6D* [$-log_{10}(p)$ = 4.1] is within ±250 kb of *TraesCS6D01G334300*, a gene that encodes for protein pelota homolog (**Table 4**), previously reported to have a role in meiotic cell division (Caryl et al., 2000).

Kernel development is wrapped up by maturation. Tang et al. (2016) indicated that late embryogenesis abundant (LEA) genes become abundant during the late stages of seed development and enable the maturing seeds to acquire the desiccation tolerance. Temporal differences in expression of these genes may be a good signal for differences in the arrest of enlargement of the growing kernels. Two loci responsible for KL, i.e.; QKL_{pur}-*3A.2* [$-log_{10}(p)$ = 3.8] and QKL_{pur}-*3A.3* [$-log_{10}(p)$ = 4.1] were linked to wheat genes TraesCS3A01G467000 and TraesCS3A01G469200, which are predicted to encode *late embryogenesis abundant protein* (**Table 4**).

The QTL on 2D, QKW_{pur}-*2D.1* [$-log_{10}(p)$ = 3.7], was found to be associated with *Apyrase* (**Table 3**). Riewe et al. (2008) silenced *apyrase* gene in potato using RNAi that led to less than 10% Apyrase activity. This ultimately changed the phenotypes in transgenic lines, including a general retardation in growth, an increase in tuber number per plant, and differences in tuber morphology.

Three genes *TraesCS2D01G020800*, *TraesCS2D01G020900*, and *TraesCS2D01G021000* encoding photosystem reaction center proteins were found near QKW_{pur}-*2D.1* with $-log_{10}(p)$ = 3.7 (**Table 3**). The photosystem II is the reaction center that uses light energy to split water into hydrogen and oxygen, and release electrons that will be transferred to the second photosynthetic reaction center called photosystem I (Caffarri et al., 2014). We also identified a gene which encodes for photosystem I reaction center subunit VIII (*TraesCS3A01G343800*) and is within ±250 kb of QKL_{pur}-*3A.1*, with $-log_{10}(p)$ = 3.6 (**Table 4**). As current assimilates filling the developing kernels are direct products of photosynthesis, the candidacy of these photosystem reaction proteins seems to be logical and is worth validation studies.

Starch accumulation accounts for 60–75% of kernel dry matter and mainly responsible for kernel size and yield (Rahman et al., 2000; De Gara et al., 2003). Sucrose is the most common form of carbohydrate transported from source to sink organs. Thirty-eight kilo base away from QKW_{pur}-*4B* [$-log_{10}(p)$ = 4.5], we identified *TraesCS4B01G193000* which encodes a *fructose-2,6-bisphosphatase* (**Table 3**) that is involved in the dephosphorylation step of *sucrose synthesis* (Lunn, 2016). Transgenic Arabidopsis plants with only 5% *fructose-2,6-bisphosphates* expression, as compared to wild-type plants, demonstrate altered partitioning of carbon between sucrose and starch (Draborg et al., 2001). McCormick and Kruger (2015) reported that the T-DNA insertional Arabidopsis mutant lines for *fructose-2,6-bisphosphates* showed reduced growth and seed yields compared with wild-type plants. This enzyme was also reported to play a role in the partitioning of photoassimilate in sorghum (Reddy, 1996) and wheat (Reddy, 2000).

A QTL was reported previously that enhances KW and GY in rice via increases in cell numbers, allowing grains to reach to higher potential sizes. This QTL, named *GW2* in rice, was found to be a *RING-type protein E3 ubiquitin ligase* activity, with loss of function mutant (Song et al., 2007). Our study resulted in identification of two loci, i.e.; *QKWpur-2D.2* [$-log_{10}(p)$ = 3.7] and QKL_{pur}-*3A.2* [$-log_{10}(p)$ = 3.8] that are associated with *E3 ubiquitin-protein ligase* via *TraesCS2D01G141100* and *TraesCS3A01G467300*, respectively (**Tables 3**, **4**).

CONCLUSION

This study utilized genome-based markers and resulted in the identification of loci and genes important to the determination

of grain traits. We have also demonstrated that GWAS results can be utilized to further investigate genomic regions to drive putative list of candidate genes that can be further validated. The immediate use of this data could be developing breeder friendly markers (i.e., KASP) that can be useful in breeding. Further functional genomic studies are crucial to validate the effect of the identified candidate genes on KW and dimension traits. Utilizing mutant resources developed recently (Krasileva et al., 2017) is one way to functionally validate the effect of these candidate genes in the determination of KW and KL.

DATA AVAILABILITY

The genotypic and phenotypic data pertaining to the analysis and conclusion are available via the link: https://de.cyverse.org/de/?type=data&folder=/iplant/home/shared/commons_repo/staging/Daba_KernelTriats_2018.

AUTHOR CONTRIBUTIONS

GB-G and PT executed genome-wide marker development at the Small Grains Genotyping Laboratory at USDA-ARS in Raleigh, NC, United States and participated in the writing of the manuscript. MM and SD designed the study, collected all the data, performed all the statistical and blast analysis, and wrote the manuscript. SD also conducted the SNP calling using IWGSv1.0.

FUNDING

This work was financially supported by Purdue College of Agriculture and the USDA Hatch grant 1013073.

ACKNOWLEDGMENTS

The authors are thankful to Dr. Harold Bockelman of USDA-NSGC for providing seed. The authors would like to thank the International Wheat Genome Sequencing Consortium (IWGSC, www.wheatgenome.org) for pre-publication access to IWGSC RefSeq v1.0 and annotation v1.0.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01045/full#supplementary-material

FIGURE S1 | Phenotypic distributions of kernel weight (top) and kernel length (bottom) measured in 2016 (left) and 2017 (right).

FIGURE S2 | Scatterplot showing correlation of BLUP values of KW and KL over the two years of study.

FIGURE S3 | Changes over the four year-groups (1 = before 1920, 2 = 1920 to 1960, 3 = 1960 to 2000, and 4 = after 2000) observed in kernel weight (top) for measurements in 2016 **(a)**, 2017 **(b)**, and for the BLUP values across the two years of study **(c)**; and in kernel length (bottom) for measurements in 2016 **(d)**, 2017 **(f)**, and the BLUP values across the two years of study **(f)**.

FIGURE S4 | Plots of F_{ST} statistics for pairs of sub-populations generated using the model-based clustering procedure.

FIGURE S5 | Manhattan plots showing negative log *p*-values of SNPs tested across the 21 chromosomes (i.e., 1 = 1A, 2 = 1B, 3 = 1D, ..., 20 = 7B, and 21 = 7D) for kernel weight (top) and kernel length (bottom) for traits measured in 2016 (left) and 2017 (right).

FIGURE S6 | Frequency of favorable alleles observed in each of the year-group for a selected number of loci controlling kernel weight.

FIGURE S7 | Frequency of favorable alleles observed in each of the year-group for a select number of loci controlling kernel length.

TABLE S1 | The accuracy of imputation at different levels of marker masking using LDKNNi procedure in TASSEL. We used 30 sites for LD estimation. The number of nearest neighbors of entries was 10.

TABLE S2 | Cluster membership of the 324 genotypes used in model-based clustering with the year in which the accession was registered at NSGC.

TABLE S3 | The GWAS statistics for each marker-trait association for kernel weight in 2016, 2017, and combined year data. The table includes variants, minor allele frequency (MAF), −logP, R2, and allelic effect.

TABLE S4 | The GWAS statistics after categorizing MTAs into QTL regions kernel weight in 2016, 2017, and combined year data. The table includes variants, minor allele frequency (MAF), −logP, R2, and allelic effect.

TABLE S5 | The GWAS statistics for each marker-trait association for kernel length in 2016, 2017, and combined year data. The table includes variants, minor allele frequency (MAF), −logP, R2, and allelic effect.

TABLE S6 | The GWAS statistics after categorizing MTAs into QTL regions kernel length in 2016, 2017, and combined year data. The table includes variants, minor allele frequency (MAF), −logP, R2, and allelic effect.

TABLE S7 | The putative candidate genes found nearby the polymorphic sites for kernel weight.

TABLE S8 | The putative candidate genes found nearby the polymorphic sites for kernel length.

REFERENCES

Acreche, M. M., and Slafer, G. A. (2006). Grain weight response to increases in number of grains in wheat in a Mediterranean area. *Field Crops Res.* 98, 52–59. doi: 10.1016/j.fcr.2005.12.005

Altenbach, S. B., and Kothari, K. M. (2004). Transcript profiles of genes expressed in endosperm tissue are altered by high temperature during wheat grain development. *J. Cereal Sci.* 40, 115–126. doi: 10.1016/j.jcs.2004.05.004

Aoki, N., Scofield, G. N., Wang, X.-D., Patrick, J. W., Offler, C. E., and Furbank, R. T. (2004). Expression and localisation analysis of the wheat sucrose transporter TaSUT1 in vegetative tissues. *Planta* 219, 176–184. doi: 10.1007/s00425-004-1232-7

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

Bednarek, J., Boulaflous, A., Girousse, C., Ravel, C., Tassy, C., Barret, P., et al. (2012). Down-regulation of the TaGW2 gene by RNA interference results in

decreased grain size and weight in wheat. *J. Exp. Bot.* 63, 5945–5955. doi: 10.1093/jxb/ers249

Bergman, C. J., Gualberto, D. G., Campbell, K. G., Sorrells, M. E., and Finney, P. L. (2000). Kernel morphology variation in a population derived from a soft by hard cross and associations with end-use quality traits. *J. Food Qual.* 23, 391–407. doi: 10.1111/j.1745-4557.2000.tb00566.x

Bidinger, F., Musgrave, R. B., and Fischer, R. A. (1977). Contribution of stored pre-anthesis assimilate to grain yield in wheat and barley. *Nature* 270, 431–433. doi: 10.1038/270431a0

Bradbury, P. J., Zhang, Z., Kroon, E. D., Casstevens, T. M., Ramdoss, Y., and EBuckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Brancourt-Hulmel, M., Doussinault, G., Lecomte, C., Bérard, P., Le Buanec, B., and Trottet, M. (2003). Genetic improvement of agronomic traits of winter wheat cultivars released in France from 1946 to 1992. *Crop Sci.* 43, 37–45. doi: 10.2135/cropsci2003.3700

Brisson, N., Gate, P., Gouache, D., Charmet, G., Ouryc, F.-X., and Huarda, F. (2010). Why are wheat yields stagnating in Europe? A comprehensive data analysis for France. *Field Crops Res.* 119, 201–212. doi: 10.1016/j.fcr.2010.07.012

Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J., and Imberty, A. (2006). Structures and mechanisms of glycosyltransferases. *Glycobiology* 16, 29R–37R. doi: 10.1093/glycob/cwj016

Bustos, D. V., Hasan, A. K., Reynolds, M. P., and Calderini, F. D. (2013). Combining high grain number and weight through a DH-population to improve grain yield potential of wheat in high-yielding environments. *Field Crops Res.* 145, 106–115. doi: 10.1016/j.fcr.2013.01.015

Caffarri, S., Tibiletti, T., Jennings, R. C., and Santabarbara, S. (2014). A comparison between plant photosystem I and photosystem II architecture and functioning. *Curr. Protein Pept. Sci.* 15, 296–331. doi: 10.2174/1389203715666140327102218

Carver, B. F. (2009). *Wheat Science and Trade*. Hoboken, NJ: Wiley-Blackwell. doi: 10.1002/9780813818832

Caryl, A. P., Lacroix, I., Jones, G. H., and Franklin, F. C. H. (2000). An *Arabidopsis* homologue of the *Drosophila* meiotic gene pelota. *Sex. Plant Reprod.* 12, 310–313. doi: 10.1007/s004970050200

Chang, C., Lu, J., Zhang, H.-P., Ma, C.-X., and Sun, G. (2016). Copy number variation of cytokinin oxidase gene Tackx4 associated with grain weight and chlorophyll content of flag leaf in common wheat. *PLoS One* 10:e0145970. doi: 10.1371/journal.pone.0145970

Chen, F., Zhu, Z., Zhou, X., Yan, Y., Dong, Z., and Cui, D. (2016). High-throughput sequencing reveals single nucleotide variants in longer-kernel bread wheat. *Front. Plant Sci.* 7:1193. doi: 10.3389/fpls.2016.01193

De Gara, L., de Pinto, M. C., Moliterni, V. M. C., and D'Egidio, M. G. (2003). Redox regulation and storage processes during maturation in kernels of Triticum durum. *J. Exp. Bot.* 54, 249–258. doi: 10.1093/jxb/erg021

Deol, K. K., Mukherjee, S., Gao, F., Brûlé-Babel, A., Stasolla, C., and Ayele, B. T. (2013). Identification and characterization of the three homeologues of a new sucrose transporter in hexaploid wheat (*Triticum Aestivum* L.). *BMC Plant Biol.* 13:181. doi: 10.1186/1471-2229-13-181

Draborg, H., Villadsen, D., and Nielsen, T. H. (2001). Transgenic Arabidopsis plants with decreased activity of fructose-6-phosphate,2-kinase/fructose-2,6-bisphosphatase have altered carbon partitioning. *Plant Physiol.* 126, 750–758. doi: 10.1104/pp.126.2.750

Eberhart, C. G., and Wasserman, S. A. (1995). The pelota locus encodes a protein required for meiotic cell division: an analysis of G2/M arrest in *Drosophila* spermatogenesis. *Development* 121, 3477–3486.

Emes, M. J., Bowsher, C. G., Hedley, C., Burrell, M. M., Scrase-Field, E. S. F., and Tetlow, I. J. (2003). Starch Synthesis and carbon partitioning in developing endosperm. *J. Exp. Bot.* 54, 569–575. doi: 10.1093/jxb/erg089

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

Forde, B. G. (2000). Nitrate transporters in plants: structure, function and regulation. *Biochim. Biophys. Acta* 1465, 219–235. doi: 10.1016/S0005-2736(00)00140-1

Gebbing, T., and Schnyder, H. (1999). Pre-anthesis reserve utilization for protein and carbohydrate synthesis in grains of wheat. *Plant Physiol.* 121, 871–878. doi: 10.1104/pp.121.3.871

Hanif, M., Gao, F., Liu, J., Wen, W., Zhang, Y., Rasheed, A., et al. (2015). TaTGW6-A1, an ortholog of rice TGW6, is associated with grain weight and yield in bread wheat. *Mol. Breed.* 36:1. doi: 10.1007/s11032-015-0425-z

Hawkesford, M. J., Araus, J.-L., Park, R., Calderini, D., Miralles, D., Shen, T., et al. (2013). Prospects of doubling global wheat yields. *Food Energy Secur.* 2, 34–48. doi: 10.1002/fes3.15

Hou, J., Jiang, Q., Hao, C., Wang, Y., Zhang, H., and Zhang, X. (2014). Global selection on sucrose synthase haplotypes during a century of wheat breeding. *Plant Physiol.* 164, 1918–1929. doi: 10.1104/pp.113.232454

Hu, M.-J., Zhang, H.-P., Cao, J.-J., Zhu, X.-F., Wang, S.-X., Jiang, H., et al. (2016). Characterization of an IAA-glucose hydrolase gene TaTGW6 associated with grain weight in common wheat (*Triticum Aestivum* L.). *Mol. Breed.* 36:25. doi: 10.1007/s11032-016-0449-z

Jahrmann, T., Bastida, M., Pineda, M., Gasol, E., Ludevid, M. D., Palacín, M., et al. (2005). Studies on the function of TM20, a transmembrane protein present in cereal embryos. *Planta* 222, 80–90. doi: 10.1007/s00425-005-1519-3

Jain, M., Tyagi, A. K., and Khurana, J. P. (2006). Molecular characterization and differential expression of cytokinin-responsive type-A response regulators in rice (*Oryza Sativa*). *BMC Plant Biol.* 6:1. doi: 10.1186/1471-2229-6-1

Jaiswal, V., Gahlaut, V., Mathur, S., Agarwal, P., Khandelwal, M. K., Khurana, J. P., et al. (2015). Identification of novel SNP in promoter sequence of TaGW2-6A associated with grain weight and other agronomic traits in wheat (*Triticum Aestivum* L.). *PLoS One* 10:e0129400. doi: 10.1371/journal.pone.0129400

Jenner, C. F., Ugalde, T. D., and Aspinall, D. (1991). The physiology of starch and protein deposition in the endosperm of wheat. *Funct. Plant Biol.* 18, 211–226. doi: 10.1071/PP9910211

Jiang, Q., Hou, J., Hao, C., Wang, L., Ge, H., Dong, Y., et al. (2011). The wheat (*T. aestivum*) sucrose synthase 2 gene (TaSus2) active in endosperm development is associated with yield traits. *Funct. Integr. Genomics* 11, 49–61. doi: 10.1007/s10142-010-0188-x

Jiang, Y., Jiang, Q., Hao, C., Hou, J., Wang, L., Zhang, H., et al. (2015). A yield-associated gene TaCWI, in wheat: its function, selection and evolution in global breeding revealed by haplotype analysis. *Theor. Appl. Genet.* 128, 131–143. doi: 10.1007/s00122-014-2417-5

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, H. 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

Krasileva, K. V., Vasquez-Gross, H. A., Howell, T., Bailey, P., Paraiso, F., Clissold, L., et al. (2017). Uncovering hidden variation in polyploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* 114, E913–E921. doi: 10.1073/pnas.1619268114

Lairson, L. L., Henrissat, B., Davies, G. J., and Withers, S. G. (2008). Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* 77, 521–555. doi: 10.1146/annurev.biochem.76.061005.092322

Leegood, R. C., and Walker R. P. (2003). Regulation and roles of phosphoenolpyruvate carboxykinase in plants. *Arch. Biochem. Biophys.* 414, 204–210. doi: 10.1016/S0003-9861(03)00093-6

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, N., and Li, Y. (2014). Ubiquitin-mediated control of seed size in plants. *Front. Plant Sci.* 5:332. doi: 10.3389/fpls.2014.00332

Ligges, U., and Maechler, M. (2003). 3D scatter plots: an r package for visualizing multivariate data. *J. Stat. Softw.* 8, 1–20. doi: 10.18637/jss.v008.i11

Lin, M., and Huybers, P. (2012). Reckoning wheat yield trends. *Environ. Res. Lett.* 7:024016. doi: 10.1088/1748-9326/7/2/024016

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

Lu, J., Chang, C., Zhang, H.-P., Wang, S.-X., Sun, G., Xiao, S.-H., et al. (2015). Identification of a novel allele of *TaCKX6a02* associated with grain size, filling rate and weight of common wheat. *PLoS One* 10:e0144765. doi: 10.1371/journal.pone.0144765

Lunn, J. E. (2016). *Sucrose Metabolism. eLS*. Chichester: John Wiley & Sons, Ltd. doi: 10.1002/9780470015902.a0021259.pub2

Ma, D., Yan, J., He, Z., Wu, L., and Xia, X. (2012). Characterization of a cell wall invertase gene TaCwi-A1 on common wheat chromosome 2A and development of functional markers. *Mol. Breed.* 29, 43–52. doi: 10.1007/s11032-010-9524-z

McCartney, C. A., Somers, D. J., Humphreys, D. G., Lukow, O., Ames, N., Noll, J., et al. (2005). Mapping quantitative trait loci controlling agronomic traits in the spring wheat cross RL4452 × 'AC domain. *Genome* 48, 870–883. doi: 10.1139/g05-055

McCormick, A. J., and Kruger, N. J. (2015). Lack of fructose 2,6-bisphosphate compromises photosynthesis and growth in Arabidopsis in fluctuating environments. *Plant J.* 81, 670–683. doi: 10.1111/tpj.12765

Micheli, F. (2001). Pectin methylesterases: cell wall enzymes with important roles in plant physiology. *Trends Plant Sci.* 6, 414–419. doi: 10.1016/S1360-1385(01)02045-3

Miralles, D. J., and Slafer, G. A. (1995). Individual grain weight responses to genetic reduction in culm length in wheat as affected by source-sink manipulations. *Field Crops Res.* 43, 55–66. doi: 10.1016/0378-4290(95)00041-N

Mohammadi, M., Blake, T. K., Budde, A. D., Chao, S., Hayes, P. M., Horsley, R. D., et al. (2015). A genome-wide association study of malting quality across eight U.S. barley breeding programs. *Theor. Appl. Genet.* 128, 705–721. doi: 10.1007/s00122-015-2465-5

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.-Y., and Myles, S. (2015). LinkImpute: fast and accurate genotype imputation for nonmodel organisms. *G3* 5, 2383–2390. doi: 10.1534/g3.115.021667

Pflieger, S., Lefebvre, V., and Causse, M. (2001). The candidate gene approach in plant genetics: a review. *Mol. Breed.* 7, 275–291. doi: 10.1023/A:1011605013259

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5, 103–113. doi: 10.3835/plantgenome2012.06.0006

Pontvianne, F., Blevins, T., and Pikaard, C. S. (2010). Arabidopsis Histone Lysine Methyltransferases. *Adv. Bot. Res.* 53, 1–22. doi: 10.1016/S0065-2296(10)53001-5

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

Rahman, S., Li, Z., Batey, I., Cochrane, M. P., Appels, R., and Morell, M. (2000). Genetic alteration of starch functionality in wheat. *J. Cereal Sci.* 31, 91–110. doi: 10.1006/jcrs.1999.0291

Rasheed, A., Wen, W., Gao, F., Zhai, S., Jin, H., Liu, J., et al. (2016). Development and validation of KASP assays for genes underpinning key economic traits in bread wheat. *Theor. Appl. Genet.* 129, 1843–1860. doi: 10.1007/s00122-016-2743-x

Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C., and Foley, J. A. (2012). Recent patterns of crop yield growth and stagnation. *Nat. Commun.* 3:1293. doi: 10.1038/ncomms2296

Reddy, R. A. (1996). Fructose 2,6-bisphosphate-modulated photosynthesis in sorghum leaves grown under low water regimes. *Phytochemistry* 43, 319–322. doi: 10.1016/0031-9422(96)00052-0

Reddy, R. R. (2000). Photosynthesis and fructose 2,6-bisphophate content in water stressed wheat leaves. *Cereal Res. Commun.* 28, 131–137.

Riewe, D., Grosman, L., Fernie, A. R., Wucke, C., and Geigenberger, P. (2008). The potato-specific apyrase is apoplastically localized and has influence on gene expression, growth, and development. *Plant Physiol.* 147, 1092–1109. doi: 10.1104/pp.108.117564

Röder, M. S., Huang, X. Q., and Börner, A. (2008). Fine mapping of the region on wheat chromosome 7D controlling grain weight. *Funct. Integr. Genomics* 8, 79–86. doi: 10.1007/s10142-007-0053-8

Sayre, K. D., Rajaram, S., and Fischer, R. A. (1997). Yield potential progress in short bread wheats in northwest Mexico. *Crop Sci.* 37, 36–42. doi: 10.2135/cropsci1997.0011183X003700010006x

Schnyder, H. (1993). The role of carbohydrate storage and redistribution in the source-sink relations of wheat and barley during seed filling-a review. *New Phytol.* 123, 233–245. doi: 10.1111/j.1469-8137.1993.tb03731.x

Simmonds, J., Scott, P., Leverington-Waite, M., Turner, A. S., Brinton, J., Korzun, V., et al. (2014). Identification and independent validation of a stable yield and thousand grain weight QTL on chromosome 6A of hexaploid wheat (*Triticum Aestivum* L.). *BMC Plant Biol.* 14:191. doi: 10.1186/s12870-014-0191-9

Song, J., Jiang, L., and Jameson, P. E. (2010). "Identification and quantitative expression of cytokinin regulatory genes during seed and leaf development in wheat," in *Proceedings of the Joint Symposium Between the Agronomy Society of New Zealand and the New Zealand Grassland Association: Seed Symposium: Seeds for Futures*, eds C. R. McGill and J. S. Rowarth (Palmerston North: Massey University).

Song, J., Jiang, L., and Jameson, P. E. (2012). Co-ordinate regulation of cytokinin gene family members during flag leaf and reproductive development in wheat. *BMC Plant Biol.* 12:78. doi: 10.1186/1471-2229-12-78

Song, X. J., Huang, W., Shi, M., Zhu, M. Z., and Lin, H. X. (2007). A QTL for rice grain width and weight encodes a previously unknown RING-Type E3 ubiquitin ligase. *Nat. Genet.* 39, 623–630. doi: 10.1038/ng2014

Su, Z., Hao, C., Wang, L., Dong, Y., and Zhang, X. (2011). Identification and development of a functional marker of TaGW2 associated with grain weight in bread wheat (*Triticum Aestivum* L.). *Theor. Appl. Genet.* 122, 211–223. doi: 10.1007/s00122-010-1437-z

Su, Z., Jin, S., Lu, Y., Zhang, G., Chao, S., and Bai, G. (2016). Single nucleotide polymorphism tightly linked to a major QTL on chromosome 7A for both kernel length and kernel weight in wheat. *Mol. Breed.* 36:15. doi: 10.1007/s11032-016-0436-4

Tang, X., Wang, H., Chu, L., and Shao, H. (2016). KvLEA, a new isolated late embryogenesis abundant protein gene from *Kosteletzkya virginica* responding to Multiabiotic stresses. *Biomed Res. Int.* 2016:9823697. doi: 10.1155/2016/9823697

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

Verger, S., Chabout, S., Gineau, E., and Mouille, G. (2016). Cell adhesion in plants is under the control of putative O-fucosyltransferases. *Development* 143, 2536–2540. doi: 10.1242/dev.132308

Vinyard, D. J., Ananyev, G. M., and Dismukes, G. C. (2013). Photosystem II: the reaction center of oxygenic photosynthesis. *Annu. Rev. Biochem.* 82, 577–606. doi: 10.1146/annurev-biochem-070511-100425

Wahl, V., Brand, L. H., Guo, Y.-L., and Schmid, M. (2010). The FANTASTIC FOUR proteins influence shoot meristem size in *Arabidopsis thaliana*. *BMC Plant Biol.* 10:285. doi: 10.1186/1471-2229-10-285

Wiersma, J. J., Busch, R. H., Fulcher, G. G., and Hareland, G. A. (2001). recurrent selection for kernel weight in spring wheat contribution from minnesota agric. *Crop Sci.* 41, 999–1005. doi: 10.2135/cropsci2001.414999x

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.* 15, 323–354. doi: 10.1111/j.1469-1809.1949.tb02451.x

Wright, S. (1978). *Evolution and the Genetics of Populations. Variability Within and Among Natural Populations*, Vol. 4. Chicago, IL: University of Chicago Press.

Yee, D., and Goring, D. R. (2009). The diversity of plant U-box E3 ubiquitin ligases: from upstream activators to downstream target substrates. *J. Exp. Bot.* 60, 1109–1121. doi: 10.1093/jxb/ern369

Zhang, J., Liu, W., Yang, X., Gao, A., Li, X., Wu, X., et al. (2010). Isolation and characterization of two putative cytokinin oxidase genes related to grain number per spike phenotype in wheat. *Mol. Biol. Rep.* 38, 2337–2347. doi: 10.1007/s11033-010-0367-9

Zhang, L., Zhao, Y.-L., Gao, L.-F., Zhao, G.-Y., Zhou, R.-H., Zhang, B.-S., et al. (2012). TaCKX6-D1, the ortholog of rice OsCKX2, is associated with grain weight in hexaploid wheat. *New Phytol.* 195, 574–584. doi: 10.1111/j.1469-8137.2012.04194.x

Zhang, Y., Liu, J., Xia, X., and He, Z. (2014). TaGS-D1, an ortholog of rice OsGS3, is associated with grain weight and grain length in common wheat. *Mol. Breed.* 34, 1097–1107. doi: 10.1007/s11032-014-0102-7

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

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 JC and handling Editor declared their shared affiliation.

frontiers in Plant Science

REVIEW
published: 23 July 2018
doi: 10.3389/fpls.2018.01054

Finger Millet [*Eleusine coracana* (L.) Gaertn.] Improvement: Current Status and Future Interventions of Whole Genome Sequence

S. Antony Ceasar[1,2], T. Maharajan[1], T. P. Ajeesh Krishna[1], M. Ramakrishnan[1], G. Victor Roch[1], Lakkakula Satish[3,4] and Savarimuthu Ignacimuthu[1]**

[1] Division of Plant Biotechnology, Entomology Research Institute, Loyola College, Chennai, India, [2] Functional Genomics and Plant Molecular Imaging Lab, University of Liege, Liege, Belgium, [3] Department of Biotechnology Engineering, Ben-Gurion University of the Negev, Beersheba, Israel, [4] The Jacob Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Beersheba, Israel

OPEN ACCESS

Edited by:
Jun Yang,
Shanghai Chenshan Plant Science Research Center (CAS), China

Reviewed by:
Ren-You Gan,
Shanghai Jiao Tong University, China
Parvathi Madathil Sreekumar,
Harvard University, United States
Gengyun Zhang,
Beijing Genomics Institute (BGI), China

***Correspondence:**
S. Antony Ceasar
antony_sm2003@yahoo.co.in
Savarimuthu Ignacimuthu
eriloyola@hotmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: 04 April 2018
Accepted: 28 June 2018
Published: 23 July 2018

Citation:
Antony Ceasar S, Maharajan T, Ajeesh Krishna TP, Ramakrishnan M, Victor Roch G, Satish L and Ignacimuthu S (2018) Finger Millet [Eleusine coracana (L.) Gaertn.] Improvement: Current Status and Future Interventions of Whole Genome Sequence. Front. Plant Sci. 9:1054. doi: 10.3389/fpls.2018.01054

The whole genome sequence (WGS) of the much awaited, nutrient rich and climate resilient crop, finger millet (*Eleusine coracana* (L.) Gaertn.) has been released recently. While possessing superior mineral nutrients and excellent shelf life as compared to other major cereals, multiploidy nature of the genome and relatively small plantation acreage in less developed countries hampered the genome sequencing of finger millet, disposing it as one of the lastly sequenced genomes in cereals. The genomic information available for this crop is very little when compared to other major cereals like rice, maize and barley. As a result, only a limited number of genetic and genomic studies has been undertaken for the improvement of this crop. Finger millet is known especially for its superior calcium content, but the high-throughput studies are yet to be performed to understand the mechanisms behind calcium transport and grain filling. The WGS of finger millet is expected to help to understand this and other important molecular mechanisms in finger millet, which may be harnessed for the nutrient fortification of other cereals. In this review, we discuss various efforts made so far on the improvement of finger millet including genetic improvement, transcriptome analysis, mapping of quantitative trait loci (QTLs) for traits, etc. We also discuss the pitfalls of modern genetic studies and provide insights for accelerating the finger millet improvement with the interventions of WGS in near future. Advanced genetic and genomic studies aided by WGS may help to improve the finger millet, which will be helpful to strengthen the nutritional security in addition to food security in the developing countries of Asia and Africa.

Keywords: finger millet, whole genome sequence (WGS), millets, nutrient transport, genomic resources

INTRODUCTION

Finger millet (*Eleusine coracana* (L.) Gaertn.) is an allotetraploid (2n = 4X = 36, AABB) belonging to the Family Poaceae and the genus *Eleusine*. The genome size of finger millet is 1,593 Mb and is a self-pollinated crop (Goron and Raizada, 2015). It is an annual herbaceous cereal crop widely grown and consumed by poor people in Africa and Asia. It contains rich amounts of protein,

mineral nutrient as compared to other major cereals like wheat, rice, and sorghum (Gupta et al., 2017; Sharma et al., 2017). Finger millet is well known for its exceptionally high calcium (Ca) content having about 0.34% in whole seeds as compared with 0.01–0.06% in most other cereals (Kumar et al., 2016; Gupta et al., 2017). The seeds are abundant source of dietary fiber, iron, essential amino acids viz., isoleucine, leucine, methionine, phenylalanine, pytates and trypsin inhibitory factors, and are also gluten-free (Chandra et al., 2016; Sood et al., 2016). Finger millet also has many health-promoting benefits such as hypoglycemic, hypocholesterolemic and anti-ulcerative effects (Chethan and Malleshi, 2007). The grain is used as flour in the preparation of cakes, bread and other pastry products, and also serves as a beneficial food for infants (Mgonja et al., 2007; Ceasar and Ignacimuthu, 2011). The seeds can be stored for more than 5 years without insect damage which makes it a most valuable crop in drought-prone areas of Africa (Latha et al., 2005). According to estimates, about 3.5 billion people were at the risk of Ca deficiency in 2011 and about 90% of these people were living in Africa and Asia (Kumssa et al., 2015). Crops such as rice and wheat can provide food security, but finger millet has nutritional properties superior to that of rice and wheat, so it has been proposed to help in strengthening the nutritional security in the developing countries of Asia and Africa (Puranik et al., 2017).

Establishment of genetic and genomic resources is a crucial step forward in improving the crop plants for specific traits. Rapid developments in the tools like Illumina sequencing in recent years have accelerated the whole genome and transcriptome sequencing in several plants (Bolger et al., 2014). As a result, the whole genome sequence (WGS) has become available for model plants and many cereals, even with more complex genomes[1]. The whole genome sequencing of finger millet has been delayed as compared to other major cereals, leaving it as one of the lastly sequenced genomes among cereals (**Figure 1**). For e.g., the first draft genome for rice was released in 2005 (International Rice Genome Sequencing, 2005) with the completion of annotation in 2013 (Kawahara et al., 2013). Foxtail millet is the only millet to have its WGS released with complete annotations till date. The WGS of 2 different foxtail millet genotypes were released in 2012 (Bennetzen et al., 2012; Zhang et al., 2012). However, the first draft genome of finger millet was released only recently (Hittalmani et al., 2017), more than a decade after the release of rice draft genome (**Figure 1**). As a result, only a few genetic and genomic studies have been performed in finger millet and the high resolution genetic and genomics studies are lagging behind due to the lack of WGS for finger millet. The recently released draft genome is expected to serve as a major resource for the accelerated studies for the improvement of finger millet in near future.

In this article, we review the details of various studies undertaken to improve the finger millet including genetic improvement, identification of quantitative trait loci (QTLs) for key traits, gene characterization and transcriptome analysis. We have collected the details on genomic resources and literature on finger millet from public database like NCBI, PubMed and major publishers' sites such as, Springer, Elsevier, etc., ranging from 1975 to till date. We discuss the past works, analyze shortfalls and provide insights on the interventions of WGS in aiding finger millet improvement in future.

FINGER MILLET GERMPLASM AND PRODUCTION

More than 28,041 finger millet germplasms are available in various organizations worldwide. Of this, the National Bureau of Plant Genetic Resources (NBPGR), India, has 10,507 germplasm and the International Crop Research Institute for the Semi-Arid Tropics (ICRISAT), India has 5957 germplasms. Other institutes like Kenya Agricultural Research Institute (KARI), Kenya (2875), Institute of Biodiversity Conservation (IBC), Ethiopia (2156), USDA Agricultural Research Service (USDA-ARS), United States (1452) and Serere Agricultural and Animal Production Research Institute (SAARI), Uganda have a reasonable collection of germplasm (Dwivedi et al., 2012; Goron and Raizada, 2015; Saha et al., 2016; Gupta et al., 2017).

Finger millet is majorly grown in the semi-arid tropics of Asia and Africa. In Asia, finger millet is mostly grown in the Southern states of India which provide favorable growth conditions (**Figure 2A**). Among the millets, finger millet ranks fourth on a global scale of production next to sorghum, pearl millet (*Cenchrus americanus*), and foxtail millet (*Setaria italica*) (Upadhyaya et al., 2007). Around 4.5 million tons of finger millet are produced worldwide every year. Africa produce 2.5 million

[1]https://phytozome.jgi.doe.gov/pz/portal.html

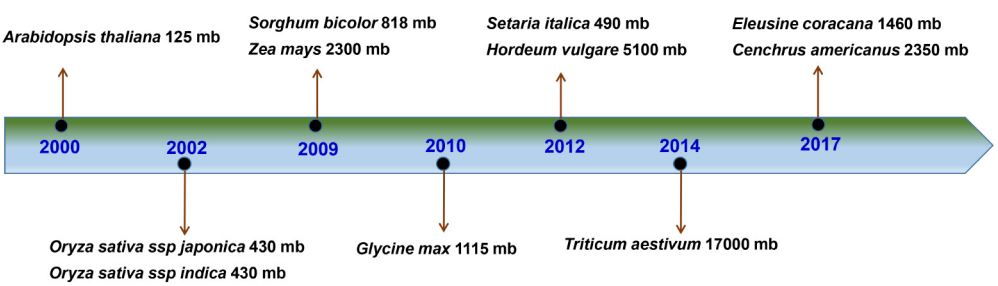

FIGURE 1 | Milestones in the genome sequencing of cereals. The scale shows the year of release of genome sequence for key cereal. The scientific name and genome size are indicated for each cereal. Source: www.plabipd.de

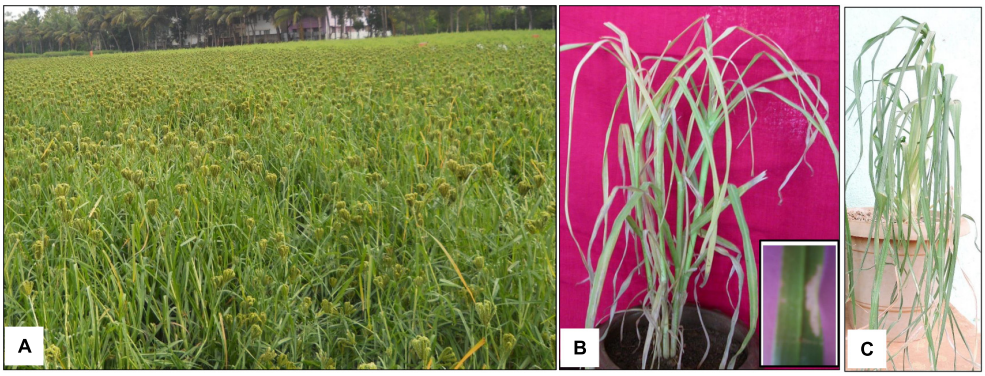

FIGURE 2 | Finger millet growth and stresses. **(A)** Photograph showing the cultivation of finger millet in Coimbatore district in Tamil Nadu state of South India; **(B)** 2 months old finger millet affected by leaf blast disease in glass house condition, the insert image shows the magnification of leaf infection by the fungus; **(C)** 2 months old finger millet affected by drought stress. Images **(B,C)** were obtained from the experiments conducted by SAC and LS.

tons and India produces 1.2 million tons annually. Finger millet accounts for about 85% all millets produced in India and is cultivated over 1.19 million hectares in India according to a recent report (Sakamma et al., 2018).

As the increase in population and industrialization throughout the world reduced the availability of agricultural land, by the end of 2050, the world is expected to face a severe food demand (Gupta et al., 2017). To overcome such a situation, there is an urgent need to increase the production of cereals like finger millet, which has to be increased up to 4.5 t ha^{-1} by 2025 (Borlaug, 2002). Finger millet will be an ideal crop for climate resilient agriculture due to its adaptation in semi-arid tropics which are characterized by unpredicted weather and erratic rainfall. So it will be a good cereal for harsh climate due to global warming. Increasing the finger millet production will make this high nutritional food available for the poor people of developing nations and will help to attain nutritional security.

CONSTRAINTS OF FINGER MILLET PRODUCTION

Finger millet production is severely affected by both biotic and abiotic stresses (Saha et al., 2016) (**Figure 2B**). Fungal blast is a major disease affecting growth and yield of finger millet (Kumar and Kumar, 2011). Blast diseases are caused by an ascomycete fungus *Magnaporthe oryzae* (anamorph: *Pyricularia grisea*) (Singh and Kumar, 2010). The fungus mostly infects young leaf and causes leaf blast, whereas under highly favorable conditions, neck and finger blasts are also formed at flowering (Babu et al., 2013). Ekwamu (1991) reported that the head blast significantly reduced the spikelet length, grain weight, number of grains per head and grain yield. The blast fungus enters and causes the breakdown of parenchymatous, sclerenchymatous, and vascular tissues of the neck region, thereby inhibiting the flow of nutrients into the grains (Rath and Mishra, 1975). Subsequently, grain formation is partially or totally inhibited (Rath and Mishra, 1975; Ekwamu, 1991). The infected spikelets were shorter than healthy spikelets, which affects the grain formation. Eventually, the high seed infection reduced the seed germination in the field (Gashaw et al., 2014). The average loss owing to the blast has been reported to be around 28–36% per hectare (Nagaraja et al., 2007) and according to an earlier study, the yield losses could be as high as 80–90% per hectare (Rao, 1990).

Major abiotic stresses such as deficiencies of nutrients [nitrogen (N), phosphorus (P), and zinc (Zn)], drought, and salinity also seem to affect the growth and yield of finger millet (Yamunarani et al., 2016; Ramakrishnan et al., 2017; Maharajan et al., 2018). According to a recent study, N deficiency decreased the tiller number in finger millet (Goron et al., 2015). Low P stress also affected the growth and biomass of finger millet seedlings in glass house conditions (Ramakrishnan et al., 2017). Zn deficiency resulted in stunted growth, delayed seed maturity, appearance of chlorosis, shortened internodes and petioles, and malformed leaves (Yamunarani et al., 2016). Drought is also one of the major abiotic constraints of finger millet production (**Figure 2C**). Parvathi et al. (2013) studied the effect of drought stress on the expression of candidate genes in genotype GPU-28. Drought stress caused wilting and leaf rolling and resulted in the reduction of leaf solute potential and chlorophyll content with the induction of many drought stress responsive genes when compared to control condition (Parvathi et al., 2013). Salinity also reduced the water content, plant height, leaf expansion, finger length and width, grain weight, and delayed the flowering (Anjaneyulu et al., 2014). Seedlings of finger millet genotype GPU-28 exposed to salinity stress, PEG and oxidative stress showed significant reduction in plant growth and shoot and root biomass (Parvathi and Nataraja, 2017).

Nutrient deficiency may be one of the major abiotic stresses affecting the finger millet production in the future. For example, the demand for fertilizers like N is expected to rise steadily, during the forecast period, from 8.8% in 2017 and reaching 9.5% in 2018 (Food and Agriculture Organization of the United States [FAO], 2015). In 2018, the global potential balance of P fertilizer is expected to rise from 6.4 to 8.5% of total demand (Khabarov and Obersteiner, 2017). Developing plants with improved P-use efficiency has been considered as essential to reduce the P

fertilizer usage (Baker et al., 2015; Ceasar, 2018). Based on the Food and Agriculture Organization (FAO) analysis, N and P demands may also affect the production of finger millet in future. This is an important issue since crops like finger millet are majorly grown by resource poor farmers in low input agricultural systems of Asia and Africa who cannot afford to buy expensive fertilizers (Thilakarathna and Raizada, 2015). Breeding of finger millet with genetic and genomic studies aided by recently released WGS may be helpful to develop new genotypes that are tolerant to multiple nutrient stresses.

CURRENT STATUS OF GENOMIC RESOURCES AVAILABLE FOR FINGER MILLET

The genomic resources available for finger millet are limited as compared with other major cereals which hampers the further improvement of this crop (Saha et al., 2016). The details of various genomic resources available for finger millet, rice, barley and maize at NCBI are listed in **Table 1**. For e.g., only a few expressed sequence tags (ESTs) are available in finger millet compared to those of rice, maize and barley. The finger millet has only 1934 ESTs which is almost 100 times lower than that of maize and rice and 50 times lower than that of barley (**Table 1**). No complete gene and Unigene sequence has yet been reported for finger millet. Several genome assemblies are available for other cereals as compared to just only one for finger millet (ASM218045v1). Similarly, limited number of proteins were reported for finger millet when compared to 3 other major cereals (**Table 1**). Till date, no single nucleotide polymorphism (SNP) has been developed in finger millet genome. The recently released WGS of finger millet will be helpful to build all these resources in the coming years to accelerate finger millet research at all spheres of studies (**Figure 3**). Finger millet also has a limited number of transcriptome sequences obtained from a few stress conditions and for grain Ca content (**Table 2**). Few efforts were made to sequence the transcriptome of specific genotypes subjected to various stresses like drought, saline and blast. However, the validation of sequence reads information and further characterization of key genes have not yet been accomplished in most of these studies and have simply been submitted as raw reads (**Table 2**). The recently released WGS of finger millet is expected to serve as a major resource for making several of these resources and for further studies. For e.g., the RNAseq reads can be validated using the WGS of finger millet to find key genes involved in each process (**Figure 3**).

WHOLE GENOME SEQUENCE

The much awaited WGS of finger millet genotype ML-365 (a drought tolerant and blast disease resistant genotype with good cooking qualities) was obtained recently using Illumina and Sequencing by Oligonucleotide Ligation and Detection SOLiD sequencing technologies (Hittalmani et al., 2017). Around 45 Gb paired end and 21 Gb mate-pair data were generated. The genome assembly consisted of 525,759 scaffolds (>200 bp) with N50 length of 23.73 Kb, and the average scaffold length of 2275 bp (Hittalmani et al., 2017). The transcriptome was also successfully sequenced and assembled in this study for well-watered (WW) (53,300 unigenes) and low moisture stressed (LMS) (100,046 unigenes) plants of genotype ML-365. Among the unigenes assembled, nearly 64% were functionally annotated with Viridiplantae protein sequences using UniProt database. The differential gene expression analysis revealed that 2,267 unigenes were specific to WW, 12,893 were specific to LMS and 111,096 unigenes were found in both WW and LMS conditions. Further, protein domain analysis predicted several functional proteins in the expressed genes. Plant transcription factors (TFs) were mined by protein-protein homology modeling and a total of 11,125 genes were predicted to have homology with 56 TF families. Overall, 2866 drought responsive genes were associated with major TF families across 19 Pfam domains. About 1766 genes were identified as R-genes for various diseases and 330 genes were found to be involved in calcium transport and accumulation (Hittalmani et al., 2017). The WGS of finger millet was found to have greater co-linearity with foxtail millet and rice as associated to other Poaceae species. This study also revealed that the genome sizes of *E. coracana* subspecies *coracana* and *E. coracana* subspecies *africana* were relatively similar (Hittalmani et al., 2017). This may be due to the fact

TABLE 1 | Details on genomic and proteomic resources available for finger millet, rice, maize, and barley.

Name of the sequence/resource	Finger millet (*Eleusine coracana*)	Rice (*Oryza sativa*)	Maize (*Zea mays*)	Barley (*Hordeum vulgare*)
EST	1,934	1,281,057	2,023,541	840,300
Gene	0	97,446	78,018	707
Unigene	0	74,892	61,577	20,224
Genome assembly	01	26	17	09
Clone	0	172,025	1,145,013	0
Nucleotide	1,095	771,335	1,059,632	3,536,399
SNP	0	13,218,961	58,915,360	0
Protein	554	1,324,842	332,077	69,529
Protein cluster	0	15,559	94	77
Protein structure	3	210	330	142

Source, NCBI; Date of collection, 5th March 2018.

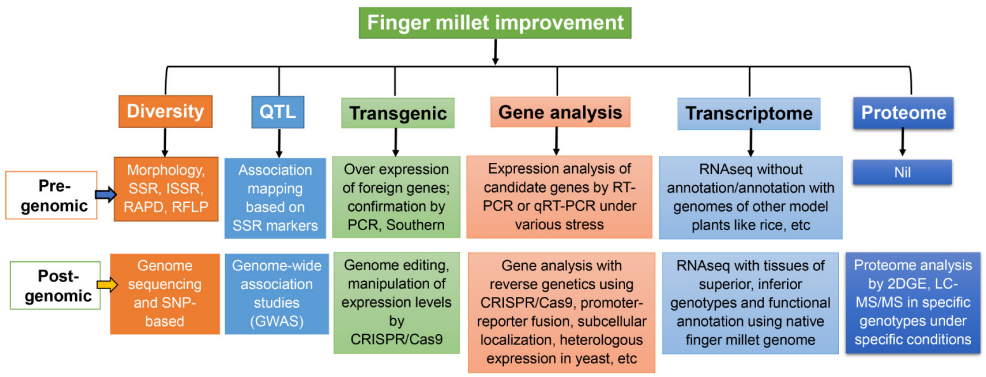

FIGURE 3 | Finger millet improvement in the pre-genomic and post genomic eras. Various approaches used for the improvement of finger millet like, diversity analysis, QTL mapping, transgenic modification, gene analysis are indicated. The use of advance tools for the improvement of finger millet in post-genomic era based on the whole genome sequence are indicated.

TABLE 2 | Details on genome and transcriptome sequences reported for finger millet under various experimental conditions.

S. No	Name of the genotype	Type of sequence	Property/trait	NICBI Accession no.
1	PR-202	Genome assembly	Drought stress	PRJDB5606
2	**	Metagenome	Blast disease	PRJNA383952
3	KNE796	Whole genome and transcriptome	Crop improvement	PRJNA377606
4	ML-365	Transcriptome	Moisture stressed	PRJNA339512
5	ML-365	Whole genome	Drought stress	PRJNA318349
6	KNE796		High throughput marker development	PRJNA317618
7	GPU-28	Transcriptome	Drought stress	PRJNA282860
8	MR-1	smallRNA analysis	Drought stress	PRJNA277250
9	CO 12 and Trichy 1	Transcriptome	Salinity stress	PRJNA236733
10	**	Transcriptome	Water deficit	PRJNA229808
11	GPU-28	Transcriptome	Drought stress	PRJNA282859
12	**	Transcriptome	Drought stress	PRJNA282578
13	**	Transcriptome	Blast disease	PRJNA268401
14	GPU-1 and GPU-45	Transcriptome	Calcium content	PRJNA236796

*Source, NCBI; Date of collection, 5th March 2018. **Details not provided.*

that finger millet was domesticated from *E. coracana* subspecies *africana* (Dida et al., 2008).

Hatakeyama et al. (2018) reported the WGS and assembly of finger millet genotype PR-202 (IC: 479099) using a novel polyploidy genome assembly workflow. Their initial analysis identified the genome size of finger millet as 1.5 Gb and the assembled genome was 1189 Mb which was estimated to cover 78.2% genome. The whole genome of genotype PR-202 consisted of 2387 scaffolds with the N50 value of 905.318 Kb having maximum sequence length of 5 Mb. The FASTA file format of final scaffolds and annotation are publicly available at NCBI (BioSample number: SAMD00076255). Overall, 62,348 genes were identified by this study, nearly 91% genes were functionally annotated and 96.5% were found to be single-copy genes. The NCBI BLAST analysis identified that a total of 57,913 genes was duplicated with more than two copies in the genome of PR-202 (Hatakeyama et al., 2018).

The availability of WGSs of ML-365 and PR-202 can be used effectively for further studies, such as SNP identification, next-generation sequencing (NGS)-based allele discovery, linkage and association map construction, identification of candidate genes for agronomically important traits, functional characterization of candidate genes using reverse genetic approaches and marker-assisted breeding programs (**Figure 3**).

IN VITRO STUDIES: A PREREQUISITE FOR GENETIC TRANSFORMATION

Establishment of an efficient *in vitro* regeneration protocol is a vital prerequisite for the transformation and regeneration of cereals (Shrawat and Lörz, 2006). *In vtiro* culture has been considered essential for finger millet improvement (Yemets et al., 2013). Several reports are available for the *in vitro* regneration of finger millet using various explants in different genotypes (Supplementary Table S1). The types of explants used include shoot tip (Eapen and George, 1990; Ceasar and Ignacimuthu, 2008, 2011), leaf sheath fragments (Eapen and George, 1990; Gupta et al., 2001), embryogenic seed (Kothari et al., 2004; Latha et al., 2005; Sharma et al., 2011; Babu et al., 2012),

mature and immature embryos (Kumar et al., 2011), undeveloped inflorescence (Eapen and George, 1990; Kumar et al., 2011), root mesocotyl (Mohanty et al., 1985), and leaf-base segments (Rangan, 1976; Mohanty et al., 1985) (Supplementary Table S1). Satish et al. (2015) made an attempt to regenerate finger millet through direct organogenesis using shoot apical meristem. The same group also developed an efficient *in vitro* regeneration protocol for indirect organogenesis in four Indian genotypes (CO (Ra)-14, GPU-25, Try-1, and Piyur-2) using plant growth regulators and polyamine compounds like spermidine (Satish et al., 2016b). Use of seaweed liquid extracts seem to promote the somatic embryogenesis and regeneration in the same genotypes of finger millet (Satish et al., 2016a). Recently, yet another direct plant regeneration protocol was developed in three genotypes [CO 9, CO (Ra)-14 and GPU-28] of finger millet (Babu et al., 2018).

Shoot apex explant is an ideal material for efficient *in vitro* regeneration owing to its easy availability, accessibility, rapid regeneration of multiple shoots, and easier to handle when compared with other explants (Arockiasamy and Ignacimuthu, 2007; Ceasar and Ignacimuthu, 2008; Dey et al., 2012). Shoot apex was used in the past for finger millet regeneration. The direct plant organogenesis is also an effective method to produce more multiple shoots with less somoclonal variation in a short time as it minimizes the culture duration for callus formation, sub-culturing cycles and quicker regeneration of transgenic plants following transformation (Satish et al., 2015). There is no report available till date for *in vitro* regeneration through anther culture, protoplast and protoplasmic fusion in finger millet. Development of anther culture in finger millet could help to develop the haploid lines. The development of protoplasmic fusion may also help to improve the hybrid variety of finger millet. The WGS may also be utilized for clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated protein 9 (Cas9)-mediated genome editing in finger millet for which protoplast mediated transformation looks very effective. So the establishment of protoplast based regeneration in finger millet may be helpful to achieve these tasks taking finger millet research into next and higher level.

GENETIC IMPROVEMENT OF FINGER MILLET

Genetic improvement of finger millet has been lagging behind when compared to the efforts made for other major cereals. Improved genetic transformation of millets including finger millet has been considered essential to improve the nutritional quality, and resistance to abiotic and biotic stresses (Ceasar and Ignacimuthu, 2009). Gupta et al. (2001) initiated the preliminary work on transformation of finger millet using biolistic method for comparing the efficiency of five gene promoters [*cauliflower mosaic virus 35s* (*CaMV35S*)*/rice actin gene promoter ActI/maize ubiquitin (UqI)/ribulose-1,5-biohosphate carboxylase small subunit gene promoter(RbcS)/Flaveria trinervia* β*-glucuronidase* (*FtuidA*) on the expression of the β*-glucuronidase* (*GUS*)] reporter gene. Following this, a few studies reported on the optimization of transformation conditions for efficient transformation and regeneration and most of these studies employed *Agrobacterium*-mediated transformation procedure (**Table 3**). The general schematic protocol used for the *Agrobacterium*-mediated transformation of finger millet is presented in **Figure 4**. Only a limited number of studies were reported on the transformation of finger millet using a functionally active transgene. The details are discussed below.

Genetic Improvement for Blast Resistance

A transgenic finger millet resistant to leaf blast disease was developed using *antifungal protein* (*PIN*) gene of prawn (Latha et al., 2005). The *PIN* gene was chemically synthesized and cloned into plasmid *pPin35S* under the control of *CaMV35S* promoter and transformed by biolistic method. Similarly, we have introduced a rice *Chitinase11* gene (*Chi11*) into genotype GPU45 of finger millet through *Agrobacterium*-mediated transformation to develop leaf blast resistance (Ignacimuthu and Ceasar, 2012). These two initial studies helped to develop the transgenic finger millet resistant to leaf blast disease. In both these reports, the transgenic plants overexpressing foreign gene exhibited resistance to leaf blast disease compared to non-transformed control plants. However, there are no reports available on transgenic finger millet resistant to neck and finger blasts. So screening of many other potential antifungal genes and gene pyramiding will be helpful to develop transgenic finger millet resistant to a wide spectrum of fungal diseases.

Genetic Improvement for Abiotic Stress Tolerance

A salt-tolerant finger millet was developed using *sorghum vacuolar* H^+*-pyrophosphatase* (*SbVPPase*) gene by *Agrobacterium*-mediated transformation. Overexpression of *SbVPPase* gene in finger millet enhanced the growth performance under salt stress. Jayasudha et al. (2014) also produced a transgenic finger millet by introducing Na^+/H^+ *antiporter* of *Pennisetum glaucum (PgNHX1)* and *Arabidopsis thaliana vacuolar H+-pyrophosphatase (AVP1)* for salinity stress tolerance through *Agrobacterium*-mediated transformation. The transgenic finger millet showed a higher level of salinity tolerance compared to wild type plants. *Porteresia coarctata*'s *serine-rich protein* (*PcSrp*) gene was overexpressed in finger millet under salinity condition (Mahalakshmi et al., 2006). The transgenic finger millet grown under 250 mM NaCl stress condition showed normal growth, flower and seed set rescuing from the saline stress (Mahalakshmi et al., 2006). Transgenic finger millet expressing a bacterial *mannitol-1-phosphate dehydrogenase* (*mtlD*) gene was developed through *Agrobacterium*-mediated transformation (Hema et al., 2014). Transgenic finger millet plants expressing *mtlD* gene had better growth under drought and salinity stress compared to wild-type. The transgenic plants also showed better osmotic stress tolerance with chlorophyll retention under drought stress compared to the wild-type plants (Hema et al., 2014).

TABLE 3 | Details of various transformation studies reported in finger millet.

Name of the genotype	Promoter/ reporter gene used	Promoter/selectable marker	Functional gene used	Methods of transformation	Type of explants used	Application	References
PR202	*CaMV35S/Actl/Uql/ RbcS/Ft GUS*	Nil	Nil	Biolistic	Leaf sheath segments	Testing the efficiency of various promoters in *GUS* expression	Gupta et al., 2001
PR202	*CaMV35S GUS*	*CaMV35S nptII*	Nil	*Agrobacterium*-mediated	Embryogenic seed	Establishment of transformation efficiency under different parameters	Sharma et al., 2011
**	*CaMV35S GUS*	*CaMV35S bar*	*Antifungal protein (PIN)* gene of prawn	Biolistic	Shoot apex	Transgenics resistant to leaf blast disease	Latha et al., 2005
GPU45 and CO14	*CaMV35S GUS*	*CaMV35S hptII*	Nil	*Agrobacterium*-mediated	Shoot apex	Optimization of transformation using shoot apex	Ceasar and Ignacimuthu, 2011
GPU45 and CO14	*Uql GUS*	*Uql hptII*	Rice *chitinase* gene	*Agrobacterium*-mediated	Shoot apex	Transgenics resistant to leaf blast disease	Ignacimuthu and Ceasar, 2012
PR202	*CaMV35S GUS*	*CaMV35S hptII*	Nil	Biolistic	Green nodular calli	Optimization of biolistic mediated transformation protocol	Jagga-Chugh et al., 2012
Tropikanka and Yaroslav8	Nil	*CaMV35S bar*	*HvTUB1* and *TUAm1*	Biolistic and *Agrobacterium*-mediated	Embryogenic callus	Resistance to herbicides of the dinitroaniline family	Bayer et al., 2014
GPU28	*CaMV35S GUS*	*CaMV35S hptII*	Bacterial *mannitol-1-phosphate dehydrogenase* gene	*Agrobacterium*-mediated	Embryogenic callus	Tolerance to drought and salinity	Hema et al., 2014
GPU28	*CaMV35S GUS*	*CaMV35S hptII*	*PgNHX1, AVP1*	*Agrobacterium*-mediated	Embryogenic callus	Salinity tolerance	Jayasudha et al., 2014
CO(Ra)-14, PR-202, Try-1 and Paiyur2	*CaMV35S GUS*	*CaMV35S hptII*	Nil	*Agrobacterium*-mediated	Shoot apex	Optimization of transformation using direct plant regeneration	Satish et al., 2017

*Name of the marker and reporter genes with promoter, genotype, explant type and method of transformation used are given for each study. Any functional gene used is also indicated. AVP1, Arabidopsis vacuolar pyrophosphatase1; bar, phosphinothricin resistance; CaM35S, cauliflower mosaic virus 35s; Ft, promoter of C4 isoform of phosphoenolpyruate carboxylase gene from Flaveriatrinervia; gus or uidA, β-glucuronidase; nos, nopaline synthase; hptII, hygromycin phosphotransferase; HvTUB1, Hordeum vulgare β1-tubulin; nptII, neomycin phosphotransferase; PgNHX1, Pennisetum glaucum Sodium hydrogen exchanger1; Uql, Ubiquitin promoter of maize); RbcS, ribulose-1,5-bisphosphate carboxylase small subunit gene promoter; TUAm1, α 1_tubulin mutant gene.**PGEC-2, IE-2576, IE-2367, IE-2366, IE-2683, IE-2684, IE-2851, IE-2861, IE-2333, IE-2995, IE-2300, IE-2675, IE-2340, IE-2983, IE-3242, IE-3020, IE-4673, IE-4683, IE-4120.*

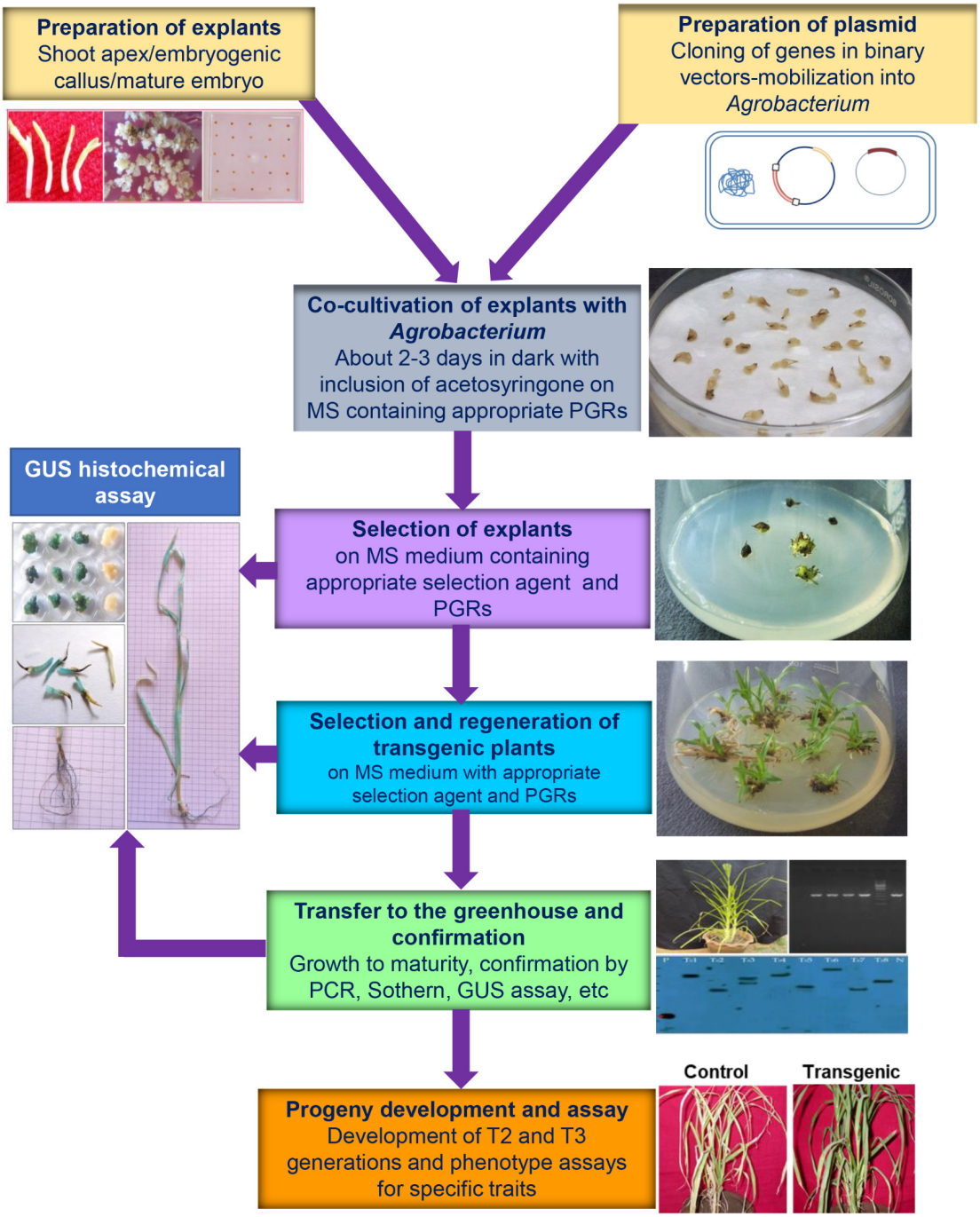

FIGURE 4 | General protocol used in the *Agrobacterium*-mediated transformation of finger millet. The stepwise protocol used in the *Agrobacterium*-mediated transformation is illustrated with respective figures. The photographs were obtained from the works performed in the labs of SI, SAC, and LS. The bio-assay photograph was obtained in the transgenic finger millet resistant to leaf blast disease and control plants by SAC.

It is evident that only a limited number of reports are available on overexpression of transgenes conferring tolerance to blast and other abiotic stresses in finger millet. More foreign genes need to be screened by overexpression for developing varieties resistant to multiple stresses. Most of these studies also focused on the introduction of foreign genes and phenotyping under a specific stress. High resolution studies, like subcellular localization of foreign gene and fusion of promoters of finger millet with reporter genes [GUS, green fluorescent protein (GFP), etc.] are yet to be performed in finger millet. The recently released WGS will be helpful to design such studies, especially those focusing on isolation of native promoters for functional analysis by fusing

them with reporter genes. This will help to perform studies in line with those performed in model plants like rice and *A. thaliana* for functional validation of key genes and their promoters which will be useful to identify key genes and signals involved in grain filling of nutrients, drought tolerance, fungal resistance, etc.

MOLECULAR MARKER-ASSISTED BREEDING

Molecular markers are one of the important tools employed for the identification and improvement of particular traits. The DNA-based markers provide foundation for a wide range of molecular marker techniques, which are being widely used in the crop breeding program (Babu et al., 2007). Plant breeding backed by molecular markers helps to track traits more precisely when compared to conventional breeding. Several reports are available for the analysis of genetic diversity and QTL in finger millet using molecular markers which are discussed below.

Genetic Diversity Analysis

The analysis of genetic diversity is crucial for crop improvement as it reveals the details of genetic relationships and provides insights for sampling of breeding populations (Mohammadi and Prasanna, 2003). Genetic diversity analysis helps to understand the relationships of genotypes around the world at genetic level and will aid in the selection of suitable genotypes for breeding programs (Babu et al., 2017). As finger millet is cultivated under diverse climatic conditions in Asia and Africa, analysis of genetic diversity helps to understand the genome variation between genotypes and subsequent population development for molecular marker analysis. The genotypes that are adapted to various biotic and abiotic stresses have more allele variation compared to susceptible genotypes. The genotypes having greater allele variation are being used for breeding programs. Randomly amplified polymorphic DNA (RAPD), restriction fragment length polymorphism (RFLP), and simple sequence repeats (SSR) markers were frequently used for the analysis of genetic diversity in finger millet (Parani et al., 2001; Fakrudin et al., 2004; Babu et al., 2014a) (Supplementary Table S2). Gupta et al. (2010) analyzed three genotypes of finger millet with variable seed coat color (brown, white, and golden) by studying morphological, physiological, and biochemical characteristics using 10 RAPD and 10 inter simple sequence repeats (ISSR) markers (Gupta et al., 2010). RAPD markers showed better polymorphism than ISSR markers. To investigate the genetic diversity of 32 finger millet genotypes, 45 RAPD primers were used (Patil and Kale, 2013). Out of 45 primers, 25 primers showed polymorphism and maximum genetic diversity was identified in VL149, KOPN 161, 338, and 929. The genetic diversity and population structure were assessed in 128 genotypes of finger millet collected from various geographical regions using 25 RAPD markers (Ramakrishnan et al., 2016b). Following this, genetic variation and population structure and relationship were evaluated between the Indian and non-Indian genotypes using 72 genomic SSR primers (Ramakrishnan et al., 2016a). Molecular variance and population structure in 42 genotypes of finger millet collected from different geographical regions of southern India were analyzed using 10 RAPD, 9 ISSR, and 36 SSR markers (Rajendran et al., 2016). These genotypes with diversity information can be used as parents of interest and may be crossed with elite material to develop new breeding population.

Although the PCR based markers were successfully used in the past for genetic diversity analysis, they have some limitations as the development of precise primers is very difficult (Arif et al., 2010). Alternatively, SNP based diversity analysis has been used in recent years for several plants for high throughput analysis of genetic diversity (Jeong et al., 2013; Ren et al., 2013; Tang et al., 2016). With no surprise, such studies have not yet been attempted in finger millet. Hopefully, the recently released WGS will aid in the development of SNP based diversity analysis in finger millet accessions (**Figure 3**). This will help to choose the genotypes for breeding and marker development based on SNP.

Identification of QTLs for Agronomical Traits

The microsatellite markers have been used to identify the agronomically important traits in finger millet such as grain yield, disease resistance, drought resistance and nutritional quality (**Figure 5**). Association mapping based identification of QTLs related to nutritional traits are thereby helpful in bio-fortification programs for ameliorating nutritional deficiencies (Kumar et al., 2015). For e.g., a total of 9 QTLs associated with Ca content were identified in 113 genotypes of finger millet using 23 anchored SSR markers (Kumar et al., 2015). Hence, determination of QTLs controlling these traits along with the candidate genes that cause deviation in Ca accumulation are essential for the successful incorporation into breeding and transgenic strategies. In another study, 46 genomic SSR markers were used to identify 4 agro-morphological traits such as basal tiller number, days to 50% flowering, flag leaf blade width, and plant height in 190 finger millet genotypes (Babu et al., 2014a) (**Figure 5**). The same group also identified four QTLs (UGEP81, UGEP24, FMBLEST32, and RM262) in the same genotypes of finger millet using 104 SSR markers (Babu et al., 2014c). In the same year, two QTLs (OM5 and FM8) were identified for tryptophan content and one QTL (FMO2EST1) for protein content in the aforementioned genotypes of finger millet using 120 SSR markers and these QTLs were linked to *opaque2 modifiers* (*Opm*) gene (Babu et al., 2014b). Tryptophan and lysine are the amino acids used in the biosynthesis of proteins. In cereal endosperm, these amino acids are deficient because it generally contains 1.5–2% lysine and 0.25–0.5% tryptophan, whereas 5% lysine and 1.1% tryptophan are required for optimal human nutrition. Finger millet contains high amount of tryptophan compared to other cereals. In view of this, identification of the QTLs linked to *Opm* gene responsible for the tryptophan content could be a major target for further improvement of the quality of finger millet germplasm. Further, 7 QTLs were found to be associated with seven agronomic traits, including productive tillers, seed yield, leaf blast resistance and number of tillers by 87 genomic SSR markers in 128 genotypes of finger millet (Ramakrishnan et al., 2016c). Recently, four QTLs (qLRDW.1, qLRDW.2, qHSDW.1, and qHRL.1) associated

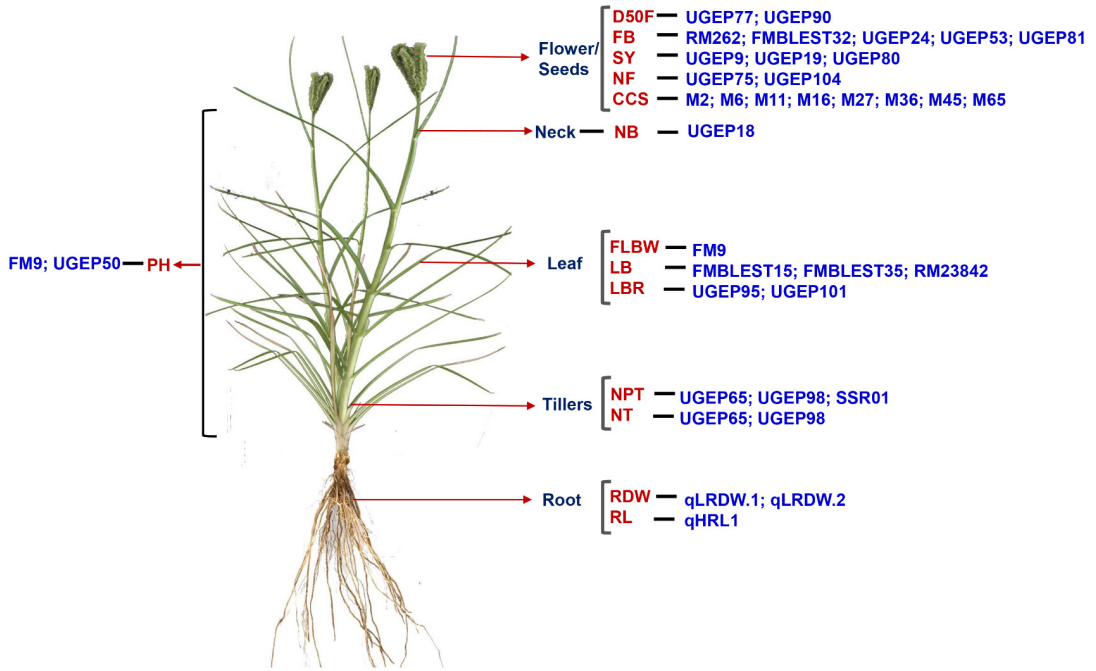

FIGURE 5 | Identification of QTLs for important agronomical traits in finger millet. QTLs identified for various traits in different tissues are indicated with a pictorial representation. CCS, calcium content in seed; D50F, days to flowering (50%); FB, finger blast; FLBW, Flag leaf blade length; LB, leaf blast; LBR, leaf blast resistant; NB, neck blast; NF, number of fingers; NPT, number of productive tillers; NT, number of tillers; PH, plant height; RDW, root dry weight; RL, root length; SY, seed yield.

with root dry weight, shoot dry weight, and root length were identified in finger millet by association mapping under P deficient and P sufficient conditions (Ramakrishnan et al., 2017). In seedling stage, shoot and root growths were severely affected by P deficiency. Hence, the P deficiency tolerance in the seedling stage is an essential trait that needs to be used in finger millet cultivars (Ramakrishnan et al., 2017). This study provides input to breed low P-tolerant genotypes in finger millet using marker-assisted selection, and selected germplasm lines can be used either as cultivars for marginal lands where P deficiency is prominent or as donors for P starvation tolerance QTLs for breeding.

In finger millet, only the association mapping of populations was used so far for QTL studies. There is a crucial need to develop the linkage maps of finger millet for the identification of QTL, since it will play a major role in identifying the agronomically important traits. Moreover, the high throughput QTL mappings have not yet been attempted in finger millet due to lack of the WGS. Genome-wide association study (GWAS) has been emerging as a powerful tool in the identification of QTLs based on WGS. Several QTLs have been identified based on GWAS in cereals like rice (Huang et al., 2010, 2011, 2016; Zhang et al., 2018), barley (Visioni et al., 2013; Matthies et al., 2014; Fan et al., 2016; Bellucci et al., 2017) and maize (Mahuku et al., 2016; Wang et al., 2016). Development of low cost and high throughput genome sequencing technologies together with the availability of WGS will aid in the development of GWAS in finger millet in the coming years (**Figure 3**). This will be highly beneficial for dissecting QTLs and associated SNPs more precisely for key traits of finger millet including grain Ca content.

FUNCTIONAL CHARACTERIZATION OF KEY GENES

Functional characterization of genes with key traits has been considered essential for developing varieties with improved traits. Only preliminary attempts have been made in finger millet for such studies (Supplementary Table S3). The recent developments in genomic research are expected to play a key role in the identification and characterization of candidate genes involved in nutrient signaling and transport in finger millet (Sood et al., 2016). As finger millet has 10-fold higher Ca in seeds compared to other cereals, dissection of key genes and signals involved in grain Ca filling will be important for nutrient enrichment of other cereals. The preliminary studies reporting the identification and expression analysis of candidate genes in finger millet are discussed below.

Genes Involved in Ca Transport

Ca is a vital macronutrient for growth and development of plants as well as humans and animals. Ca is the third most important nutrient available in the soil and is required for normal growth of plants. In finger millet, maximum Ca is present in aleurone layer, followed by seed coat and embryo (Nath et al., 2013). Elevated level of Ca is also associated with higher expression of Ca-signaling transporter genes (Carter et al., 2004). Although there is

no active transpiration stream within cells of the mature embryo, nutrient transfer between maternal and filial tissues is restricted to the apoplast (Patrick and Offler, 2001); therefore, changes in apoplastic Ca levels of the maternal plant could be reflected in the mature embryo or seed coat, which may be governed by Ca^{2+} transporter genes. So characterization of key genes involved in Ca accumulation will be helpful in transferring the same trait to other millets and non-millet cereals. Expression levels of key genes involved in Ca transport such as *Ca2*$^{+}$*/H*$^{+}$ *antiporter* (*CAX1*), *two pore channel1* (*TPC1*), *calmodulin (CaM)-stimulated type IIB Ca2+ ATPase* and *two CaM dependent protein kinase* (*CaMK1* and *CaMK2*) have been analyzed in 2 finger millet genotypes of contrasting Ca traits (GP-1, low Ca and GP-45, high Ca) (Mirza et al., 2014). The same group also identified 82 Ca sensor *genes* from the transcriptome of developing spikes of both genotypes GP-1 and GP-45 (Singh U.M. et al., 2014). As an outcome of this, the expression of 24 genes was higher in the pooled spike sample of genotype GP-45 while the expression of 11 genes was higher in the pooled spike sample of genotype GP-1. Twenty-four genes were highly expressed in the developing spikes of GP-45, seven encoded for *CaML*, two for *CRK*, five for *CBL*, seven for *CIPK*, and four for *CDPK* genes. Another report in the following year by the same group reported the characterization of *Ca*$^{2+}$ *transporter* gene family in these two genotypes of finger millet. Whole genome and transcriptome profiling was also performed in the developing spikes of finger millet to find key genes involved in Ca^{2+} transport (Singh U.M. et al., 2015). More recently, *CIPk24* gene was also characterized in these two genotypes of finger millet (Chinchole et al., 2017). This gene was overexpressed in root, shoot, leaf and developing spike tissues of GP-45 compared to GP1. Nine SNPs and one extra beta sheet domain as well as differences in vacuolar localization were identified through *in silico* analyses using the genomes of other model plants. Both EcCBL4 and EcCBL10 were found to show strong binding affinity with *EcCIPK24* (GP-1) compared to *EcCIPK24* (GP-45). It has been predicted that by activating EcCAX1b protein, EcCIPK24 can play an important role in high seed Ca accumulation (Chinchole et al., 2017).

Most of these studies were performed before the release of WGS of finger millet. So the complete list of genes involved in Ca sensing and transport has been analyzed in the primary article reporting the details of finger millet genome (Hittalmani et al., 2017). With no doubt, the complete genome sequence will aid in the functional characterization of key genes involved in Ca^{2+} transport especially those mediate grain filling. High resolution studies involving reverse genetics approaches will help to dissect the complex mechanisms involved in Ca^{2+} transport in finger millet. To this end, recently popularized tools like CRISPR/Cas9 may be helpful to develop mutants with defects in key genes of Ca transport and grain filling (**Figure 3**) since this technique demands WGS to avoid any off-target effects (Ceasar et al., 2016). CRISPR/Ca9 has been successfully applied in many plants for such studies.

Genes Involved in N Metabolism

A few studies also reported the analysis of key genes involved in N transport in finger millet. Expression of *prolamin-binding factor DNA binding with one finger only* (*PBF Dof*) TF involved in regulation of seed protein storage was analyzed in different tissues like root, stem and flag leaf at vegetative stage and developing spikes of three finger millet genotypes (PRM-1, PRM-701, and PRM-801) with differing seed protein content and color (Gupta et al., 2011). The expression of this gene was relatively higher in developing spikes than in other tissues in all three genotypes. Interestingly, the grain protein content of these genotypes is directly related to higher expression of *PBF Dof* at early stages of growth (Gupta et al., 2011). Expression profile of key genes *Eleusine coracana high-affinity nitrate transporter* (*EcHNRT2*), *Ec low-affinity nitrate transporter* (*EcLNRT1*), *Ec nitrate reductase* (*EcNADH-NR*), *Ec glutamine synthetase (EcGS), Ec glutamine oxoglutarate aminotransferase* (*EcFd-GOGAT*) and *Ec DNA binding with one finger* 1 (*EcDof1*), involved in N uptake and assimilation were analyzed in two genotypes with contrasting (GE-1437, low-protein and GE-3885, high-protein) grain protein content (Gupta et al., 2013). Except *EcHNRT2*, remaining 5 genes were induced in the leaves of GE-3885 within 30 min of exposure to N deficiency. *EcNADH-NR* was found to be overexpressed in roots of GE3885 when the plans were exposed to increasing nitrate concentrations but not in GE-1437. This study revealed that GE-3885 might be a quick sensor of nitrogen compared to low-protein genotype (Gupta et al., 2013). In the following year, the same group also analyzed the expression pattern of *EcDof1* and *EcDof2* in the same genotypes (GE3885 and GE1437) (Gupta et al., 2014). *Dof1* and *Dof2* are TFs having opposite roles in regulation of genes related to C and N metabolism. The *EcDof1*/*EcDof2* ratio was higher in the roots of GE-3885 than in GE-3885 indicating higher activation of genes involved in N uptake and assimilation resulting in high grain protein accumulation (Gupta et al., 2014).

Genes Involved in Carbon (C) Metabolism

Expression analysis was performed for some of the genes involved in C metabolism, such as *chlorophyll a/b binding protein* (*Cab*), *Rubisco* (*RBCS*), *phosphoenol pyruvate carboxylase* (*PEPC*), *phosphoenol pyruvate carboxykinase* (*PEPC*-k), *malic enzyme* (*ME*), *sucrose phosphate synthase* (*SPS*), *pyruvatekinase* (*PK*), *pyruvate dikinase* (*PPDK*), *14-3-3* and *sensor protein kinase 1* (*SnRK1*) and co-expression of these genes with *Dof1,* in the same genotypes (GE-1437 and GE-3885) used in the above studies (Kanwal et al., 2014). Oscillations of expression of these genes were studied under light-dark conditions. The expression of these genes in both genotypes oscillated confirming their control by an endogenous clock. But the genes such as *Cab*, *RBCS* and *PPDK* showed no oscillations which might be due to induction by light. Expression of *Dof1* was higher in GE-3885 (higher grain protein genotype) along with other genes involved in C metabolism suggesting that *Dof1* regulates the expression of light inducible genes and controls the grain protein content in finger millet (Kanwal et al., 2014). This is the only report available on validation of genes involved in C metabolism. The WGS of finger millet will help to identify and characterize more genes in involved in C metabolism in near future.

Genes Involved in Phosphate Transport

Four *phosphate transporter 1* (*EcPT1* to *EcPT4*) genes were identified and their expression was analyzed in three genotypes (RagiKorchara, Khairna, and VHC 3611) of finger millet (Pudake et al., 2017). The expression of these genes was validated under different regimes of inorganic phosphate (Pi) and under the colonization of *arbuscular mycorrhizae fungus* (AMF). It was found that *EcPT1* transcript levels were about fivefold higher in roots and leaves under deplete Pi than control. *EcPT3* gene was induced under phosphate stress in both leaves and roots. *EcPT4* genes was found to be induced by AMF in root tissues (Pudake et al., 2017). So far, only 4 *EcPT1* genes have been identified in finger millet. But each plant seems to possess more than 10 such genes (Baker et al., 2015). Even a close relative, foxtail millet has been reported to possess 12 PT genes which have been characterized for expression pattern, P transport assay in yeast and *in planta* function by downregulation through RNAi (Ceasar et al., 2014, 2017). These 4 PT genes were identified in finger millet based on the partial transcript sequences. Recently released WGS will be helpful for the genome-wide identification and functional characterization of all PT genes of finger millet.

Genes Involved in Abiotic Stress Tolerance

Finger millet has been considered as a drought-hardy crop due to its adaptation for semi-arid tropical climate. Efforts have been made to characterize the key genes involved in drought tolerance and to utilize them for further applications. Drought stress is one of the most important abiotic factors affecting plant growth and productivity. Singh R.K. et al. (2015) made an effort to characterize the drought-responsive gene *EcDehydrin7* of finger millet by isolating and overexpressing it in tobacco. Tobacco plants overexpressing *EcDehydrin7* conferred tolerance to drought. Seven drought responsive genes (including *metallothionein*, *farnesylated protein ATFP6*, *protein phosphatase 2A*, *RISBZ4* and *farnesyl pyrophosphate synthase*) were found to be overexpressed in genotype GPU-28 under drought stress (Parvathi et al., 2013). These genes are believed to play crucial roles in drought tolerance and further characterization of these genes will help to identify any novel signals involved in drought tolerance in finger millet. A drought response regulatory gene of finger millet, *TBP Associated Factor6* (*EcTAF6*) was identified by screening cDNA library of finger millet and its expression in response to various stresses was analyzed in finger millet genotype GPU-28 (Parvathi and Nataraja, 2017). When the seedlings were exposed to NaCl, PEG and methyl viologen (oxidative stress), the normal growth was inhibited and *EcTAF6* was found to be significantly induced under these abiotic stresses when compared to the control (Parvathi and Nataraja, 2017). Drought responsive genes have also been identified and validated using drought responsive transcriptome by cDNA subtraction in finger millet (Ramegowda et al., 2017). One such potential gene, *EcGBF3* was characterized by ectopic expression in *A. thaliana*. Overexpression of *EcGBF3* in *A. thaliana* improved tolerance to osmotic, saline and drought stresses in *Atgbf3* mutant lines (Ramegowda et al., 2017). This study also indicated the difficulty in generating mutant lines in finger millet for such functional genomics studies; so it was analyzed using a model plant *A. thaliana*.

Several salinity stress responsive genes were identified in leaves of two contrasting finger millet genotypes viz., Co-12 (susceptible) and Trichy 1 (tolerant) under salinity condition through RNAseq (Rahman et al., 2014). The same group also reported that overexpression of *EcNAC67* TF in rice improved salinity and drought tolerances (Rahman et al., 2016). A stress responsive NAC gene (*EcNAC1*) was found to be highly up-regulated in response to salinity stress and was reported to be involved in tolerance against salinity and other abiotic stresses (Ramegowda et al., 2012). Two abiotic stress responsive TFs belonging to bZIP family (*EcbZIP60*) (Babitha et al., 2015a) and Basic helix-loop-helix (bHLH) family (*EcbHLH57*) (Babitha et al., 2015b) were identified in GPU-28 genotype of finger millet under drought, osmotic, salt and methyl viologen (MV) stresses. Nagarjuna et al. (2016) identified and characterized *CBL interacting protein kinase31* (*EcCIPK31*-like) gene responsible for drought tolerance in finger millet. A *TATA box Binding Protein* (TBP)-*Associated Factors* (TAFs) gene (*EcTAF6*) was identified in GPU-28 genotypes of finger millet under drought stress (Parvathi and Nataraja, 2017). More recently, a novel endoplasmic reticulum specific bZIP TF gene of finger millet (*EcbZIP17*) was isolated and overexpressed in tobacco (Ramakrishna et al., 2018). The tobacco plants overexpressing *EcbZIP17* exhibited tolerance to saline and heat stresses as compared to wild type plants.

These are the preliminary studies reported in finger millet on identification and validation of candidate genes. Unfortunately, these genes have not yet been characterized further in finger millet using reverse genetic tools as in model plants like rice and *A. thaliana*. It may be due to lack of WGS as one needs to design precise genomic targets for such studies. The recently released WGS is expected to help for such reverse genetic approaches like development of mutants using CRISPR, functional characterization by promoter reporter fusions, localization studies and heterologous expression in yeast mutants, etc. Overall, WGS of finger millet is expected to help to perform many high resolution studies to understand the function of genes involved in nutrient signaling and abiotic stress responses and could be tapped for breeding programs to develop improved finger millet.

CONCLUSION AND FUTURE PROSPECTS

Finger millet is a nutrient rich and drought hardy crop majorly cultivated and consumed by resource poor farmers in the developing countries of Asia and Africa. Only a limited number of genomic resources are available till date due to lack of WGS. Although finger millet has been considered as a climate resilient crop for the developing world, recent studies indicated that this crop is also vulnerable to drought, saline and low nutrient stresses in addition to fungal blast. Only a limited number of

studies has been performed on characterization of functionally important genes of finger millet, before the release of WGS. WGS of two different finger millet genotypes were released recently (Hittalmani et al., 2017; Hatakeyama et al., 2018). This will help to design many high resoultion studies like those performed in other model plants such as rice and *A. thaliana* and WGS may change the course of finger millet research in future. The new genomic resource is expected to enrich the finger millet research in many spheres including dissection of key traits involved in nutrient enrichment and dorught tolerance using GWAS, genetic diversity analysis based on SNP, characterization of genes by reverse genetic studies using precise mutants using genome editing techniques like CRISPR/Cas9, accelerated functional gemomics studies such as promoter fusion of key genes with reporters like GFP for localication and spatial expression analysis, tissues specific transcriptome analysis to identify key regulatory genes of nutriant signaling and high thoroughput proteomics research to idenify the proteins associated with key agronomical functions. Overall, the recently released WGS of finger millet is expected to augment the finger millet research for its breeding and improvement. Many genes and proteins involved in the transport of key nutrients viz. Ca, P, N can be characterized in finger millet with the help of WGS. This will help to understand the key genes and regulatory networks involved in nutrient transport and can be harnessed for nutrient enrichment of other millets and non-millet cereals which will help to conserve nutrient security of growing world population.

AUTHOR CONTRIBUTIONS

SAC conceptualized the manuscript. TM, TPAK, and SAC wrote the manuscript. MR, LS, and GVR assisted, edited, and updated the manuscript. SI contributed critically in revising and improving the manuscript for publication.

FUNDING

SAC was supported by European Union through a Marie Curie International Incoming Fellowship (No: FP7-People-2-11-IIF-Acronym IMPACT-No: 300672 and 921672). TM and TAK were supported by Loyola College-Times of India grant (No: 7LCTOI14ERI001).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01054/full#supplementary-material

REFERENCES

Anjaneyulu, E., Reddy, P. S., Sunita, M. S., Kishor, P. B., and Meriga, B. (2014). Salt tolerance and activity of antioxidative enzymes of transgenic finger millet overexpressing a vacuolar H+-pyrophosphatase gene (SbVPPase) from *Sorghum bicolor*. *J. Plant Physiol.* 171, 789–798. doi: 10.1016/j.jplph.2014.02.001

Arif, I. A., Bakir, M. A., Khan, H. A., Al Farhan, A. H., Al Homaidan, A. A., Bahkali, A. H., et al. (2010). A brief review of molecular techniques to assess plant diversity. *Int. J. Mol. Sci.* 11, 2079–2096. doi: 10.3390/ijms11052079

Arockiasamy, S., and Ignacimuthu, S. (2007). Regeneration of transgenic plants from two indica rice (*Oryza sativa* L.) cultivars using shoot apex explants. *Plant Cell Rep.* 26, 1745–1753. doi: 10.1007/s00299-007-0377-9

Babitha, K. C., Ramu, S. V., Nataraja, K. N., Sheshshayee, M. S., and Udayakumar, M. (2015a). EcbZIP60, a basic leucine zipper transcription factor from *Eleusine coracana* L. improves abiotic stress tolerance in tobacco by activating unfolded protein response pathway. *Mol. Breed.* 35, 181–197. doi: 10.1007/s11032-015-0374-6

Babitha, K. C., Vemanna, R. S., Nataraja, K. N., and Udayakumar, M. (2015b). Overexpression of EcbHLH57 transcription factor from *Eleusine coracana* L. in tobacco confers tolerance to salt, oxidative and drought stress. *PLoS One* 10:e0137098. doi: 10.1371/journal.pone.0137098

Babu, A., Geetha, K., Manjunatha, V., and Shankar, A. (2012). An efficient high throughput plant regeneration and transformation protocol for production of transgenics tolerant to salt in finger millet. *Int. J. For. Crop Improv.* 3, 16–20.

Babu, B. K., Agrawal, P., Pandey, D., Jaiswal, J., and Kumar, A. (2014a). Association mapping of agro-morphological characters among the global collection of finger millet genotypes using genomic SSR markers. *Mol. Biol. Rep.* 41, 5287–5297. doi: 10.1007/s11033-014-3400-6

Babu, B. K., Agrawal, P. K., Pandey, D., and Kumar, A. (2014b). Comparative genomics and association mapping approaches for opaque2 modifier genes in finger millet accessions using genic, genomic and candidate gene-based simple sequence repeat markers. *Mol. Breed.* 34, 1261–1279. doi: 10.1007/s11032-014-0115-2

Babu, B. K., Dinesh, P., Agrawal, P. K., Sood, S., Chandrashekara, C., Bhatt, J. C., et al. (2014c). Comparative genomics and association mapping approaches for blast resistant genes in finger millet using SSRs. *PLoS One* 9:e99182. doi: 10.1371/journal.pone.0099182

Babu, B. K., Senthil, N., Gomezm, S. M., Biji, K. R., Rajendraprasad, N. S., Kumar, S. S., et al. (2007). Assessment of genetic diversity among finger millet (*Eleusine coracana* (L.) Gaertn.) accessions using molecular markers. *Genet. Resour. Crop Evol.* 54, 399–404. doi: 10.1007/s10722-006-0002-8

Babu, B. K., Sood, S., Agrawal, P. K., Chandrashekara, C., Kumar, A., and Kumar, A. (2017). Molecular and phenotypic characterization of 149 finger millet accessions using microsatellite and agro-morphological markers. *Proc. Natl. Acad. Sci. India Sect. B Biol. Sci.* 87, 1217–1228. doi: 10.1007/s40011-015-0695-6

Babu, G. A., Vinoth, A., and Ravindhran, R. (2018). Direct shoot regeneration and genetic fidelity analysis in finger millet using ISSR markers. *Plant Cell Tissue Organ Cult.* 132, 157–164. doi: 10.1007/s11240-017-1319-z

Babu, T. K., Thakur, R. P., Upadhyaya, H. D., Reddy, P. N., Sharma, R., Girish, A. G., et al. (2013). Resistance to blast (*Magnaporthe grisea*) in a mini-core collection of finger millet germplasm. *Eur. J. Plant Pathol.* 135, 299–311. doi: 10.1007/s10658-012-0086-2

Baker, A., Ceasar, S. A., Palmer, A. J., Paterson, J. B., Qi, W., Muench, S. P., et al. (2015). Replace, reuse, recycle: improving the sustainable use of phosphorus by plants. *J. Exp. Bot.* 66, 3523–3540. doi: 10.1093/jxb/erv210

Bayer, G. Y., Yemets, A., and Blume, Y. B. (2014). Obtaining the transgenic lines of finger millet *Eleusine coracana* (L.). with dinitroaniline resistance. *Cytol. Genet.* 48, 139–144. doi: 10.3103/S0095452714030025

Bellucci, A., Tondelli, A., Fangel, J. U., Torp, A. M., Xu, X., Willats, W. G. T., et al. (2017). Genome-wide association mapping in winter barley for grain yield and culm cell wall polymer content using the high-throughput CoMPP technique. *PLoS One* 12:e0173313. doi: 10.1371/journal.pone.0173313

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., et al. (2012). Reference genome sequence of the model plant Setaria. *Nat. Biotech.* 30, 555–561. doi: 10.1038/nbt.2196

Bolger, M. E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., and Mayer, K. F. X. (2014). Plant genome sequencing-applications for crop improvement. *Curr. Opin. Biotechnol.* 26, 31–37. doi: 10.1016/j.copbio.2013.08.019

Borlaug, N. E. (2002). Feeding a world of 10 billion people: the miracle ahead. *In Vitro Cell. Dev. Biol. Plant* 38, 221–228. doi: 10.1079/IVP2001279

Carter, C., Pan, S., Zouhar, J., Avila, E. L., Girke, T., and Raikhel, N. V. (2004). The vegetative vacuole proteome of *Arabidopsis thaliana* reveals predicted and unexpected proteins. *Plant Cell* 16, 3285–3303. doi: 10.1105/tpc.104.027078

Ceasar, S. A. (2018). Feeding world population amidst depleting phosphate reserves: the role of biotechnological interventions. *Open Biotechnol. J.* 12, 51–55. doi: 10.2174/1874070701812010051

Ceasar, S. A., Baker, A., and Ignacimuthu, S. (2017). Functional characterization of the PHT1 family transporters of foxtail millet with development of a novel Agrobacterium-mediated transformation procedure. *Sci. Rep.* 7:14064. doi: 10.1038/s41598-017-14447-0

Ceasar, S. A., Hodge, A., Baker, A., and Baldwin, S. A. (2014). Phosphate concentration and arbuscular mycorrhizal colonisation influence the growth, yield and expression of twelve PHT1 family phosphate transporters in foxtail millet (*Setaria italica*). *PLoS One* 9:e108459. doi: 10.1371/journal.pone.0108459

Ceasar, S. A., and Ignacimuthu, S. (2008). Efficient somatic embryogenesis and plant regeneration from shoot apex explants of different Indian genotypes of finger millet (*Eleusine coracana* (L.) Gaertn.). *In Vitro Cell. Dev. Biol. Plant* 44, 427–435. doi: 10.1007/s11627-008-9153-y

Ceasar, S. A., and Ignacimuthu, S. (2009). Genetic engineering of millets: current status and future prospects. *Biotechnol. Lett.* 31, 779–788. doi: 10.1007/s10529-009-9933-4

Ceasar, S. A., and Ignacimuthu, S. (2011). Agrobacterium-mediated transformation of finger millet (*Eleusine coracana* (L.) Gaertn.) using shoot apex explants. *Plant Cell Rep.* 30, 1759–1770. doi: 10.1007/s00299-011-1084-0

Ceasar, S. A., Rajan, V., Prykhozhij, S. V., Berman, J. N., and Ignacimuthu, S. (2016). Insert, remove or replace: a highly advanced genome editing system using CRISPR/Cas9. *Biochim. Biophys. Acta* 1863, 2333–2344. doi: 10.1016/j.bbamcr.2016.06.009

Chandra, D., Chandra, S., Pallavi, and Sharma, A. K. (2016). Review of finger millet (*Eleusine coracana* (L.) Gaertn): a power house of health benefiting nutrients. *Food Sci. Hum. Welln.* 5, 149–155. doi: 10.1016/j.fshw.2016.05.004

Chethan, S., and Malleshi, N. (2007). Finger millet polyphenols: optimization of extraction and the effect of pH on their stability. *Food Chem.* 105, 862–870. doi: 10.1016/j.foodchem.2007.02.012

Chinchole, M., Pathak, R. K., Singh, U. M., and Kumar, A. (2017). Molecular characterization of *EcCIPK24* gene of finger millet (*Eleusine coracana*) for investigating its regulatory role in calcium transport. *3 Biotech* 7:267. doi: 10.1007/s13205-017-0874-7

Dey, M., Bakshi, S., Galiba, G., Sahoo, L., and Panda, S. K. (2012). Development of a genotype independent and transformation amenable regeneration system from shoot apex in rice (*Oryza sativa* spp. indica) using TDZ. *3 Biotech* 2, 233–240. doi: 10.1007/s13205-012-0051-y

Dida, M. M., Wanyera, N., Dunn, M. L. H., Bennetzen, J. L., and Devos, K. M. (2008). Population structure and diversity in finger millet (*Eleusine coracana*) germplasm. *Trop. Plant Biol.* 1, 131–141. doi: 10.1007/s12042-008-9012-3

Dwivedi, S. L., Upadhyaya, H. D., Senthilvel, S., Hash, C. T., Fukunaga, K., Diao, X., et al. (2012). "Millets: genetic and genomic resources," in *Plant Breeding Reviews*, ed. J. Janick (Hoboken, NJ: Wiley), 247–375.

Eapen, S., and George, L. (1990). Influence of phytohormones, carbohydrates, aminoacids, growth supplements and antibiotics on somatic embryogenesis and plant differentiation in finger millet. *Plant Cell Tissue Organ Cult.* 22, 87–93. doi: 10.1007/BF00043683

Ekwamu, A. (1991). Influence of head blast infection on seed germination and yield components of finger millet (*Eleusine coracana* L. Gaertn). *Int. J. Pest Manage.* 37, 122–123. doi: 10.1080/09670879109371556

Fakrudin, B., Shashidhar, H., Kulkarni, R., and Hittalmani, S. (2004). Genetic diversity assessment of finger millet. *Eleusine coracana* (Gaertn.), germplasm through RAPD analysis. *PGR Newslett.* 138, 50–54.

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

Food and Agriculture Organization of the United States [FAO] (2015). *World Fertilizer Trends and Outlook to 2018*. Rome: Food and Agriculture Organization of the United States.

Gashaw, G., Alemu, T., and Tesfaye, K. (2014). Morphological, physiological and biochemical studies on *Pyricularia grisea* isolates causing blast disease on finger millet in Ethiopia. *J. Appl. Biosci.* 74, 6059–6071. doi: 10.4314/jab.v74i1.2

Goron, T. L., Bhosekar, V. K., Shearer, C. R., Watts, S., and Raizada, M. N. (2015). Whole plant acclimation responses by finger millet to low nitrogen stress. *Front. Plant Sci.* 6:652. doi: 10.3389/fpls.2015.00652

Goron, T. L., and Raizada, M. N. (2015). Genetic diversity and genomic resources available for the small millet crops to accelerate a new green revolution. *Front. Plant Sci.* 6:157. doi: 10.3389/fpls.2015.00157

Gupta, A. K., Gaur, V. S., Gupta, S., and Kumar, A. (2013). Nitrate signals determine the sensing of nitrogen through differential expression of genes involved in nitrogen uptake and assimilation in finger millet. *Funct. Integr. Genomics* 13, 179–190. doi: 10.1007/s10142-013-0311-x

Gupta, N., Gupta, A. K., Singh, N., and Kumar, A. (2011). Differential expression of PBF Dof transcription factor in different tissues of three finger millet genotypes differing in seed protein content and color. *Plant Mol. Biol. Rep.* 29, 69–76. doi: 10.1007/s11105-010-0208-y

Gupta, P., Raghuvanshi, S., and Tyagi, A. K. (2001). Assessment of the efficiency of various gene promoters via biolistics in leaf and regenerating seed callus of millets, *Eleusine coracana* and *Echinochloa crus-galli*. *Plant Biotechnol.* 18, 275–282. doi: 10.5511/plantbiotechnology.18.275

Gupta, R., Verma, K., Joshi, D. C., Yadav, D., and Singh, M. (2010). Assessment of genetic relatedness among three varieties of finger millet with variable seed coat color using RAPD and ISSR markers. *Genet. Eng. Biotechnol. J.* 2, 1–9.

Gupta, S., Gupta, S. M., Gupta, A. K., Gaur, V. S., and Kumar, A. (2014). Fluctuation of Dof1/Dof2 expression ratio under the influence of varying nitrogen and light conditions: involvement in differential regulation of nitrogen metabolism in two genotypes of finger millet (*Eleusine coracana* L.). *Gene* 546, 327–335. doi: 10.1016/j.gene.2014.05.057

Gupta, S. M., Arora, S., Mirza, N., Pande, A., Lata, C., Puranik, S., et al. (2017). Finger Millet: a "certain" crop for an "uncertain" future and a solution to food insecurity and hidden hunger under stressful environments. *Front. Plant Sci.* 8:643. doi: 10.3389/fpls.2017.00643

Hatakeyama, M., Aluri, S., Balachadran, M. T., Sivarajan, S. R., Patrignani, A., Grüter, S., et al. (2018). Multiple hybrid de novo genome assembly of finger millet, an orphan allotetraploid crop. *DNA Res.* 25, 39–47. doi: 10.1093/dnares/dsx036

Hema, R., Vemanna, R. S., Sreeramulu, S., Reddy, C. P., Senthil Kumar, M., and Udayakumar, M. (2014). Stable expression of mtlD gene imparts multiple stress tolerance in finger millet. *PLoS One* 9:e99110. doi: 10.1371/journal.pone.0099110

Hittalmani, S., Mahesh, H., Shirke, M. D., Biradar, H., Uday, G., Aruna, Y., et al. (2017). Genome and transcriptome sequence of finger millet (*Eleusine coracana* (L.) Gaertn.) provides insights into drought tolerance and nutraceutical properties. *BMC Genomics* 18:465. doi: 10.1186/s12864-017-3850-z

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., Yang, S., Gong, J., Zhao, Q., Feng, Q., Zhan, Q., et al. (2016). Genomic architecture of heterosis for yield traits in rice. *Nature* 537, 629–633. doi: 10.1038/nature19760

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., et al. (2011). 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

Ignacimuthu, S., and Ceasar, S. A. (2012). Development of transgenic finger millet (*Eleusine coracana* (L.) Gaertn.) resistant to leaf blast disease. *J. Biosci.* 37, 135–147. doi: 10.1007/s12038-011-9178-y

International Rice Genome Sequencing (2005). The map-based sequence of the rice genome. *Nature* 436, 793–800. doi: 10.1038/nature03895

Jagga-Chugh, S., Kachhwaha, S., Sharma, M., Kothari Chajer, A., and Kothari, S. (2012). Optimization of factors influencing microprojectile bombardment-mediated genetic transformation of seed-derived callus and regeneration of transgenic plants in *Eleusine coracana* (L.) Gaertn. *Plant Cell Tissue Organ Cult.* 109, 401–410. doi: 10.1007/s11240-011-0104-7

Jayasudha, B. G., Sushma, A. M., Prashantkumar, H. S., and Sashidhar, V. R. (2014). An efficient in-vitro agrobacterium mediated transformation protocol for raising salinity tolerant transgenic finger millet (*Eleusine coracana* (L.) Gaertn). *Plant Arch.* 14, 823–829.

Jeong, I.-S., Yoon, U.-H., Lee, G.-S., Ji, H.-S., Lee, H.-J., Han, C.-D., et al. (2013). SNP-based analysis of genetic diversity in anther-derived rice by whole genome sequencing. *Rice* 6:6. doi: 10.1186/1939-8433-6-6

Kanwal, P., Gupta, S., Arora, S., and Kumar, A. (2014). Identification of genes involved in carbon metabolism from *Eleusine coracana* (L.) for understanding their light-mediated entrainment and regulation. *Plant Cell Rep.* 33, 1403–1411. doi: 10.1007/s00299-014-1625-4

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

Khabarov, N., and Obersteiner, M. (2017). Global phosphorus fertilizer market and national policies: a case study revisiting the 2008 price peak. *Front. Nutr.* 4:22. doi: 10.3389/fnut.2017.00022

Kothari, S., Agarwal, K., and Kumar, S. (2004). Inorganic nutrient manipulation for highly improved in vitro plant regeneration in finger millet-*Eleusine coracana* (l.) Gaertn. *In Vitro Cell. Dev. Biol. Plant* 40, 515–519. doi: 10.1079/IVP2004564

Kumar, A., Metwal, M., Kaur, S., Gupta, A. K., Puranik, S., Singh, S., et al. (2016). Nutraceutical value of finger millet [*Eleusine coracana* (L.) Gaertn.], and their improvement using omics approaches. *Front. Plant Sci.* 7:934. doi: 10.3389/fpls.2016.00934

Kumar, A., Yadav, S., Panwar, P., Gaur, V. S., and Sood, S. (2015). Identification of anchored simple sequence repeat markers associated with calcium content in finger millet (*Eleusine coracana*). *Proc. Natl. Acad. Sci. India Sect. B Biol. Sci.* 85, 311–317. doi: 10.1007/s40011-013-0296-1

Kumar, B., and Kumar, J. (2011). Management of blast disease of finger millet (*Eleusine coracana*) through fungicides, bioagents and varietal mixture. *Indian Phytopathol.* 64:272.

Kumar, S., Mangal, M., Dhawan, A., and Singh, N. (2011). Assessment of genetic fidelity of micropropagated plants of *Simmondsia chinensis* (Link) Schneider using RAPD and ISSR markers. *Acta Physiol. Plant.* 33, 2541–2545. doi: 10.1007/s11738-011-0767-z

Kumssa, D. B., Joy, E. J. M., Ander, E. L., Watts, M. J., Young, S. D., Walker, S., et al. (2015). Dietary calcium and zinc deficiency risks are decreasing but remain prevalent. *Sci. Rep.* 5:10974. doi: 10.1038/srep10974

Latha, A. M., Rao, K. V., and Reddy, V. D. (2005). Production of transgenic plants resistant to leaf blast disease in finger millet (*Eleusine coracana* (L.) Gaertn.). *Plant Sci.* 169, 657–667. doi: 10.1016/j.plantsci.2005.05.009

Mahalakshmi, S., Christopher, G., Reddy, T., Rao, K., and Reddy, V. (2006). Isolation of a cDNA clone (PcSrp) encoding serine-rich-protein from *Porteresia coarctata* T. and its expression in yeast and finger millet (*Eleusine coracana* L.) affording salt tolerance. *Planta* 224, 347–359. doi: 10.1007/s00425-005-0218-4

Maharajan, T., Ceasar, S. A., Ajeesh Krishna, T. P., Ramakrishnan, M., Duraipandiyan, V., Naif Abdulla, A.-D., et al. (2018). Utilization of molecular markers for improving the phosphorus efficiency in crop plants. *Plant Breed.* 137, 10–26. doi: 10.1111/pbr.12537

Mahuku, G., Chen, J., Shrestha, R., Narro, L. A., Guerrero, K. V. O., Arcos, A. L., et al. (2016). Combined linkage and association mapping identifies a major QTL (qRtsc8-1), conferring tar spot complex resistance in maize. *Theor. Appl. Genet.* 129, 1217–1229. doi: 10.1007/s00122-016-2698-y

Matthies, I. E., Malosetti, M., Röder, M. S., and Van Eeuwijk, F. (2014). Genome-wide association mapping for kernel and malting quality traits using historical European barley records. *PLoS One* 9:e110046. doi: 10.1371/journal.pone.0110046

Mgonja, M. A., Lenne, J. M., Manyasa, E., and Sreenivasaprasad, S. (2007). "Finger millet blast management in East Africa. Creating Opportunities for Improving Production and Utilization of Finger Millet," in *Proceedings of the First International Finger Millet Stakeholder Workshop, Projects R8030 & R8445 UK Department for International Development—Crop Protection Programme*, (Patancheru: International Crops Research Institute for the Semi-AridTropics), 196.

Mirza, N., Taj, G., Arora, S., and Kumar, A. (2014). Transcriptional expression analysis of genes involved in regulation of calcium translocation and storage in finger millet (*Eleusine coracana* L. Gartn.). *Gene* 550, 171–179. doi: 10.1016/j.gene.2014.08.005

Mohammadi, S., and Prasanna, B. (2003). Analysis of genetic diversity in crop plants-salient statistical tools and considerations. *Crop Sci.* 43, 1235–1248. doi: 10.2135/cropsci2003.1235

Mohanty, B., Gupta, S. D., and Ghosh, P. (1985). Callus initiation and plant regeneration in ragi (*Eleusine coracana* Gaertn.). *Plant Cell Tissue Organ Cult.* 5, 147–150. doi: 10.1007/BF00040311

Nagaraja, A., Jagadish, P., Ashok, E., and Gowda, K. K. (2007). Avoidance of finger millet blast by ideal sowing time and assessment of varietal performance under rain fed production situations in Karnataka. *J. Mycopathol. Res.* 45, 237–240.

Nagarjuna, K. N., Parvathi, M. S., Sajeevan, R. S., Pruthvi, V., Mamrutha, H. M., and Nataraja, K. N. (2016). Full-length cloning and characterization of abiotic stress responsive CIPK31-like gene from finger millet, a drought-tolerant crop. *Curr. Sci.* 111, 890–894. doi: 10.18520/cs/v111/i5/890-894

Nath, M., Roy, P., Shukla, A., and Kumar, A. (2013). Spatial distribution and accumulation of calcium in different tissues, developing spikes and seeds of finger millet genotypes. *J. Plant Nutr.* 36, 539–550. doi: 10.1080/01904167.2012.748072

Parani, M., Rajesh, K., Lakshmi, M., Parducci, L., Szmidt, A., and Parida, A. (2001). Species identification in seven small millet species using polymerase chain reaction-restriction fragment length polymorphism of trn S-psb C gene region. *Genome* 44, 495–499. doi: 10.1139/gen-44-3-495

Parvathi, M., Nataraja, K. N., Yashoda, B., Ramegowda, H., Mamrutha, H., and Rama, N. (2013). Expression analysis of stress responsive pathway genes linked to drought hardiness in an adapted crop, finger millet (*Eleusine coracana*). *J. Plant Biochem. Biotechnol.* 22, 193–201. doi: 10.1007/s13562-012-0135-0

Parvathi, M. S., and Nataraja, K. N. (2017). Discovery of stress responsive TATA-box binding protein associated Factor6 (TAF6) from finger millet (*Eleusine coracana* (L.) Gaertn). *J. Plant Biol.* 60, 335–342. doi: 10.1007/s12374-016-0574-6

Patil, J., and Kale, A. (2013). Study of genetic diversity in finger millet (*Elesuine coracana* L.) genotypes using RAPD markers. *Int. J. Integr. Sci. Innov. Technol.* 2, 31–36.

Patrick, J. W., and Offler, C. E. (2001). Compartmentation of transport and transfer events in developing seeds. *J. Exp. Bot.* 52, 551–564. doi: 10.1093/jexbot/52.356.551

Pudake, R. N., Mehta, C. M., Mohanta, T. K., Sharma, S., Varma, A., and Sharma, A. K. (2017). Expression of four phosphate transporter genes from finger millet (*Eleusine coracana* L.) in response to mycorrhizal colonization and Pi stress. *3 Biotech* 7:17. doi: 10.1007/s13205-017-0609-9

Puranik, S., Kam, J., Sahu, P. P., Yadav, R., Srivastava, R. K., Ojulong, H., et al. (2017). Harnessing finger millet to combat calcium deficiency in humans: challenges and prospects. *Front. Plant Sci.* 8:1311. doi: 10.3389/fpls.2017.01311

Rahman, H., Jagadeeshselvam, N., Valarmathi, R., Sachin, B., Sasikala, R., Senthil, N., et al. (2014). Transcriptome analysis of salinity responsiveness in contrasting genotypes of finger millet (*Eleusine coracana* L.) through RNA-sequencing. *Plant Mol. Biol.* 85, 485–503. doi: 10.1007/s11103-014-0199-4

Rahman, H., Ramanathan, V., Nallathambi, J., Duraialagaraja, S., and Muthurajan, R. (2016). Over-expression of a NAC 67 transcription factor from finger millet (*Eleusine coracana* L.) confers tolerance against salinity and drought stress in rice. *BMC Biotechnol.* 16:35. doi: 10.1186/s12896-016-0261-1

Rajendran, H. A. D., Muthusamy, R., Stanislaus, A. C., Krishnaraj, T., Kuppusamy, S., Ignacimuthu, S., et al. (2016). Analysis of molecular variance and population structure in southern Indian finger millet genotypes using three different molecular markers. *J. Crop Sci. Biotechnol.* 19, 275–283. doi: 10.1007/s12892-016-0015-6

Ramakrishna, C., Singh, S., Raghavendrarao, S., Padaria, J. C., Mohanty, S., Sharma, T. R., et al. (2018). The membrane tethered transcription factor EcbZIP17 from finger millet promotes plant growth and enhances tolerance to abiotic stresses. *Sci. Rep.* 8:2148. doi: 10.1038/s41598-018-19766-4

Ramakrishnan, M., Ceasar, S. A., Duraipandiyan, V., Al-Dhabi, N., and Ignacimuthu, S. (2016a). Assessment of genetic diversity, population structure and relationships in Indian and non-Indian genotypes of finger millet (*Eleusine coracana* (L.) Gaertn) using genomic SSR markers. *Springerplus* 5:120. doi: 10.1186/s40064-015-1626-y

Ramakrishnan, M., Ceasar, S. A., Duraipandiyan, V., Al-Dhabi, N., and Ignacimuthu, S. (2016b). Using molecular markers to assess the genetic diversity

and population structure of finger millet (*Eleusine coracana* (L.) Gaertn.) from various geographical regions. *Genet. Resour. Crop Evol.* 63, 361–376. doi: 10.1007/s10722-015-0255-1

Ramakrishnan, M., Ceasar, S. A., Duraipandiyan, V., Vinod, K., Kalpana, K., Al-Dhabi, N., et al. (2016c). Tracing QTLs for leaf blast resistance and agronomic performance of finger millet (*Eleusine coracana* (L.) Gaertn.) genotypes through association mapping and in silico comparative genomics analyses. *PLoS One* 11:e0159264. doi: 10.1371/journal.pone.0159264

Ramakrishnan, M., Ceasar, S. A., Vinod, K., Duraipandiyan, V., Krishna, T. A., Upadhyaya, H. D., et al. (2017). Identification of putative QTLs for seedling stage phosphorus starvation response in finger millet (*Eleusine coracana* L. Gaertn.) by association mapping and cross species synteny analysis. *PLoS One* 12:e0183261. doi: 10.1371/journal.pone.0183261

Ramegowda, V., Gill, U. S., Sivalingam, P. N., Gupta, A., Gupta, C., Govind, G., et al. (2017). GBF3 transcription factor imparts drought tolerance in *Arabidopsis thaliana*. *Sci. Rep.* 7:9148. doi: 10.1038/s41598-017-09542-1

Ramegowda, V., Senthil-Kumar, M., Nataraja, K. N., Reddy, M. K., Mysore, K. S., and Udayakumar, M. (2012). Expression of a finger millet transcription factor. EcNAC1, in tobacco confers abiotic stress-tolerance. *PLoS One* 7:e40397. doi: 10.1371/journal.pone.0040397

Rangan, T. (1976). Growth and plantlet regeneration in tissue cultures of some Indian millets: *Paspalum scrobiculatum* L., *Eleusine coracana* Gaertn. and *Pennisetum typhoideum* Pers. *Z. Pflanzenphysiol.* 78, 208–216. doi: 10.1016/S0044-328X(73)80003-0

Rao, A. (1990). Estimates of losses in finger millet (*Eleusine coracana*) due to blast disease (*Pyricularia grisea*). *Mysore J. Agric. Sci.* 24, 57–60.

Rath, G., and Mishra, D. (1975). Nature of losses due to neck blast infection in ragi. *Sci. Cult.* 41, 322–323.

Ren, J., Sun, D., Chen, L., You, F. M., Wang, J., Peng, Y., et al. (2013). Genetic diversity revealed by single nucleotide polymorphism markers in a worldwide germplasm collection of durum Wheat. *Int. J. Mol. Sci.* 14, 7061–7088. doi: 10.3390/ijms14047061

Saha, D., Gowda, M. V. C., Arya, L., Verma, M., and Bansal, K. C. (2016). Genetic and genomic resources of small millets. *Crit. Rev. Plant Sci.* 35, 56–79. doi: 10.1080/07352689.2016.1147907

Sakamma, S., Umesh, K. B., Girish, M. R., Ravi, S. C., Satishkumar, M., and Bellundagi, V. (2018). Finger millet (*Eleusine coracana* L. Gaertn.) production system: status, potential, constraints and implications for improving small farmer's welfare. *J. Agric. Sci.* 10, 162–179. doi: 10.5539/jas.v10n1p162

Satish, L., Ceasar, S. A., and Ramesh, M. (2017). Improved Agrobacterium-mediated transformation and direct plant regeneration in four cultivars of finger millet (*Eleusine coracana* (L.) Gaertn.). *Plant Cell Tissue Organ Cult.* 131, 547–565. doi: 10.1007/s1124

Satish, L., Ceasar, S. A., Shilpha, J., Rency, A. S., Rathinapriya, P., and Ramesh, M. (2015). Direct plant regeneration from in vitro-derived shoot apical meristems of finger millet (*Eleusine coracana* (L.) Gaertn.). *In Vitro Cell. Dev. Biol. Plant* 51, 192–200. doi: 10.1007/s11627-015-9672-2

Satish, L., Rathinapriya, P., Rency, A. S., Ceasar, S. A., Pandian, S., Rameshkumar, R., et al. (2016a). Somatic embryogenesis and regeneration using *Gracilaria edulis* and *Padina boergesenii* seaweed liquid extracts and genetic fidelity in finger millet (*Eleusine coracana*). *J. Appl. Phycol.* 28, 2083–2098. doi: 10.1007/s10811-015-0696-0

Satish, L., Rency, A. S., Rathinapriya, P., Ceasar, S. A., Pandian, S., Rameshkumar, R., et al. (2016b). Influence of plant growth regulators and spermidine on somatic embryogenesis and plant regeneration in four Indian genotypes of finger millet (*Eleusine coracana* (L.) Gaertn). *Plant Cell Tissue Organ Cult.* 124, 15–31. doi: 10.1007/s11240-015-0870-8

Sharma, D., Jamra, G., Singh, U. M., Sood, S., and Kumar, A. (2017). Calcium biofortification: three pronged molecular approaches for dissecting complex trait of calcium nutrition in finger millet (*Eleusine coracana*) for devising strategies of enrichment of food crops. *Front. Plant Sci.* 7:2028. doi: 10.3389/fpls.2016.02028

Sharma, M., Kothari-Chajer, A., Jagga-Chugh, S., and Kothari, S. (2011). Factors influencing *Agrobacterium tumefaciens*-mediated genetic transformation of *Eleusine coracana* (L.) Gaertn. *Plant Cell Tissue Organ Cult.* 105, 93–104. doi: 10.1007/s11240-010-9846-x

Shrawat, A. K., and Lörz, H. (2006). Agrobacterium-mediated transformation of cereals: a promising approach crossing barriers. *Plant Biotechnol. J.* 4, 575–603. doi: 10.1111/j.1467-7652.2006.00209.x

Singh, R. K., Singh, V. K., Raghavendrarao, S., Phanindra, M. L. V., Raman, K. V., Solanke, A. U., et al. (2015). Expression of finger millet EcDehydrin7 in transgenic tobacco confers tolerance to drought stress. *Appl. Biochem. Biotechnol.* 177, 207–216. doi: 10.1007/s1201

Singh, U. M., Chandra, M., Shankhdhar, S. C., and Kumar, A. (2014). Transcriptome wide identification and validation of calcium sensor gene family in the developing spikes of finger millet genotypes for elucidating its role in grain calcium accumulation. *PLoS One* 9:e103963. doi: 10.1371/journal.pone.0103963

Singh, U. M., Metwal, M., Singh, M., Taj, G., and Kumar, A. (2015). Identification and characterization of calcium transporter gene family in finger millet in relation to grain calcium content. *Gene* 566, 37–46. doi: 10.1016/j.gene.2015.04.021

Singh, Y., and Kumar, J. (2010). Study of genomic fingerprints profile of *Magnaporthe grisea* from finger millet (*Eleusine coracana*) by random amplified polymorphic DNA-polymerase chain reaction (RAPD-PCR). *Afr. J. Biotechnol.* 9, 7798–7804. doi: 10.5897/AJB09.1648

Sood, S., Kumar, A., Babu, B. K., Gaur, V. S., Pandey, D., Kant, L., et al. (2016). Gene discovery and advances in finger millet (*Eleusine coracana* (L.) Gaertn.) genomics-an important nutri-cereal of future. *Front. Plant Sci.* 7:1634. doi: 10.3389/fpls.2016.01634

Tang, W., Wu, T., Ye, J., Sun, J., Jiang, Y., Yu, J., et al. (2016). SNP-based analysis of genetic diversity reveals important alleles associated with seed size in rice. *BMC Plant Biol.* 16:93. doi: 10.1186/s12870-016-0779-3

Thilakarathna, M. S., and Raizada, M. N. (2015). A review of nutrient management studies involving finger millet in the semi-arid tropics of Asia and Africa. *Agronomy* 5, 262–290. doi: 10.3390/agronomy5030262

Upadhyaya, H., Gowda, C., and Reddy, V. G. (2007). Morphological diversity in finger millet germplasm introduced from Southern and Eastern Africa. *J. SAT Agric. Res.* 3, 1–3.

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, X., Wang, H., Liu, S., Ferjani, A., Li, J., Yan, J., et al. (2016). Genetic variation in ZmVPP1 contributes to drought tolerance in maize seedlings. *Nat. Genet.* 48, 1233–1241. doi: 10.1038/ng.3636

Yamunarani, R., Govind, G., Ramegowda, V., Thammegowda, H. V., and Guligowda, S. A. (2016). Genetic diversity for grain Zn concentration in finger millet genotypes: Potential for improving human Zn nutrition. *Crop J.* 4, 229–234. doi: 10.1016/j.cj.2015.12.001

Yemets, A. I., Bayer, G. Y., and Blume, Y. B. (2013). An effective procedure for in vitro culture of *Eleusine coracana* (L.) and its application. *ISRN Bot.* 2013:853121. doi: 10.1155/2013/853121

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., et al. (2012). Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* 30, 549–554. doi: 10.1038/nbt.2195

Zhang, M., Ye, J., Xu, Q., Feng, Y., Yuan, X., Yu, H., et al. (2018). Genome-wide association study of cold tolerance of Chinese indica rice varieties at the bud burst stage. *Plant Cell Rep.* 37, 529–539. doi: 10.1007/s00299-017-2247-4

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 July 2018
doi: 10.3389/fpls.2018.00934

Dissecting Key Adaptation Traits in the Polyploid Perennial *Medicago sativa* Using GBS-SNP Mapping

Laxman Adhikari[1], *Orville M. Lindstrom*[2], *Jonathan Markham*[1] *and Ali M. Missaoui*[1]*

[1] Crop and Soil Sciences and Institute of Plant Breeding Genetics and Genomics, Center for Applied Genetic Technologies, University of Georgia, Athens, GA, United States, [2] Department of Horticulture, University of Georgia, Athens, GA, United States

Understanding key adaptation traits is crucial to developing new cultivars with broad adaptations. The main objective of this research is to understand the genetic basis of winter hardiness (WH) and fall dormancy (FD) in alfalfa and the association between the two traits. QTL analysis was conducted in a pseudo-testcross F1 population developed from two cultivars contrasting in FD (3010 with $FD = 2$ and CW 1010 with $FD = 10$). The mapping population was evaluated in three replications at two locations (Watkinsville and Blairsville, GA). FD levels showed low to moderate correlations with WH (0.22–0.57). Assessing dormancy in winter is more reliable than in the fall in southern regions with warm winters. The mapping population was genotyped using Genotyping-by-sequencing (GBS). Single dose allele SNPs (SDA) were used for constructing linkage maps. The parental map (CW 1010) consisted of 32 linkage groups spanning 2127.5 cM with 1377 markers and an average marker density of 1.5 cM/SNP. The maternal map (3010) had 32 linkage groups spanning 2788.4 cM with 1837 SDA SNPs with an average marker density of 1.5 cM/SNP. Forty-five significant ($P < 0.05$) QTLs for FD and 35 QTLs for WH were detected on both male and female linkage maps. More than 75% (22/28) of the dormancy QTL detected from the 3010 parent did not share genomic regions with WH QTLs and more than 70% (12/17) dormancy QTLs detected from CW 1010 parent were localized in different genomic regions than WH QTLs. These results suggest that the two traits have independent inheritance and therefore can be improved separately in breeding programs.

Keywords: alfalfa, genetic map, QTL, genotype X environment interaction, fall dormancy, winter hardiness

OPEN ACCESS

Edited by:
Shuizhang Fei,
Iowa State University, United States

Reviewed by:
Xuehui Li,
North Dakota State University, United States
Joseph Robins,
United States Department of Agriculture, United States

***Correspondence:**
Ali M. Missaoui
cssamm@uga.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *13 April 2018*
Accepted: *11 June 2018*
Published: *04 July 2018*

Citation:
Adhikari L, Lindstrom OM, Markham J and Missaoui AM (2018) Dissecting Key Adaptation Traits in the Polyploid Perennial Medicago sativa Using GBS-SNP Mapping. Front. Plant Sci. 9:934. doi: 10.3389/fpls.2018.00934

INTRODUCTION

Alfalfa (*Medicago sativa* L.) is a perennial cool-season forage legume grown worldwide for hay, pasture and silage (Duke, 1981; Adhikari and Missaoui, 2017). It is native to the southwestern and central Asia, near southern Caucasus Mountains (Duke, 1981; Li and Brummer, 2012). Alfalfa is well-regarded for providing high-quality forage with high protein content and nutritive value (Duke, 1981). Like other legumes, alfalfa fixes atmospheric nitrogen (N), up to 130–220 lbs per acre per year, thereby supplying N to itself and succeeding crops in rotation[1]. In the US., alfalfa

[1]http://nmsp.cals.cornell.edu/publications/factsheets/factsheet39.pdf, Cornell University. 2008

and its mixtures contribute a major part of haylage production, where the productivity varies from 1.1 ton/acre (Rhode Island) to 7 ton/acre (California) with an average national productivity of 3.45 ton/acre in 2016[2]. Alfalfa is cross-pollinated and highly heterozygous. It is a polyploid ($2n = 4x = 32$) with tetrasomic inheritance and a genome size near 1 Gb (Li and Brummer, 2012). Alfalfa grows best in cool sub-tropical and warm temperate environments (Duke, 1981). Growth and yield are remarkably affected by seasonal dormancy and low temperature stress in winter (Adhikari et al., 2017).

Alfalfa evolved FD as an important adaptation strategy to survive in latitudes with harsh winter conditions. The short growth cycle of fall-dormant alfalfa varieties limits not only the amounts of biomass accumulated but also the seasonal distribution, which is reduced to a few harvests per year in summer. FD rating (FDR) of alfalfa cultivars is assigned based on fall regrowth height, after clipping, to 11 groups ranging from FD 1 to FD 11 with lower numbers indicating more dormant (Teuber et al., 1998). These groups are very dormant, (FD 1, 2); dormant (FD 3, 4), moderately dormant (FD 5), semi-dormant (FD 6, 7), non-dormant (FD 8, 9), and very non-dormant (FD 10, 11)[3]. Dormancy classes are assigned based on standard check cultivars. Diminishing day length and temperature in fall season are the two major environmental factors triggering physiological dormancy in alfalfa (McKenzie et al., 1988; Brummer et al., 2000). FD is a strongly expressed trait where certain genotypes exhibit slow growth leading to a short stature and decumbent plant architecture after autumn clipping (Teuber et al., 1998; Adhikari et al., 2017). In order to assign FD accurately in the field, it is suggested to collect information from multiple locations for least for 2 years (Teuber et al., 1998).

The genetic control of FD in alfalfa is not known and investigation into the endogenous factors influencing FD will be valuable for developing cultivars with no- or short-fall dormancy. The molecular basis of dormancy has been studied mostly in woody species adapted to temperate environments. There are few reports on QTLs associated with the dormancy trait in herbaceous forage species. Some QTLs associated with fall growth and WH were mapped using an interspecific hybrid population developed by crossing annual x perennial ryegrass (Xiong et al., 2007). Day length and temperature are most likely the two major environmental cues that plants use to sense the environmental changes (Olsen, 2010; Tanino et al., 2010). Genomic studies have identified a number of genes involved in the control of dormancy induction and growth cessation, including circadian clock regulators (Ibáñez et al., 2010). However, McKenzie et al. (1988) argued that alfalfa FD is not physiologically similar to that of higher trees since the plant exhibits dormancy due to decreasing day length and temperature but it is reversible when alfalfa is switched to an environment with warmer temperature and longer photoperiod. Research investigating the genetic and physiological basis of FD in alfalfa, in the context of genes, quantitative trait loci (QTL) and hormones regulating the process of alfalfa FD (Brouwer et al., 2000; Li and Brummer, 2012) suggested that FD in alfalfa is correlated with winter survival and very often, fall dormant alfalfa is considered more winter-hardy (Stout and Hall, 1989; Li et al., 2014a). In northern latitudes, mostly dormant germplasm is grown because they have better chances of completing the development cycle and go dormant before the onset of freezing temperatures in early fall. There is a lack of consensus regarding the relationship between fall regrowth and WH even though alfalfa breeders have been routinely using FD as a surrogate to select for cold tolerance in northern latitudes. A strong phenotypic as well as genetic correlation between dormancy and WH was observed in alfalfa breeding populations developed from wide dormancy crosses involving parents with contrasting dormancy ratings (Li et al., 2015). Cunningham et al. (1998) examined the impact of selection for differences in FD on carbohydrate and protein accumulation in roots and crown buds as well as its effect on winter survival and bud development in four alfalfa parents and their progeny. They concluded, after three cycles of selection, that imposing selection on FD will lead to improved cold acclimation and winter survival. Brummer et al. (2000) stressed the need for reexamining the relationship between FD and WH because contrary to the traditional concept, they found weak association between the two traits (Brummer et al., 2000). Similarly, quantitative trait loci (QTLs) independently controlling autumn plant growth and winter survival were reported indicating the possibility of independent improvement of the two traits through marker-assisted selection (MAS) (Li et al., 2015). In a recent study, scientists have identified differentially expressed genes such as C-repeat binding factors (CBF) in response to freezing stress in alfalfa which may be induced regardless of the genotype dormancy (Shu et al., 2017). Zhang et al. (2015) also observed several differentially expressed genes in fall dormant lines in leaf transcriptome analysis (Zhang et al., 2015). Similarly, alfalfa cold acclimation specific (*CAS*) genes such as *cas15* and other cold related genes are also potential genetic factors controlling WH without affecting dormancy (Castonguay et al., 2011; Li et al., 2015). There has been a limited progress in developing non-dormant alfalfa varieties with improved cold and freeze tolerance. Most of the studies have been conducted in Northern latitudes on dormant germplasm or in growth chambers rather than in the field under real winter conditions. Significant differences are known to exist between natural and artificial cold acclimation conditions and therefore plants that are cold acclimated in growth chambers may react differently compared to those acclimated naturally (Dhanaraj et al., 2007). Field grown plants are often exposed to varying light spectrum and intensities compared to the constant conditions in growth chambers. Plants in the field are also frequently exposed to strong winds that influence gene expression and plant structure (Gusta and Wisniewski, 2013). Dhanaraj et al. (2007) documented a large number of genes that were induced in a growth chamber but not under field conditions (Dhanaraj et al., 2007). An understanding of the interconnection between genetic factors and networks that control winter dormancy and WH will provide fundamental knowledge needed for the development of genomic resources that will enable selection of non-dormant

[2]http://usda.mannlib.cornell.edu/usda/current/CropProdSu/CropProdSu-01-12-2018.pdf

[3]https://www.alfalfa.org/pdf/2017%20NAFA%20Variety%20Leaflet.pdf

alfalfa germplasm that persist well under occasional freezing temperatures. Therefore, dissecting the relationship between alfalfa FD and WH at the genomics level would be valuable to improving alfalfa.

Genetic analysis of FD and WH in alfalfa through QTL mapping requires adequate genome coverage with molecular markers. A large number of SNPs can be obtained cost effectively through next generation sequencing methods like genotyping-by-sequencing (GBS) even in species with no prior genome assemblies. The GBS method developed by Elshire et al. (2011) comprises selective fragmentation of DNA by specific enzymes, ligation of common and barcode adapters, PCR, clean up and sequencing (Elshire et al., 2011). The GBS method has been used successfully in discovering SNP markers in several diploids and autotetraploids crop species such as potato (*Solanum tuberosum* L.), rose (*Rosa hybrida*), and alfalfa (Gar et al., 2011; Li et al., 2014b; Boudhrioua et al., 2017). However, in species with tetrasomic inheritance, only certain biallelic SNPs (simplex, duplex, double simplex) can be mapped. Since most of the mapping software are designed for diploid genomes, mapping autopolyploids is cumbersome. Some new software applications can handle this issue, but they still have limitations. TetraploidMap seems useful in adjusting markers segregating in various ratios (simplex, duplex, double simplex), but it can fit only about 800 markers and works better when each linkage group has <50 markers (Hackett et al., 2007; Li et al., 2014b). Similarly, TetraploidSNPMap can support a higher number of SNPs, but requires SNP dosage data from SNP array (Hackett et al., 2017). However, most of the autotetraploid QTL maps available so far such as potato (da Silva et al., 2017) and alfalfa (Li et al., 2015) maps were constructed using TetraploidMap or TetraploidSNPMap. Mapping autotetraploids with unique kinds of markers using software like JoinMap is also common. Often the pseudo-testcross simplex markers (AB x BB), i.e., markers heterozygous in one parent and not the other, are used to construct autotetraploid genetic maps in software like JoinMap[4]. The pseudo-tetstcross strategy allows the use of several thousand single dose SNPs and is considered a simple method of linkage mapping (Li et al., 2014b). Identifying quantitative trait loci (QTLs) underlying FD and WH will enable understanding the genetic factors controlling these traits and helps in discovering markers associated with each trait. Manipulation of these alleles through MAS will enable the development of non-dormant alfalfa cultivars with improved WH. The objective of this study was to understand the genetic basis of alfalfa FD and WH via genetic linkage analysis and QTL mapping.

MATERIALS AND METHODS

Mapping Population

An F1 population was developed by crossing a tetraploid dormant (FD = 2) winter-hardy alfalfa cultivar (3010, ♀) with a tetraploid non-dormant (FD = 10) winter susceptible cultivar (CW 1010, ♂). The cross was made in the greenhouse using hand pollination in isolation under 18 hr. light and 6 hr. dark. About 384 F1 seeds were harvested, scarified, inoculated with rhizobium strain N-dure (INTX Microbials LLC, *Sinorhizobium meliloti* and *Rhizobium leguminosarum*), and grown in the greenhouse. In order to confirm the true hybrids, 24 simple sequence repeat (SSR) markers were screened for polymorphism between the parents. These markers were developed from *M. truncatula* (Eujayl et al., 2004) and were previously used to genotype tetraploid alfalfa (Li et al., 2011). From the set of 24 SSR markers, three markers with the strongest amplification and highest polymorphic index between the two parents were used to genotype the F1 progeny. Two hundred true F1 hybrids were retained but sufficient numbers of clones for the target locations and replications were obtained only from 184 hybrids. Twenty-four clones per entry were generated through stem cuttings and propagated in the field.

The two parents, 184 F1 progenies, 11 standard check cultivars for FD, and six checks for winter survival[5] were planted at two locations in Georgia. The first was the J. Phil Campbell Sr. Research and Education Center (JPC) in Watkinsville (33°52′17.8″N 83°27′05.5″W) and the other was the Georgia Mountain Research and Education Center at Blairsville (BVL) (34°50′21.4″N 83°55′20.5″W). The BVL location experiences frequently harsh winters and therefore is considered an ideal location to test alfalfa WH and persistence under cold stress. The average annual precipitation in the BVL location is 55.9 in, the average high temperature in July is 29°C, and the lowest temperature in January is −4°C. At the Watkinsville location, the average annual precipitation is 48 in, the highest temperature in July is 32.2°C while the lowest temperature in January is 0°C. The experimental design at each location was a randomized complete block, with three replications, where four clones from each progeny were planted in a single row plot. Plants were spaced 45 cm within each row, and the rows were spaced 90 cm from each other. Irrigation, fertilization and weed control were applied as necessary.

Genotyping-by-Sequencing (GBS) and Marker Discovery

Single nucleotide polymorphisms (SNPs) markers were identified using genotyping-by-sequencing of the parents and progeny. DNA of each progeny and parents was extracted using CTAB method with some modifications (Doyle and Doyle, 1987). Alfalfa tissue was collected in 50-ml tubes, freeze dried for 48 h, and grinded using 6–8 zinc-plated copper balls in a Genogrinder (SPEX SamplePrep 2010 Geno/Grinder®) for 6 min at 1,600 rpm. Then, 150 mg of the powder was transferred to 2.0 ml tubes, 900 μl of CTAB buffer was added, vortexed, and the mixture was incubated at 65°C for 1 h. Nine hundred microliter of phenol:chlorofom:isoamyl alcohol (PCI) mix (25:24:1) with pH 5.0 was added to each tube, incubated for 15 min and centrifuged at 12,500 rpm for 15 min and the clear supernatants were pipetted to new 2.0 ml tubes. Equal volume of chloroform:isoamyl alcohol (CIA) mix (24:1) was added to each tube, mixed gently and centrifuged at 12,500 rpm for 15 min. The aqueous upper phase was pipetted into 2.0 ml tube. About 0.6 volumes of

[4]https://www.kyazma.nl/docs/JM5Features.pdf

[5]https://www.alfalfa.org/pdf/2014%20NAFA%20Variety%20Leaflet.pdf

chilled isopropanol was added to the tubes and left for 10 min, centrifuged at 13,000 rpm for 20 min. The DNA pellet was washed using 500 μl 70% ethanol, centrifuged at 7,500 rpm for 5 min. The supernatant was discarded, and the DNA pellet was air dried for 1–2 h under the airflow. Then, 100 μl of sterile 10 mM Tris-HCl (pH 8.0) was added and incubated at 4°C for overnight.

The DNA solution was treated with 4 μl (10 mg/ml) of RNase A, followed by 5.0 μl (20 mg/ml) proteinase K, and incubated at 37°C in a water bath for 30 min after each addition. Sterile Millipore grade H_2O was added (400 μl) and treated with PCI and CIA as described earlier. The supernatant was pipetted into 1.5 ml tubes and 1/10 volume of 3 M Na-acetate pH 5.2 (stored at 4°C) and 2.5 volumes of absolute ethanol was added, left for 10 min and centrifuged for 20 min at 13,000 rpm. The supernatant was discarded and the pellet was washed with 70% ethanol, then air dried. The DNA was dissolved into 50–100 μl of sterile 10 mM Tris-Cl (pH 8.0). High quality DNA was ensured by quantification in Qubit® 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA USA) and running DNA samples in 1% agarose gel.

Two GBS libraries were constructed for 184 F1 progenies and the 2 parents. Both libraries were 96-plex including 92 F1 progenies and 2 replications of each parent. The barcode adapters, common adapters, and two PCR primers were ordered from Integrated DNA Technologies (Coralville, IA, USA). The library was constructed using the protocol described in Li et al. (2014b). The DNA samples were digested with methylation sensitive enzyme *ApeKI* and both common as well as barcode adapters were ligated. The step was followed by pooling the libraries (multiplexing) and cleaning up with Qiagen PCR (Qiagen, Germantown, MD) cleanup kit using the protocol provided with the kit. Moreover, the steps were followed by simple PCR using Kapa Library Amplification Readymix (Kapa Biosystems, Wilmington, MA) and two PCR primers. Finally, PCR products were purified using QIAquick PCR purification kit (Qiagen, Germantown, MD). Both libraries were submitted to Georgia Genomics and Bioinformatics Core (GGBC), UGA, for removing short fragments by solid phase reversible immobilization (SPRI), cleanup, and sequencing. Sequencing was performed on an Illumina Next Seq (150 Cycles) 75 PE High Output flow cell with four lanes. The raw sequence data was processed using two pipelines; GBS SNP Calling Reference Optional Pipeline (GBS-SNP-CROP) version 2.0 (Melo et al., 2016) and Tassel 3.0 Universal Network Enabled Analysis Kit (UNEAK) pipelines (Lu et al., 2013) for de novo SNP discovery. These bioinformatics computational steps were performed on in the Unix platform "Zcluster" at the Georgia Advanced Computing Resource Center (GACRC), UGA.

The GBS-SNP-CROP was a useful tool for de novo SNP calling. The raw reads were parsed and trimmed for quality using Trimmomatic software version 0.36 (Bolger and Giorgi, 2014). The trimmed reads were demultiplexed producing high-quality reads for each genotype. The GBS specific mock reference was generated from parsed high-quality reads. The processed reads were mapped to generate standard alignment files using BWA-mem (Santhosh, 1989) and SAMtool version 1.3.1 (Li et al., 2009). Subsequently the SNP master matrix was produced followed by SNP and genotype calling. The raw sequence data was deposited at NCBI SRA website under the accession number SRP150116 and it can be accessed at https://www.ncbi.nlm.nih.gov/sra/SRP150116.

Similarly, UNEAK pipeline was used to process the high quality R1 reads of pair-end data. In UNEAK, the raw R1 reads were filtered, de-multiplexed and trimmed to 64 bp. Similar reads were grouped as a tag, where tags with >10 reads were used for alignment in SNP calling. The parameters used to call and filter the homozygote alleles, heterozygote alleles, minor alleles etc. in UNEAK were as described (Li et al., 2014b). The HapMap output files obtained from UNEAK were processed in Microsoft Excel for separating parental genotypes, removing missing data and testing for segregation ratios using chi-square.

Linkage Map Construction

Polymorphic SNPs unique to either 3010 (AB x AA) or CW 1010 (AAx AB) were screened as single dose allele (SDA) markers (Li et al., 2014b). Parental genotypes that were heterozygous (AB) at any replication were considered heterozygous and parental genotype homozygous (AA) in all replications were considered homozygote. Markers that were missing in more than 30% of the progeny were culled. The SDA markers obtained from both pipelines were added and input files were formatted as required by JoinMap 5.0[6]. The SDA segregation ratio (1:1) was confirmed by chi-square tests ($p > 0.05$). The SNPs that were present in more than 30% progenies but have segregation ratios other than 1:1 were considered in segregation distortion (Li et al., 2014b).

Both male and female SDA markers were loaded to JoinMap 5.0 separately. The markers were grouped using minimum independence LOD of 10. The grouped markers were mapped using regression mapping with minimum LOD value of one, maximum recombination frequency of 0.40, and Kosambi mapping function. Linkage maps were generated using Map Chart and the map files were exported. The tags of mapped markers from UNEAK pipeline were separated for each linkage group of both parents. Linkage groups were assigned chromosome numbers using the basic local alignment search tool (BLAST) for querying the consensus tags of SNPs with M. *truncatula* reference genome, *M. truncatula* V4.1 genome as described in Li et al. (2014b).

Phenotypic Data Analysis and QTL Mapping

Alfalfa dormancy data was collected as regrowth height after clipping in the fall and winter. In the fall, canopy height data was taken at 4 weeks after clipping the plants on 21st September according to NAAIC protocol (Teuber et al., 1998). The plant height data at the Watkinsville (JPC) and Blairsville (BVL) locations were taken in subsequent days. The mild winter of 2016/2017 in Georgia allowed taking an early winter and late winter regrowth data. FD was phenotyped in the parents and the pseudo-testcross progeny in fall 2015, fall 2016 and winter (2016/2017). We collected two winter data sets in the season (2016/2017). Because of the mild winter in the Southeastern U.S.,

[6]https://www.kyazma.nl/docs/JM5Features.pdf

it is possible to phenotype seasonal dormancy in field conditions later than northern environments. The regrowth height data was converted to FDR based on the regrowth of 11 standard checks according to NAAIC protocol. FDR of the progeny were assigned based on a regression equation derived from the relationship between standard dormancy ratings of the check cultivars and their regrowth height in each environment. The standard regression lines for each location were established using average dormancy values of three years. The equations were derived for all growing environments and seasons using the Proc Reg procedure in SAS 9.4 (SAS Institute, 2004).

The dormancy phenotypic data consisted of two fall datasets (FD/2015 and FD/2016) for both locations, JPC and BVL, a winter data set collected in the first week of January (referred to as WD/2016 data set, and a second winter data set collected in last week of February (referred to as WD/2017 dataset). WH was evaluated on the F1 population and the two parents on a scale of 0–5 according to NAAIC protocol (McCaslin et al., 2003). Visual scores of winter damage were recorded after each freezing occurrence in winter months. In the case of mild winter, we visually scored plants once a month. Standard checks for winter survival were scored and photographed, and the images were used to guide in the scoring of F1 plants to minimize bias. The visual scores ranged from 1 to 5, where 1 indicates extremely winter-hardy genotypes and 5 indicates non-winter-hardy as described in NAAIC (McCaslin et al., 2003). Phenotypic data for all traits was analyzed using SAS 9.4 (SAS Institute, 2004). The least square (LS) means for all genotypes across environments and within individual environments were estimated for each dataset using PROC GLM (Li et al., 2015). For each trait, a linear additive model was used to perform the analysis of variance (ANOVA) for randomized block design:

$$\begin{aligned}\text{Trait value} = {} & \text{genotype} + \text{environment} \\ & + \text{block (environment)} \\ & + \text{genotype}*\text{environment} + \text{Error}\end{aligned}$$

where the trait value refers to the trait phenotypic value estimated by combining the effects of genotype, environment, block, and genotype by environment interaction. The block (environment) was considered random (Haggard et al., 2015). The LS means of all traits for both parents were also obtained within each and across environments. The LS means of the progeny were used as trait value for QTL detection. Pearson correlation coefficients (r) were calculated for both FD and WH trait means within each environment.

QTLs were detected using composite interval mapping (CIM) algorithm on Windows QTL Cartographer version 2.5 (Statistical Genetics, NC State University). The model and parameters used for CIM analysis were as described (Wan et al., 2017). We calculated trait-specific LOD scores using 1000 permutations at genome wide statistical threshold ($P \leq 0.05$). A QTL was declared significant when the peak LOD value exceeded a conservative LOD threshold of 3.0. In the case of more than one peak, multiple QTLs were declared if LOD values between the peaks falls below 3.0 for more than one contiguous segment for at least one dataset analyzed (Haggard et al., 2015).

The QTL detected in both parents for all traits were classified into two types, stable QTL and potential QTL. The QTL that were detected in more than one season, one environment, or across environments were considered stable QTL. QTL detected either in only one season or one environment were considered as potential QTL. The genomic positions of some major stable QTLs detected for each parental map were indicated on linkage maps using MapChart 2.3 (Voorrips, 2002). The QTL detected for dormancy on the linkage map of the dormant parent, 3010, were given name as "dorm." Similarly, the QTL detected for non-dormant parent, CW 1010, were named as "ndorm." The QTL for WH that were detected in the winter-hardy parent 3010 were named as "wh." The QTL for WH trait detected in the cold susceptible parent (CW 1010) were labeled "ws." The QTL span was delimited using LOD-1 confidence interval and the QTL were considered identical when the 1-LOD support intervals for QTL overlaps as in previous report (Haggard et al., 2015).

RESULTS

GBS and SNPs Discovery

A total of 100 Gb raw reads were generated using Illumina NextSeq High Output Flowcell (Illumina, Inc.) amounting to one billion usable paired-reads. Using the GBS-SNP-CROP pipeline for *de novo* SNP calling resulted in 4822 raw SNPs in the pseudo-testcross F1 population. There were 838 single dose allele (SDA) SNPs segregating in the maternal parent 3010 and 794 SNPs segregating in the paternal parent CW 1010. Among these, 423 SDA SNPs from 3010 to 220 SNPs from CW 1010 were filtered as high-quality SDA after chi-square test ($\alpha = 0.05$) for the segregation ratio of 1:1 (AB: AA). The SNPs obtained from this pipeline were identified with suffix MRG referring "Mock reference genome" followed by SNPs physical position with reference to MRG created using the reads of two parents.

Using the Tassel UNEAK pipeline, 500 million high-quality R1 reads were identified and processed using default parameter settings. A total of 65101 biallelic SNPs were identified. After filtering for missing data (<30%), 34122 (52.4%) SNPs were retained. Additionally, we removed 1625 loci with missing marker information in either parent retaining 32497 SNPs. Therefore, about 50% of the raw SNPs obtained from the UNEAK pipeline were filtered out in the initial screening because of missing data. Among the 32497 SNPs obtained from UNEAK, 4925 SNPs were single dose SNPs for 3010 parent and 2121 SNPs were identified as single dose SNPs for CW 1010 based on chi-square ($\alpha = 0.05$) test for segregation ratio (1:1).

Genetic Mapping

After merging the GBS SDA-SNP obtained from both Tassel UNEAK and GBS-SNP-CROP pipelines for each parent, we generated a total of 5348 SNPs for the maternal parent 3010 and 2340 SDA for the paternal parent CW 1010 (File S1). Further screening of the SDA loci on JoinMap 5.0[7] showed that three F1 individuals (ALF107, ALF255 and ALF302) did have several missing loci and were removed from further analysis leaving

[7]https://www.kyazma.nl/docs/JM5Features.pdf

181 progenies for mapping. Similarly, 26 loci from 3010 parent to 13 loci of CW 1010 parent were excluded from further analysis because they were identical. Consequently, 5322 SNPs from 3010 to 2327 SNPs from CW 1010 were used in genetic mapping.

SNPs from both parents were assembled into 32 linkage groups using LOD for independence of 12 or above in the 3010 parent and LOD of 10 or above in CW 1010. Based on the SNP physical positions determined from BLAST analysis, each four linkage groups or haplotypes of each parent were assigned to a corresponding chromosome of *M. truncatula* as described in **Figure 1** (Li et al., 2014b). Since the majority of the SDA SNPs (92% SDA of 3010 and 90% SDA of CW 1010) were obtained from UNEAK, we only used the SNPs from this pipeline to query the physical locations of markers. The consensus sequence of tag pairs of all mapped SNPs of both parents used to query in BLAST nucleotide.

Thirty-two linkage groups of the maternal parent 3010 consisting of 1837 SDA SNPs were assembled in a linkage map spanning about 2788.4 cM with an average marker density of 1.5 cM/SNP (**Table 1**, Figure S1). The number of SNPs per linkage group in 3010 parent ranged from 10 to 116. Most of the linkage groups ranged in length from 50 to 100 cM (**Table 1**, Figure S1). Marker density of the individual LG varied from 0.9 to 5.3 cM/SNP.

The 32 linkage groups of CW 1010 spanned 2127.5 cM with 1377 mapped markers (**Table 1**, Figure S2). The average marker density was 1.5 cM/SNP. The number of SNPs mapped on CW 1010 linkage groups varied from 7 to 139 (**Table 1**, Figure S2). Most of the CW 1010 LG had genetic lengths of 40 to 90 cM. The shortest LG (3D) was 26.7 cM and the longest LG (4B) was 121.9 cM. The individual group marker density in CW 1010 linkage map varied between 0.2 and 6.6 cM/SNP (**Table 1**).

BLAST analysis showed that alfalfa genetic loci mapped in this study were syntenic with *M. truncatula* reference genome (**Figures 1**, **2**). From 1837 SNPs mapped in the 3010 parent, 967 (53%) SNPs were aligned to *Medicago* reference genome with 84-100 % identity. On *Medicago* reference genome, 3010 SDA SNPs were aligned within range of 3.3 Kb to 56.4 Mb. The cut-off value used in BLAST analysis ranged from 2.06 E^{-06} to $2E^{-26}$. Similarly, 741 (53%) SNPs from the parent CW 1010 exhibited similarity with the *M. truncatula* genome with sequence identity of 85 to 100% using the same cut-off value as for 3010 SNPs. CW 1010 SNPs and *M. truncatula* genome similarity were obtained within a range of 0.5 Kb–56.3 MB.

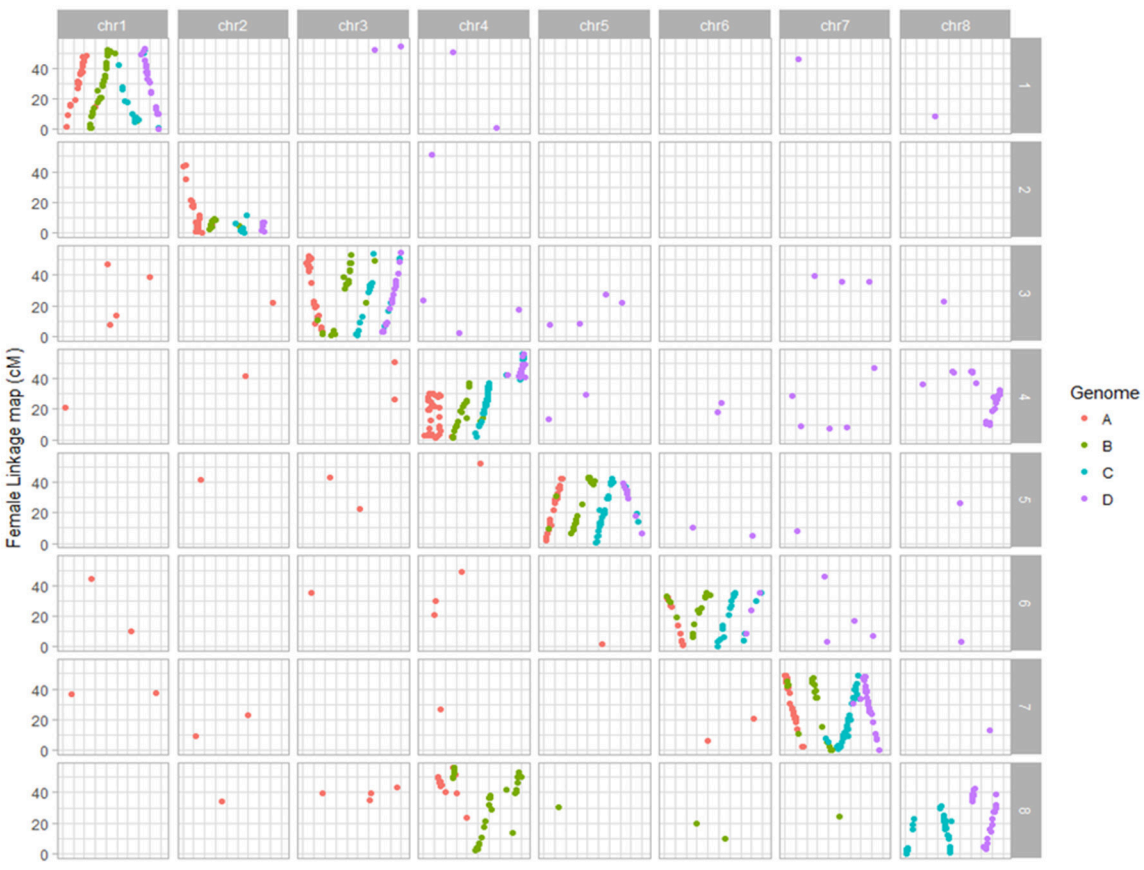

FIGURE 1 | Dot plot displaying the grouping pattern and positions of SNPs on 32 linkage groups of alfalfa 3010 linkage map. Of the 32 groups, each four homologs groups were assigned to a chromosome based on synteny with *Medicago truncatula* genome.

TABLE 1 | Distribution of SNP markers on 32 linkage groups of each of two alfalfa parents (CW 1010 and 3010).

Chr+	Homologs group	CW 1010			3010		
		No. SNPs	Length cM	MD¥	No. SNPs	Length cM	MD¥
1	A	84	56.7	0.7	59	108.5	1.8
1	B	77	73.1	0.9	75	95.1	1.3
1	C	22	71.3	3.2	30	109.6	3.7
1	D	12	40.3	3.4	58	91.6	1.6
2	A	53	114.5	2.2	48	88.9	1.9
2	B	61	78.5	1.3	20	32.7	1.6
2	C	26	76.1	2.9	10	53.3	5.3
2	D	11	73.1	6.6	16	13.6	0.9
3	A	39	70.7	1.8	79	91.9	1.2
3	B	41	50.9	1.2	45	106.7	2.4
3	C	27	62.1	2.3	51	85.9	1.7
3	D	22	26.7	1.2	55	85.7	1.6
4	A	7	48.75	7.0	64	82.1	1.3
4	B	56	121.9	2.2	53	101.2	1.9
4	C	82	72.6	0.9	70	85.5	1.2
4	D	31	71.6	2.3	116	110.4	1.0
5	A	65	88.2	1.4	65	83.1	1.3
5	B	49	84.3	1.7	30	112.3	3.7
5	C	45	47.1	1.0	72	79.2	1.1
5	D	9	51.1	5.7	44	91.2	2.1
6	A	26	55.5	2.1	64	90.3	1.4
6	B	24	42.1	1.8	80	87.9	1.1
6	C	48	91.8	1.9	69	84.9	1.2
6	D	74	91.2	1.2	42	84.2	2.0
7	A	64	83.1	1.3	74	92.3	1.2
7	B	71	46.9	0.7	37	97.3	2.6
7	C	33	82.3	2.5	86	112.1	1.3
7	D	11	42.7	3.9	82	75.9	0.9
8	A	139	28.1	0.2	44	97.9	2.2
8	B	7	46.5	6.6	76	92.1	1.2
8	C	35	66.1	1.9	65	77.8	1.2
8	D	26	71.7	2.8	58	87.2	1.5
Total		1377	2127.55	1.5	1837	2788.4	1.5

Number of markers, genetic length, and marker density for each homologs group are indicated. The homologs groups (A, B, C, and D) were assigned randomly within each chromosome based on BLAST search result. Chr+, Chromosomes; MD¥, Marker Density (cM/SNPs).

Dot plot maps constructed for each parent using mapped SNPs syntenic to *M. truncatula* clearly displayed the grouping of markers on the 32 LG groups to the corresponding eight *Medicago truncatula* chromosomes (**Figures 1**, **2**; Li et al., 2014b). In the female parent 3010, a translocation of a segment of chromosome four into eight was observed in all four homologs of chromosome eight. Three homologs (4B, 4C, 4D) of chromosome four also possessed segment from chromosome eight, indicating the reciprocal translocation between chromosomes four and eight (**Figure 1**; Li et al., 2014b). Such translocation was also observed for parent CW 1010, more clearly on haplotypes 4B, 4D, 8A, 8B, and 8D (**Figure 2**). Several other minor genome rearrangement events such as inversions and other translocations were present in several haplotypes in the maps of both parents. However, this study is focused on marker-trait association rather than structural analysis of the alfalfa genome.

Phenotypic Evaluation and Correlation Between Traits

Eight regression equations were generated to estimate FD of the mapping population at two locations in fall 2015 (FD/2015), fall 2016 (FD/2016), winter 2016 (WD/2016) and winter 2017 (WD/2017). The regression models suggested that the relationship between standard FDR and canopy regrowth height of alfalfa checks was strong and positive. The regression coefficients (R^2) ranged from 0.37 for fall 2015 at BVL to 0.73 of winter 2016 dormancy of JPC environment with six out of the eight regression models having coefficient of determinations

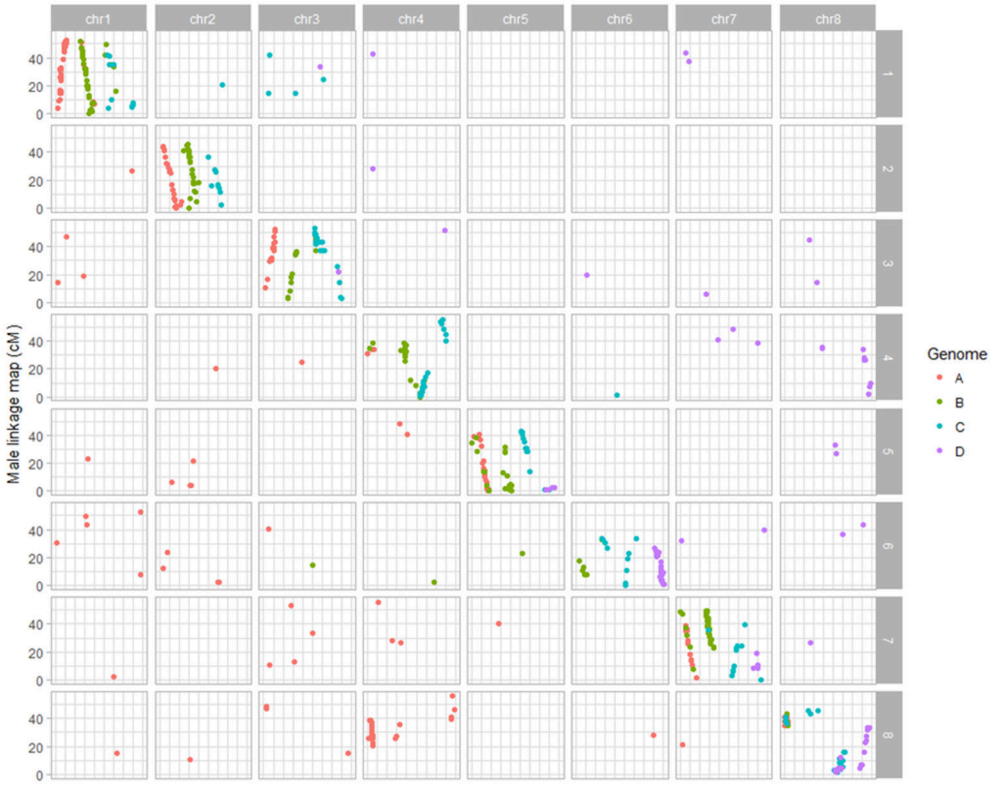

FIGURE 2 | Dot plot displaying the grouping pattern and positions of SNPs on 32 linkage groups of alfalfa CW 1010 linkage map. Of the 32 groups, each four homologs groups were assigned to a chromosome based on synteny with *Medicago truncatula* genome.

FIGURE 3 | Image showing dormant (left) and non-dormant (right) progeny rows from the pseudo-testcross F1 population (3010 × CW 1010) after frost occurrence in March 2017 at the JPC environment. Frost damage symptoms are clearly visible on the non-dormant progeny.

$R^2 > 0.50$. The regression coefficients of the winter rating were higher than fall ratings at both sites.

There were significant differences between the genotypes ($P \leq 0.01$) and significant G x E for FD ratings in most dates except for FD/2015 data. Because of the significant G x E, the LS means for each trait were estimated separately for each location (**Table 2**). The R^2 values for each trait, derived from the ANOVA, varied from 0.59 to 0.87 indicating a good fit of the data to the respective linear model for individual tests (**Table 2**). Dormancy measured in winters (WD/2016 and WD/2017) were highly correlated to each other than the dormancy measured in fall (FD/2015 and FD/2016) in JPC environment (**Table 3**). The FD trait for winter 2017 rating showed the highest R^2 (0.87) and fall 2016 exhibited the least R^2 (0.59). The LS means estimated for traits and parents revealed the presence of transgressive segregation on both sides of the parents for both FD and WH traits (**Table 2**). Some past studies also reported the presence of transgressive segregation in alfalfa pseudo-testcross progeny for such traits (Li et al., 2015).

There were significant differences between the genotypes ($P \leq 0.01$) for WH and significant G x E. Because of the significant G x E, the LS means were estimated for each location in addition to across locations (**Table 2**). The R^2 values for WH, varied from 0.63 to 0.79 indicating a good fit of the data to the respective linear model for individual tests (**Table 2**). The LS means estimated for F1 progenies and parents revealed the presence of transgressive segregants for WH on both sides of the parents (**Table 2**).

Pearson's correlation coefficient using trait means showed moderate degrees of correlation between all traits at both locations (**Tables 3**, **4**). Overall, there were stronger positive correlations between dormancy and WH when dormancy was

TABLE 2 | Phenotypic means of F1 progeny and parents for FDR and WH scores.

Trait/Year	Location	F1 phenotype range (LS means)	LS means (3010)	LS means (CW 1010)	(ANOVA, F1) R^2
FD/2015	JPC	2.3–9.0	6.4	7.5	0.73
FD/2016	JPC	1.9–7.1	4.6	4.7	0.59
WD/2016	JPC	2.0–7.3	2.3	5.7	0.82
WD/2017	JPC	1.2–8.8	2.7	5.3	0.87
FD/2015	BVL	2.2–9.1	5.7	6.6	0.71
FD/2016	BVL	2.89–10.6	4.5	6.5	0.61
WD/2016	BVL	2.5–8.4	4.5	7.9	0.77
WD/2017	BVL	1.6–9.4	4.8	7.5	0.73
FD/2015	JPC & BVL	2.6–8.5	6.1	7.0	0.67
FD/2016	JPC & BVL	3.4–8.0	4.5	5.6	0.62
WD/2016	JPC & BVL	2.3–7.6	3.8	6.8	0.80
WD/2017	JPC & BVL	1.6–8.3	3.7	6.5	0.80
WH/2015	JPC	1–3.2	1	2	0.70
WH/2016	JPC	1.2–5.0	2.5	4.2	0.79
WH/2017	JPC	1–4	1	2.5	0.79
WH/2015	BVL	1–5	2	2.7	0. 71
WH/2016	BVL	1–4	1.3	4	0.63
WH/2017	BVL	1–4.9	1.7	4.5	0.70
WH/2015	JPC & BVL	1–3.4	1.5	2.3	0.79
WH/2016	JPC & BVL	1.4–4.1	1.8	4.1	0.78
WH/2017	JPC & BVL	1–4.3	1.4	3.5	0.78

Coefficient of determination (R^2) are indicated for each data set. FD, dormancy assessed in fall; WD, dormancy assessed in winter; WH, winter hardiness.

TABLE 3 | Phenotypic correlations (r) among traits based on data collected for Watkinsville (JPC) environment on a pseudo-testcross F1 population (3010 × CW 1010).

	FD/2015	FD/2016	WD/2016	WD/2017	WH/2015	WH/2016	WH/2017
FD/2015		0.50**	0.62**	0.60**	0.39**	0.52**	0.57**
FD/2016			0.39**	0.43**	0.12^{NS}	0.31**	0.50**
WD/2016				0.92**	0.22**	0.65**	0.80**
WD/2017					0.23**	0.71**	0.85**
WH/2015						0.16*	0.10^{NS}
WH/2016							0.68**
WH/2017							

Dormancy was assessed twice in the fall (FD/2015 and FD/2016) and twice in the winter (WD/2016 and WD/2017). The WH data was collected in three consecutive winters (WH/2015, WH/2016 and WH/2017).
$^*P \leq 0.05$, $^{**}P \leq 0.01$, ^{NS}non-significant.

assessed in winter compared to fall assessment (**Figure 3**, **Tables 3**, **4**). In the JPC environment, the coefficient of correlations between dormancy rating and WH ranged from 0.12 to 0.57 when dormancy was assessed in the fall, while it ranged from 0.22 to 0.85 when dormancy was assessed in winter (**Table 3**). The same trend was observed in the BVL location. The coefficient of correlations between dormancy rating and WH ranged from 0.16 to 0.50 when dormancy was assessed in the fall, while it ranged from 0.22 to 0.57 when dormancy was assessed in winter (**Table 4**).

QTL Mapping of FD and WH

Within the 32 homologs of the 8 alfalfa chromosomes, we detected 45 significant ($P \leq 0.05$) QTLs for FD and 35 QTLs for WH on both male and female linkage maps (**Tables 5**–**8**). Most of the QTLs detected using phenotypic data across environments matched QTLs detected for individual environments with slight variation in their LOD magnitude and interval. Seven QTLs for dormancy and three QTLs for WH detected across environments were exclusively different from QTLs detected for individual environments indicating a potential effect of G x E on trait values (**Table 8**).

Fall Dormancy (FD)

Seven stable QTL for FD were identified in the dormant parent 3010. These QTLs were consistently and repeatedly detected across data sets within overlapping 1-LOD support intervals (**Table 5**, **Figure 4**). The seven dormancy QTL for

TABLE 4 | Phenotypic correlations among traits based on data collected at the BVL location on a pseudo-testcross F1 population (3010 × CW 1010).

	FD/2015	FD/2016	WD/2016	WD/2017	WH/2015	WH/2016	WH/2017
FD/2015		0.42**	0.6**	0.58**	0.16*	0.16*	0.33**
FD/2016			0.6**	0.64**	0.25**	0.43**	0.50**
WD/2016				0.92**	0.24**	0.27**	0.57**
WD/2017					0.22**	0.25**	0.51**
WH/2015						0.34**	0.46**
WH/2016							0.54**
WH/2017							

Dormancy was assessed twice in the fall (FD/2015 and FD/2016) and twice in the winter (WD/2016 and WD/2017). Winter hardiness (WH) data were collected in three consecutive winters (WH/2015, WH/2016 and WH/2017).
$^{*}P \leq 0.05$, $^{**}P \leq 0.01$, ^{NS}non-significant.

TABLE 5 | Stable QTLs for alfalfa FD detected in an F1 (3010 × CW 1010) pseudo-testcross population based on phenotypic data assessed in fall and winter at two locations.

Parent	QTL code	Chr	Year/Location	Peak markers	Peak LOD	R^2	Allele dir.	LSI (cM)	Flanking markers
3010	dorm1	1A	(π1), π4, β2, $1, $2	TP995	10.9	0.16	(–)	90.6–92.9	TP995–TP78651
3010	dorm2	1A	π1, π3, (π4), β3, β4, $1, $4	TP86274	6.2	0.11	(–)	98.2–104	TP65855–TP86274
3010	dorm3	7A	π1, (β3), β4, $1	TP24733	6.2	0.11	(–)	36.9–38.8	TP55743–TP34483
3010	dorm4	4C	(π1), β1, $1	TP56893	7.5	0.11	(–)	58.6–61.0	TP55689–TP56893
3010	dorm5	7B	(π4), β3	TP31689	4.1	0.08	(–)	34.6–48.7	TP31689–TP33803
3010	dorm6	7A	(π1) π3, π4, β3, β4, $1, $3	TP69889	5.5	0.07	(–)	47.3–52	TP59349–TP71458
3010	dorm7	3A	β1, β2, (β3)	TP32327	3.8	0.06	(–)	25.8–26.3	TP3895–TP54529
CW 1010	ndorm1	8D	(π1) π3, π4, β1, $1	TP2543	9.9	0.13	(+)	44.6–46.3	TP2543–TP88682
CW 1010	ndorm2	7C	π3, (π4), β3	TP38417	9.0	0.12	(+)	46.9–51.1	TP38417–TP54614
CW 1010	ndorm3	5B	(β1), $1	MRG_32692305	4.4	0.08	(+)	15.2–17.9	TP79886–MRG_32692305
CW 1010	ndorm4	8D	π3, (π4), β1, $3	TP24024	6.1	0.07	(+)	53.7–54.8	TP24024–TP25406
CW 1010	ndorm5	1B	π1, (β4), $4	TP57411	5.1	0.07	(+)	14.2–15.3	TP63551–TP32288
CW 1010	ndorm6	5B	(π1), β2	TP26770	5.5	0.06	(+)	48–49.5	MRG_37364973–TP26770
CW 1010	ndorm7	8D	π3, (π4)	TP67491	3.9	0.06	(+)	63.9–65.4	TP67491–TP71707
CW 1010	ndorm8	1B	π1, π3, β1, (β4), $4	TP41786	3.2	0.05	(+)	21.1–23.7	TP83000–TP7086
CW1011	ndorm9	1B	(π3), π4, β3	TP52371	4.2	0.05	(+)	31.9–33.8	TP35547–TP23336
CW 1010	ndorm10	7B	π3, (π4)	TP14107	3.8	0.04	(+)	23.3–24.2	TP14107–TP9019
CW 1010	ndorm11	7B	π3, (π4)	TP7325	3.1	0.03	(+)	11.7–12.7	TP32866–TP42483

Seven QTLs for 3010 and eleven QTLs for CW 1010 were mapped on respective genetic linkage maps for phenotypic datasets of more than one environment and/or year.
The symbol with bracket in the column "year/location" indicates the source dataset from which the other parameters in the same row were pulled.
JPC environment dormancy data for the period: π1 = 2015, fall; π2 = 2016, fall; π3 = 2016, winter; π4 = 2017, winter.
BVL environment dormancy data for the period: β1 = 2015, fall; β2 = 2016, fall; β3 = 2016, winter; β4 = 2017, winter.
Across environment dormancy for the period: $1 = 2015, fall; $2 = 2016, fall; $3 = 2016, winter; $4 = 2017, winter.
Chr., Chromosome; Dir., Direction; LSI, 1-LOD support interval in cM unit.

3010 (dorm1, dorm2, ……., dorm7) were detected on homologs 1A, 3A, 4C, 7A and 7B. Another 21 potential QTLs (dorm8, dorm9,.....…, dorm28) were detected in various homologs of 3010 chromosomes: 1, 2, 3, 4, 5, 6, and 7 (**Tables 5**, **7**, **8**). Five of the seven stable QTLs were detected also across environments. The most important dormancy QTL (dorm1) for 3010 parent ($R^2 = 0.16$) was detected on homolog 1A and was located at 90.6–92.9 cM. The same homolog harbors another QTL (dorm2) with a LOD value of 6.2 and a peak at the interval 98.2–104 cM (**Table 5**, **Figure 4**). Besides these two stable QTLs for 3010 parent, other potential QTLs (dorm 8, dorm 12, dorm 27 and dorm 28) were detected on homologs of 3010 chromosome 1, suggesting that this chromosome is important for the dormancy trait.

Further, dorm1 and dorm2 QTLs of 3010 parent mapped on homolog 1A were located within similar genomic location of alfalfa fall height QTLs (92a and 104a) in the WISFAL-6 cultivar reported previously (Li et al., 2015). (Li et al., 2015) mapped fall height on eight alfalfa linkage groups assigned using eight *M. truncatula* chromosomes. Unlike 3010 QTLs dorm1 and dorm2 detected in our study, WISFAl-6 dormancy QTL of LG1 had positive effect on trait value because the source parent WISFAL-6 had higher FD levels (Li et al., 2015).

TABLE 6 | Stable QTLs for alfalfa WH identified in an F1 (3010 × CW 1010) pseudo-testcross population based on phenotypic data assessed in three consecutive winters at two locations.

Parent	QTL code	Chr.	Year/Location	Peak marker	Peak LOD	R^2	Allele dir.	LSI (cM)	Flanking markers
3010	wh1	1A	(λ2), λ4, ϕ3	TP995	7.1	0.13	(+)	90.8–93.2	TP995–TP6492
3010	wh2	7A	(λ3), ψ1, ϕ1	TP24733	7.1	0.12	(+)	37.5–39.0	TP55743–TP34483
3010	wh3	1A	(λ2), ϕ3	TP65855	5.4	0.11	(+)	98.2–104.5	TP65855–TP86274
3010	wh4	8A	(ψ2), ϕ2	TP10810	5.1	0.09	(–)	74.1–74.8	TP58070–TP10810
3010	wh5	3A	(ψ3), ϕ3	TP71671	5.1	0.09	(+)	47.6–50.1	TP52425–TP67563
3010	wh6	1C	(λ1), ψ1, ϕ1	TP37162	4.1	0.07	(+)	96.3–99	TP37162–TP57104
3010	wh7	7A	(λ1), λ3, ϕ1	TP58371	3.2	0.07	(+)	26.4–29.3	TP58371–TP34795
3010	wh8	4C	(λ2), ψ1, ϕ1	TP2323	3.5	0.06	(+)	27.1–31.1	TP6532–TP4218
CW 1010	ws1	7C	λ2, (λ3), ψ1	TP54614	9.8	0.14	(–)	48.4–51.4	TP38417–TP54614
CW 1010	ws2	8D	(λ3), ψ3, ϕ3	TP52817	7.5	0.10	(–)	40.8–43.6	TP52817–TP46951
CW 1010	ws3	7A	(ψ3), ϕ2	TP71946	5.7	0.10	(+)	8.6–17.6	TP78230–TP71946
CW 1010	ws4	7A	ψ2, ϕ2	TP81779	5.3	0.10	(+)	31–32.1	TP16325–TP70376
CW 1010	ws5	8D	(λ2), ϕ3	TP2543	5.2	0.08	(–)	44.6–45.8	TP2543–TP6748
CW 1010	ws6	7B	(λ3), ψ1, ϕ1	TP87913	5.6	0.07	(–)	35.9–37	TP87913–TP85708
CW 1010	ws7	7B	(λ3), ψ1, ϕ3	MRG_41805356	4.9	0.07	(–)	24.5–25.3	TP49165–TP74211
CW 1010	ws8	8D	(λ1), λ2	TP8426	3.3	0.06	(–)	58.6–63.1	TP69982–MRG_7512818
CW 1010	ws9	1A	(λ3), ϕ3	TP40020	3.4	0.04	(+)	47.2–48	TP60690–TP40020

Eight QTLs from 3010 to 9 QTLs from CW 1010 were mapped on respective genetic linkage maps using phenotypic datasets of more than one environment and/or year. The symbol with bracket in the column "year/location" indicates the source dataset from which the other parameters in the same row were generated.
JPC environment winter hardiness data for the period: λ1 = 2015; λ2 = 2016; λ3 = 2017.
BVL environment winter hardiness data for the period: ψ1= 2015; ψ2 = 2016; ψ3 = 2017.
Across environment winter hardiness data for the period: ϕ1 = 2015; ϕ2 = 2016; ϕ3 = 2017.

Another two stable dormancy QTLs from 3010 were detected on chromosome 7A (R^2 = 0.07–0.11). Homologs of this chromosome also harbor QTL dorm3, dorm5, dorm6, and dorm13 (**Tables 5**, **7**). The QTL dorm6 also falls within similar genomic regions of a fall height QTL in LG 7 as reported previously (Li et al., 2015). Two other potential QTL on homologs of this chromosome at LOD = 3.1 include dorm21 and dorm22 that were located on two homologs (7D and 7C) of 3010 chromosome7. Further, the homologs of 3010 chromosomes: 3, 4 5, 6, and 2 also harbored significant ($P \leq 0.05$) QTLs for dormancy (**Tables 5**, **7**). All dormancy QTLs detected on 3010 parent had negative effects on trait value since the parent was a dormant type.

In the CW 1010 parent, 11 stable QTLs and six potential QTLs for FD were detected. All the stable QTLs for CW 1010 were detected on homologs of the chromosomes 1, 7, 5, and 8 (**Table 5**). The CW 1010 chromosome 8 exhibited a broader QTL peak extending from ∼44 to ∼ 66 cM. However, there is the possibility of presence of more than one QTL in the region because of decreasing LOD value between multiple QTL peaks. Therefore, we reported three different stable QTLs for this region to ensure the accuracy of QTL and corresponding phenotypic values of markers in the region. A past study (Li et al., 2015) also reported a QTL (46a) positively affecting fall plant height in the same genomic region (40–56 cM) of LG 8 of the alfalfa cultivar ABI408 providing more supportive evidence for this QTL. The QTL (ndorm1) from CW 1010 with positive effect on the trait value (R^2 =0.13) was detected on homolog 8D at the interval 44.6–46.3 cM (**Table 5**). Other stable QTLs for dormancy detected from CW 1010 parent including ndorm2, ndorm3, ndorm4 and ndorm6, shared common genomic regions with QTLs for fall height reported previously (Li et al., 2015). However, two CW 1010 dormancy QTLs, ndorm3 and nodorm6, had contrasting effects in trait value than previously reported QTLs of the corresponding linkage groups. Of 17 total CW 1010 dormancy QTLs, 16 QTLs had additive effect in favor of trait value and one potential QTL (ndorm13) had negative effect on trait value (**Table 7**).

Although we detected dormancy QTLs for both 3010 and CW 1010 parents in most of the datasets, a higher number of stable QTLs were repeatedly detected using winter dormancy data compared to fall data of all environments (**Tables 4**, **5**, **7**). For the 3010 parent, only two stable QTLs were observed for each 2015 and 2016 fall datasets of BVL location, while two and four stable QTLs were detected for JPC winter datasets WD/2016 and WD/2017, respectively (**Table 5**). Only two stable dormancy QTLs for 3010 parent (dorm1 and dorm4) were repeatedly detected in fall. However, four stable dormancy QTLs (dorm2, dorm3, dorm5 and dorm6) were repeatedly detected in more than one winter data (**Table 5**). For CW 1010 parent, out of 11 stable dormancy QTLs, only six stable QTLs were identified for all FD data of both locations, and nine stable QTLs were detected for winter datasets of both locations (**Table 5**). Five and three potential QTLs were detected only in cross environments analysis for both parents 3010 and CW 1010 indicating the presence of G x E (**Table 5**). There was also slight shift of QTL peaks identified in across environments compare to those QTL identified in individual environment data sets.

TABLE 7 | Potential QTLs for dormancy and WH identified in an F1 pseudo-testcross (3010 × CW 1010) population.

Trait	Parent	QTL code	Chr.	Year/Location	Peak marker	Peak LOD	R^2	Allele dir.	LSI (cM)	Flanking markers
FD	3010	dorm8	1A	β1	TP73186	7.7	0.12	(–)	69.1–71.3	TP73186–TP70400
FD	3010	dorm9	5A	π1	MRG_28485316	5.3	0.08	(–)	36.1–45.8	MRG_28485316–TP63204
FD	3010	dorm10	6D	π3	MRG_2402742	4.7	0.08	(–)	16.7–24.0	MRG_2402742–TP18699
FD	3010	dorm11	4D	π2	TP64707	4.5	0.08	(–)	101.3–101.9	TP64707–TP61536
FD	3010	dorm12	1C	π2	TP78612	3.4	0.08	(–)	62–70.3	TP78612–TP68882
FD	3010	dorm13	7B	β4	TP43449	4.2	0.07	(–)	15.4–18.5	TP14416–TP25746
FD	3010	dorm14	2C	π2	TP29084	3.8	0.07	(–)	32.7–39.5	TP29084–TP82709
FD	3010	dorm15	4C	π4	TP64526	4.2	0.06	(–)	16.3–18.8	MRG_18042076–TP64526
FD	3010	dorm16	2B	π3	TP69826	4.0	0.06	(–)	25.3–28.7	TP78664–TP59834
FD	3010	dorm17	5A	β1	TP63107	3.8	0.06	(–)	24.1–27.6	TP89078–TP46688
FD	3010	dorm18	3A	β1	TP32175	3.3	0.05	(–)	9–15	TP32175–TP44970
FD	3010	dorm19	3D	π3	TP67190	3.2	0.05	(–)	5.6–13.3	TP67190–TP58690
FD	3010	dorm20	3A	β2	TP32136	3.1	0.05	(–)	18.6–22	TP48316–MRG_4754683
FD	3010	dorm21	7D	β4	TP32437	3.1	0.05	(–)	5.2–7.7	TP79530–TP53493
FD	3010	dorm22	7C	π2	MRG_30285700	3.1	0.05	(–)	19.9–22.4	MRG_30285700–TP49176
FD	3010	dorm23	5D	π1	TP31552	3.2	0.04	(–)	36.5–41.6	TP31552–TP28126
FD	CW 1010	ndorm12	4D	β4	TP32802	6.7	0.12	(+)	64.5–66.1	TP32802–TP5506
FD	CW 1010	ndorm13	7A	β2	TP63954	5.9	0.10	(–)	28.1–29.4	TP63954–TP87998
FD	CW 1010	ndorm14	4B	β4	TP11836	3.8	0.05	(+)	71.2–72.4	TP11836–TP10328
FD	CW 1010	ndorm15	7A	β1	MRG_31595966	3.2	0.04	(+)	60.1–61.4	MRG_7180813–TP51152
WH	3010	wh9	7C	ψ2	TP84244	4.9	0.09	(–)	72.7–73.9	TP84244–TP44147
WH	3010	wh10	7C	λ1	TP74326	4.5	0.09	(+)	106.5–110.6	TP74326–TP30485
WH	3010	wh11	8B	ψ3	TP34659	3.9	0.08	(+)	43.3–47	TP24160–TP34659
WH	3010	wh12	8D	ψ3	TP15842	3.7	0.07	(+)	77–80.1	TP33611–TP15842
WH	3010	wh13	3B	λ1	TP63723	3.2	0.06	(+)	44.2–50.3	TP63723–TP46610
WH	3010	wh14	3D	ψ2	TP26775	3.3	0.06	(+)	28.5–33.6	TP88373–TP16429
WH	3010	wh15	2B	λ1	TP6025	3.1	0.05	(+)	8.7-11.1	TP19047–TP6025
WH	3010	wh16	2C	λ3	TP29084	3.2	0.05	(+)	33.5–40	TP29084–TP82709
WH	CW 1010	ws10	4D	λ1	TP88199	10.1	0.18	(–)	33.9–37.8	TP54779–TP88199
WH	CW 1010	ws11	1B	ψ2	TP66690	3.8	0.076744	(+)	41.8–43.1	TP64641–TP66690
WH	CW 1010	ws12	5A	ψ3	TP33164	3.9	0.066623	(–)	62.7–63.6	TP33164–TP30048
WH	CW 1010	ws13	1B	ψ2	TP7086	3.1	0.068061	(–)	22.9–24.1	TP7086–TP65701
WH	CW 1010	ws14	8A	ψ3	MRG_12811807	3.3	0.055791	(–)	20.1–20.3	MRG_12811807–TP29734
WH	CW 1010	ws15	1A	λ3	TP81842	3.7	0.047139	(+)	54.6–55.8	TP6332–TP81842
WH	CW 1010	ws16	6A	λ3	TP60069	3.4	0.044139	(–)	52.5–54.7	TP60069–TP5275

*These QTLs were detected only for a single location and a single year. The symbols used in this table have exactly same abbreviations as given for **Tables 5**, **6**.*

Winter Hardiness (WH)

Eight stable and 11 potential QTLs were identified from the 3010 parent for WH trait (**Tables 6**–**8**). The stable QTLs (wh1, wh2, ……, wh8) were detected on homologs of chromosomes: 1, 3, 4, 7, and 8, and 11 potential QTLs (wh9, wh10, ……, wh20) were detected on homologs of chromosomes: 2, 3, 4, 7 and 8. The QTL wh1 on homolog 1A (position 90–93.2 cM) has the largest R^2 (0.13) followed by QTL wh2 on homolog 7A (0.12) and wh3 on homolog 1A (0.11) (**Table 6**). The wh1 QTL was located in the same genomic region of previously identified QTL (100a) in WISFAL-6 alfalfa, but with opposite effect (Li et al., 2015). Similarly, other two potential QTLs, wh10 on homolog 7C and wh16 of homolog 2C were also detected within similar genomic locations of previously identified QTLs for winter injury for ABI408 (LG7, 109a) and WISFAL-6 cultivar (LG2, 36b), respectively (Li et al., 2015). With the exceptions of a stable QTL (wh4) on homolog 8A and one potential QTL (wh9) on homolog 7C, which had negative effects (−) on WH, all other WH QTLs detected on 3010 possessed positive effects (+) on WH (**Tables 6**–**8**).

Sixteen (nine stable and seven potential) QTLs for WH were detected for the winter susceptible parent CW 1010 and were coded as (ws1, ws2, …, ws16) (**Tables 6**, **7**). Major stable QTLs for WH in CW 1010 were detected on homologs of chromosomes 1, 7 and 8. The QTL ws1 had the highest R^2 (0.14) followed by ws2, ws3, and ws4 (R^2 = 0.10). However, contrary to ws1 and ws2, the ws3 and ws4 QTLs had positive effects (+) on WH (**Table 6**). The three stable QTLs, detected for WH trait on

TABLE 8 | Potential QTLs for dormancy and WH identified in a pseudo-testcross F1 population (3010 × CW 1010) using phenotypic data generated across two environments (JPC and BVL).

Trait	Parent	QTL code	Chr.	Year	Peak marker	Peak LOD	R^2	Allele dir.	LSI (cM)	Flanking markers
FD	3010	dorm24	3A	$4	TP67563	5.4	0.09	(−)	49.4–51.8	TP71671–TP76041
FD	3010	dorm25	3D	$4	MRG_38650252	3.6	0.09	(−)	70.1–74.4	MRG_31477229–TP57603
FD	3010	dorm26	3D	$2	TP16817	4.3	0.09	(−)	76.3–79.5	TP16817–TP5092
FD	3010	dorm27	1C	$1	TP32721	4.8	0.07	(−)	81.8–85.1	TP32721–TP40620
FD	3010	dorm28	1A	$3	TP46942	3.6	0.06	(−)	72–73	TP72089–TP73780
FD	CW 1010	ndorm16	4D	$3	TP69818	3.1	0.09	(+)	8.3–10.7	TP69818–TP82286
FD	CW 1010	ndorm17	4D	$2	TP80681	3.9	0.08	(+)	19.7–23.5	TP80681–TP83595
WH	3010	wh17	4C	ϕ1	TP81375	4.6	0.10	(+)	37.6–39.5	TP80478–TP81375
WH	3010	wh18	3D	ϕ3	TP43991	5.1	0.09	(+)	58.9-63.2	TP50808–TP87713
WH	3010	wh19	7A	ϕ3	TP71458	3.7	0.06	(+)	50.5–52	TP15177–TP71458

Across environment dormancy:
$1 = 2015, fall; $2 = 2016, fall; $3 = 2016, winter; $4 = 2017, winter.
Across environment winter hardiness:
ϕ1 = 2015; ϕ2 = 2016; ϕ3 = 2017.
Chr., Chromosome; Dir., Direction; LSI, 1-LOD support interval in cM unit.

homolog 8D, appeared in a single span for JPC data. However, for BVL and across environments, the QTL on 8D was separated into three different QTLs and were reported as such. Of the total 16 WH QTLs in CW 1010, ws12 on homolog 5A and ws16 on homolog 6A were detected within similar genomic regions reported previously for winter injury in the cultivars WISFAL-6 and ABI408 (Li et al., 2015). Most of the WH QTLs for CW 1010 have negative effects (−) on the trait except QTL ws3, ws4, ws9, ws11 and ws15 (**Tables 6**, **7**).

Association Between Dormancy and Winter Hardiness

Among the seven stable dormancy QTLs detected in the 3010 the dorm1 and dorm2 on homolog 1A shared the same genomic location as WH stable QTLs wh1 and wh3, respectively (**Figure 4**). Similarly, the dormancy QTL dorm3 overlapped with WH QTL wh2 on 7A with <1 cM shift (**Figure 4**). Another stable QTL dorm6 also shared the same genomic location with a potential winter hardiness QTL wh20 on 7A. Three stable QTLs (dorm4, dorm5, and dorm7) in the 3010 parent were unique and located on different chromosomes than winter hardiness. Of the 21 potential dormancy QTLs in the 3010 parent, except dorm14 and dorm24, other 19 were also located in different genomic regions from the QTLs of WH (**Tables 5**–**8**). Therefore, of the 28 dormancy QTLs detected in 3010 parent, 22 QTLs were located in separate genomic positions than the QTLs of WH indicating differences of two traits at the genomic level.

In the CW 1010 parent, the stable QTLs ndorm1 and winter hardiness QTL ws5 shared the same genomic location on homolog 8D (**Figure 5**). Likewise, the QTL ndorm2 and ws1 also reside on same position on homolog 7C. Another stable dormancy QTL of CW 1010 ndorm8 shares genomic location with a potential QTL for winter hardiness ws13 on homolog 1B. Moreover, a stable dormancy QTL ndorm7 and a winter hardiness QTL ws8 were located at nearby positions on the homolog 8D (**Figure 5**). Similarly, ndorm10 and ws7 also shared partial genomic position on homolog 7B of CW 1010 (**Figure 5**). Other stable dormancy QTLs from the parent CW 1010 such as ndorm3, ndorm4, ndorm5, ndorm6, ndorm9, and ndorm11 were located in separate genomic positions than those QTLs for WH. All potential QTL detected for CW 1010 for dormancy as well as for WH were also located in distinct genomic regions (**Tables 5**–**8**). Therefore, of the 17 dormancy QTLs detected in CW 1010 parent, 12 QTLs were located in separate genomics regions than the QTLs for WH (**Tables 5**–**8**).

DISCUSSION

Segregation and Phenotypic Relationship Between Traits

The regrowth height and dormancy ratings of the NAAIC standard checks used in this study showed a better fit to the regression model in winter height data with R^2 up to 0.73 compared to the height data taken in fall ($R^2 \sim 0.50$) (**Tables 3**, **4**). The fall data is collected around the third week of October according to NAAIC protocol. In southern environments, temperatures at this time of the year are still very favorable for active growth of alfalfa. Fall 2016 was a very unusual season in Georgia because of the historical drought in the region (Adhikari et al., 2018), which led to a very limited growth and erratic regrowth after clipping in both experimental sites. This could be the main reason that the two parents did not exhibit differences in their heights for this season (**Table 2**) and a few QTLs were detected based on the 2016 fall data. The positive R^2 for the regression of standard checks regrowth height data on their FDR for all seasons, suggests that that determining dormancy of alfalfa genotypes using regrowth height after clipping is a reliable approach.

The non-dormant parent CW 1010 exhibited slightly lower dormancy level (4.7–7.9), than it supposed to be in most of the years. The 3010 parent mostly exhibited the expected dormancy level between 2.3and 6.4 (**Table 2**). Such fluctuations of estimated dormancy level from their standard dormancy are primarily due to environmental and seasonal variations. The largest deviations

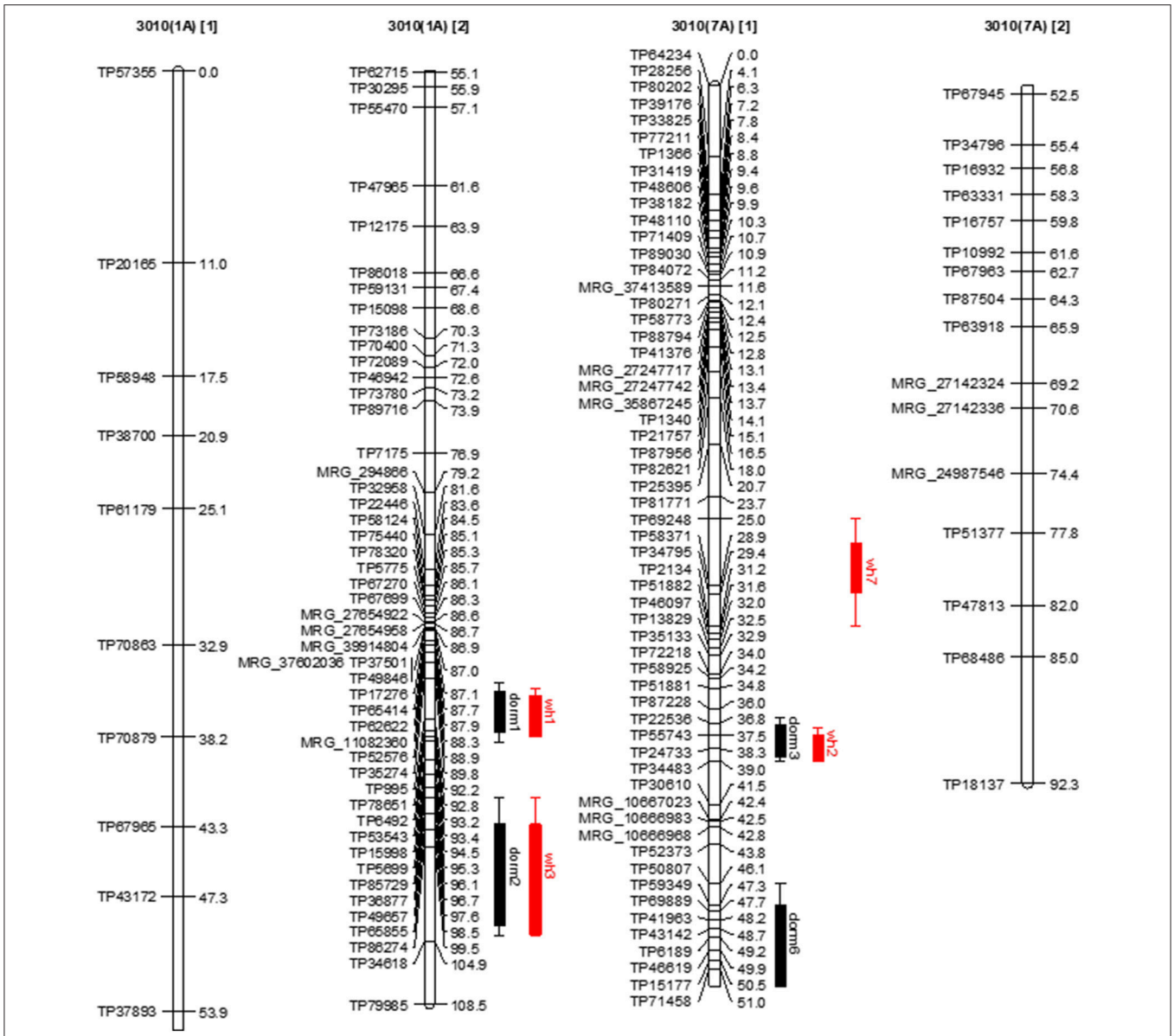

FIGURE 4 | Dormancy (black bar) and WH (red bar) stable QTLs mapped on linkage maps of homolog 1A **(left)** and 7A **(right)** for 3010 parent. The QTL bars have two intervals, an inner (1-LOD support) interval and an outer (2-LOD support) interval, where the rectangle represents inner interval and the line represents the outer. Some stable QTLs for dormancy were co-localized with WH in the same genomic regions.

from the standard dormancy ratings of the parents were observed in the data collected in fall season, suggesting that rating alfalfa dormancy in the Southeastern U.S.A based on regrowth height after clipping in third week of September is not reliable. Previous reports also suggested that FD assignment should be done in sites where the cultivars are broadly adapted (Teuber et al., 1998). Dormancy assessment in winter months showed a better approximation of the expected values with less variation compared to the fall assessment. Winter assessed dormancy also showed a better repeatability in both locations as suggested by the high correlation between WD/2016 and WD/2017 ($r = 0.92$, $p < 0.01$) (**Tables 3**, **4**). The LS means of dormancy ratings of the F1 progeny varied from 1.2 to 10.6 and showed transgressive segregation around the parental values (**Table 2**).

WH rating scores for 3010 parent varied from 1.4 to 1.5 across locations, which is within the range of the known score 2 for this cultivar. For the non-winter-hardy parent CW 1010, the scores varied between 2.3 and 4.1 across locations (**Table 2**). The F1 progeny also varied in their WH level with the largest differences observed in 2017 (WH/2017) at BVL location. Transgressive segregants were observed for both dormancy and WH similar to previous alfalfa reports (Li et al., 2015). This suggests the presence of complementary alleles for both traits in each of the parents (Li et al., 2015). The moderate positive correlations (r)

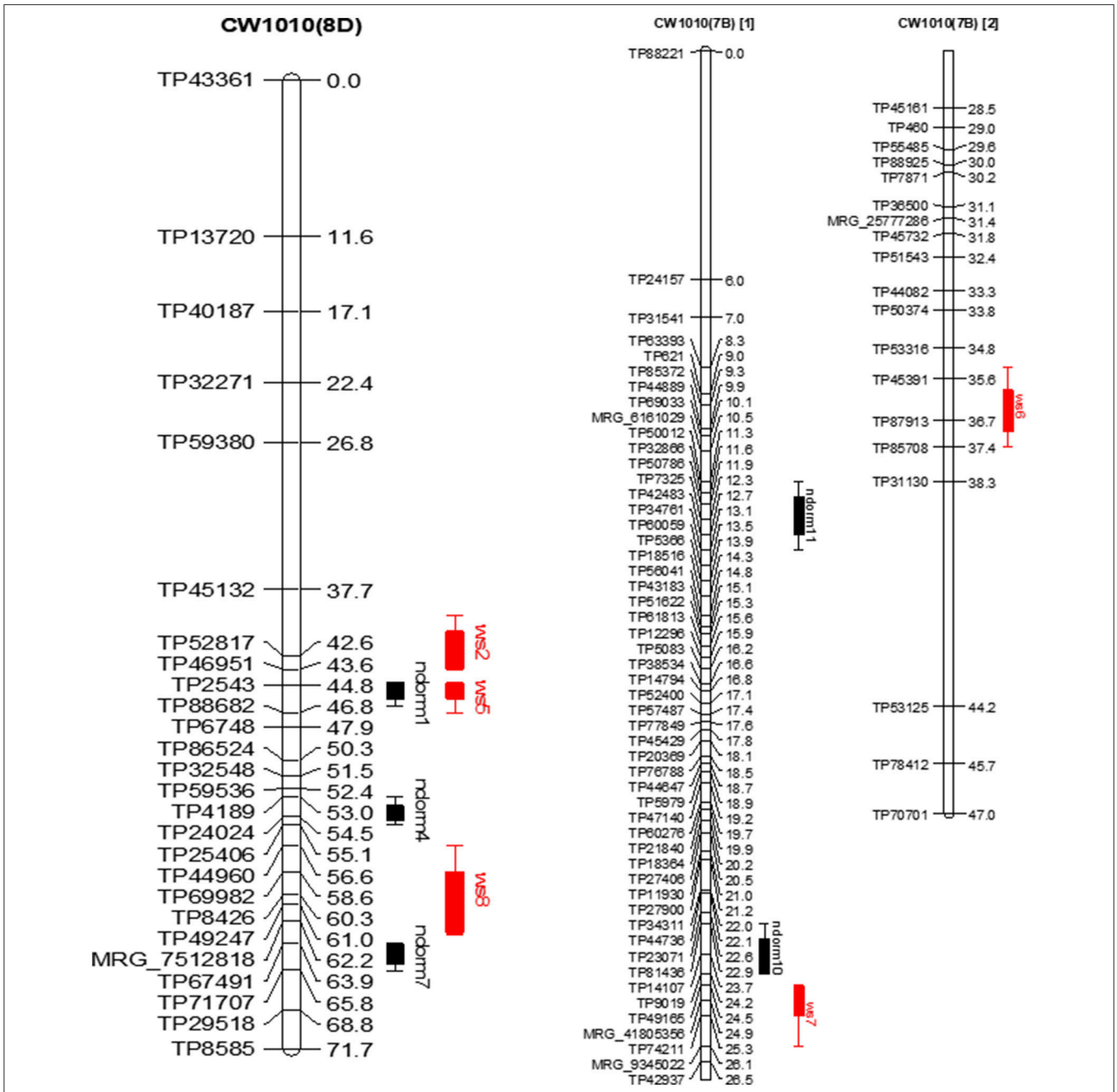

FIGURE 5 | Dormancy (black bar) and WH (red bar) stable QTLs mapped on linkage maps of homolog 8D **(left)** and 7B **(right)** for CW 1010 parent. The QTL bars have two intervals, an inner (1-LOD support) interval and an outer (2-LOD support) interval, where the rectangle represents inner interval and the line represents the outer. Some stable QTLs for dormancy were co-localized with WH in the same genomic regions.

observed between dormancy and WH irrespective of the time of assessment is a clear indication of the weak association between the two traits (**Tables 3**, **4**).

Genetic Linkage Map

Genetic maps are important tools for genetic analysis of quantitative traits through QTL and comparative mapping. They are also useful for genome assembly and marker development for MAS (Moriguchi et al., 2012; Li et al., 2014b). In alfalfa, genomic resources are very scarce and even though a limited number of genetic maps were published, there is no consensus map to date. Alfalfa linkage maps reported so far have variable sizes and marker density depending on the type of mapping population, maker types, and software used. Most studies reported tetraploid alfalfa maps either for 8 linkage groups or for 32 linkage groups for each parent (Julier et al., 2003; Li et al., 2015). An early

reported tetraploid alfalfa linkage map covered only 443 cM for seven linkage groups (Brouwer and Osborn, 1999). Julier et al. (2003) constructed alfalfa genetic maps using SSR and AFLP for eight linkage groups containing four homologs, where the parental maps spanned 2649 and 3045 cM (Julier et al., 2003). The authors argued that their alfalfa genetic maps were close to saturation and exhibited high level of collinearity with other maps of alfalfa and *M. truncatula* (Julier et al., 2003). However, the genetic maps they constructed were relatively less dense (7–9 cM/marker). Musial et al. (2007) reported alfalfa linkage maps using a backcross (BC) population, where the eight linkage groups spanned 794.1 cM with 3.9 cM/marker (Musial et al., 2007). Li et al. (2015) constructed linkage maps of WISFAL-6 and ABI408 with respective lengths of 898 cM and 845 cM for 8 linkage groups (Li et al., 2015). The linkage map constructed in this study is the second high-density genetic map, published so far, after the alfalfa linkage maps described by Li et al. (2014b) for two alfalfa genotypes DM3 and DM5. Moreover, our linkage maps for both 3010 and CW 1010 parents have almost similar average marker density ($\sim$ 1.5 cM/SDA-SNP) to the linkage map of parent DM3 (Li et al., 2014b). Our linkage maps also exhibited high level of synteny with the *M. truncatula* genome as in (Julier et al., 2003; Li et al., 2014b). The total length of the 3010 map (2788.5 cM) was slightly higher than the CW 1010 map (2127.55 cM). Such differences in parental linkage maps were also observed in previous alfalfa genetic linkage maps (Julier et al., 2003). The difference might be due to more SDA markers segregating in 3010 parents (5348) compared to (2340) segregating for CW 1010. The higher number of markers in 3010 may be resulted because of higher number of recombinations that lead to a longer linkage map. In *Brassica oleracea*, the recombination frequency in female meiosis is higher than the male, which obviously generates more markers for the female linkage map and therefore a longer map than the male one (Kearsey et al., 1996). However, in alfalfa there are no reports available regarding sex related differences in meiosis frequency. As we obtained a lower number of raw reads for CW 1010 parents than 3010 in either replications, we believe that these differences could result in low number of markers for CW 1010. Furthermore, since GBS is a reduced representation approach, the number and quality of genotype calls may vary between individuals (Gorjanc et al., 2015).

Mapping and QTL Detection

Constructing linkage maps based on single dose markers (1:1) in outcrossing polyploid species and using the maps for linkage analysis of quantitative traits has been a common practice for decades (Wu et al., 1992). The pseudo-testcross strategy uses the heterozygous markers for one parent and double null in other parents for mapping, which further uses inbred backcross configuration in mapping software (Scott, 2004; Wu et al., 2010). This mapping strategy has been successfully used before in tree plants such as *Eucalyptus grandis* and *Eucalyptus urophylla* (Grattapaglia and Sederoff, 1994), *Pinus elliottii* and *P. caribaea* (Shepherd et al., 2003), and grass such as orchardgrass (*Dactylis glomerata* L.) (Xie et al., 2010). This mapping strategy possesses however some drawbacks (Pastina et al., 2012). Dominant and additive effects on QTL are confounded and the effects of alleles that were substituted only from other parents are obtainable (Weller, 1992; Boopathi, 2013). Since the parents of the pseudo-testcross population are heterozygous, the marker and QTL alleles may be in different state and linkage phases, which makes the strategy less powerful than the classical QTL analysis in inbred populations (Boopathi, 2013) in addition to mapping only a portion of markers (Semagn et al., 2010). Nevertheless, this strategy is still useful in detecting QTL, displaying the direction and magnitude of QTL effect and the position of QTL in species with complex genome such as polyploids.

In this study, we used the pseudo-testcross strategy with GBS SNPs in alfalfa for QTL mapping of dormancy and WH in alfalfa. A total of 45 QTLs associated with FD and 35 QTLs for WH were mapped on two alfalfa cultivars CW 1010 (male) and 3010 (female) genetic maps. Even though previous studies reported QTL mapping of dormancy in alfalfa, these maps were based on parents relatively close in dormancy and constructed with traditional markers (Zúñiga et al., 2004; Robins et al., 2007; Li et al., 2015). Furthermore, with 11 dormancy classes and 6 classes of WH it is very unlikely to capture the majority of loci underlying these traits in a bi-parental population. The latest alfalfa QTL map reported QTLs for winter injury and FD, but with large QTL intervals (>10 cM) (Li et al., 2015) leading the authors to suggest the need for further research to narrow down QTLs positions. The mapping population was also generated from a dormant × a semi-dormant and winter hardy parents (WISFAL-6 x ABI408) which makes it difficult to capture the alleles underlying non-dormancy and cold susceptibility. In this study, the QTLs were detected within 1-LOD support interval with flanking and peak markers for all QTLs identified (**Tables 5–8**). Some of the QTLs detected in this study were located in the same genomic locations as previous studies (Li et al., 2015).

Because of the genotype by environment interactions, the QTLs for FD and WH were categorized into stable QTLs that were consistently expressed in more than one environment and potential QTLs that were detected just in an individual environment for one season or only across environments. Considering QTL × E in the analysis enhances the precision of QTL study since the multi-environment QTL test is more powerful in comparison to single-environment analysis (El-Soda et al., 2019). Therefore, to verify the alfalfa WH and FD QTLs detected in our analysis validation studies need to be conducted in other environments using different alfalfa backgrounds.

Association Between FD and WH

Phenotypic and genetic relationship between alfalfa FD and WH has been a matter of debate for a longtime. Earlier studies reported that alfalfa dormancy and WH are phenotypically correlated ($r = 0.90$) and most likely genetically associated leading breeders to use one trait as surrogate to select for the other (Perry et al., 1987; Brummer et al., 2000). Recent studies re-examining this relationship argued for weaker genetic linkages between the traits and suggested that improving one trait by selecting for the other may not be successful (Brummer et al., 2000; Li et al., 2015). Other findings suggested that the

relationship between WH and FD in alfalfa depends on the type of germplasm tested (Brummer et al., 2000), implying that the two traits could be manipulated independently (Brummer et al., 2000; Weishaar et al., 2005).

In this study, we observed moderate positive phenotypic correlations between WH and FD in the F1 pseudo-testcross population. The magnitude of correlation however varied with the assessment time of FD. Dormancy measured in fall after clipping on 21 September showed weaker relation to WH than dormancy assessed in winter. Assessing regrowth height for FD in areas with warmer late autumn temperatures may not be reliable and should be delayed to early winter. Nevertheless, for reliable ratings of FD and WH, multi-years data is necessary (Teuber et al., 1998).

The QTL analysis performed in this study revealed that more than 75% (22/28) of the dormancy QTL detected for the 3010 parent did not share genomic regions with winter hardiness QTLs (**Tables 5–8**). Similarly, for the CW 1010 parent, more than 70% (12/17) dormancy QTLs detected were localized in different genomic regions than winter hardiness QTLs. These results clearly suggest that the two traits are inherited separately and therefore can be genetically manipulated independently in breeding programs (Brummer et al., 2000; Li et al., 2015). The dormancy QTLs (dorm1, dorm2; ndorm1, ndorm4) sharing common genomic regions with winter hardiness QTLs (wh3, ws2, ws5) in the two parents might have been the result of pseudo-linkage resulting from the simultaneous long-term selection for the two traits. It is important to note that a pseudo-testcross population does not provide enough recombination to break apart closely linked loci. Previous QTL mapping work also found few overlapping QTLs for dormancy and winter injury suggesting genetic relation between the traits (Li et al., 2015). Since the two parents used in our study are more phenotypically divergent (FDR 2 for 3010 vs. FDR 10 for CW 1010) in both dormancy and WH than any of the parents used in previous studies, the genetic linkage between the two traits is most likely tighter (Weishaar et al., 2005). The QTLs detected in this study will be valuable addition to the genomic resources for alfalfa breeding programs and to the understanding of the genetic basis of seasonal dormancy and winter hardiness. The segregating non-dormant genotypes with low winter injury generated in this study constitute a valuable germplasm resource to develop winter-hardy non-dormant cultivars.

AUTHOR CONTRIBUTIONS

LA: contributed to the data collection, analysis, and writing the manuscript. OL: contributed to the winter hardiness data collection. JM: contributed to the phenotypic data collections at the two locations. AM: contributed to the experimental design of the study and writing the manuscript.

FUNDING

Funding support provided by start up funds from the CAES and UGA cultivar research and development program.

ACKNOWLEDGMENTS

The authors acknowledge the research technician Joseph Young, former lab member Franco V. Chirinos and Shiva Makaju for their assistance in the field work and data collection. Our special thanks goes to Mr. Dev Paudel for his assistance with data analysis in bioinformatics related part. This research was partly supported by UGA Cultivar Research and Development Program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00934/full#supplementary-material

REFERENCES

Adhikari, L., and Missaoui, A. M. (2017). Nodulation response to molybdenum supplementation in Alfalfa and it's correlation with root and shoot growh in low Ph soil. *J. Plant Nutr.* 40, 2290–2302. doi: 10.1080/01904167.2016.1264601

Adhikari, L., Mohseni-Moghadam, M., and Missaoui, A. (2018). Allelopathic effects of cereal rye on weed suppression and forage yield in Alfalfa. *Am. J. Plant Sci.* 9, 685–700. doi: 10.4236/ajps.2018.94054

Adhikari, L., Razar, R. M., Paudel, D., Ding, R., and Missaoui, A. M. (2017). Insights into seasonal dormancy of perennial herbaceous forages. *Am. J. Plant Sci.* 8, 2650. doi: 10.4236/ajps.2017.811179

Bolger, A., and Giorgi, F. (2014). *Trimmomatic: A Flexible Read Trimming Tool for illumina NGS Data*.Available online at: http://www.usadellab.org/cms/index.php.

Boopathi, N. M. (2013). *Genetic Mapping and Marker Assisted Selection: Basics, Practice and Benefits*. New Delhi: Springer India.

Boudhrioua, C., Bastien, M., Légaré, G., Pomerleau, S., St-Cyr, J., Boyle, B., et al. (2017). "Genotyping-by-sequencing in potato," in *The Potato Genome,* eds S. Kumar Chakrabarti, C. Xie, and J. Kumar Tiwari (Cham: Springer International Publishing), 283–296.

Brouwer, D. J., and Osborn, T. C. (1999). A molecular marker linkage map of tetraploid alfalfa (*Medicago sativa L.*). *Theor. Appl. Genet.* 99, 1194–1200. doi: 10.1007/s001220051324

Brouwer, D. J., Duke, S. H., and Osborn, T. C. (2000). Mapping genetic factors associated with winter hardiness, fall growth, and freezing injury in autotetraploid alfalfa. *Crop Sci.* 40, 1387–1396. doi: 10.2135/cropsci2000.4051387x

Brummer, E. C., Shah, M. M., and Luth, D. (2000). Reexamining the relationship between fall dormancy and winter hardiness in alfalfa. *Crop Sci.* 40, 971–977. doi: 10.2135/cropsci2000.404971x

Castonguay, Y., Bertrand, A., Michaud, R., and Laberge, S. (2011). Cold-induced biochemical and molecular changes in alfalfa populations selectively improved for freezing tolerance. *Crop Sci.* 51, 2132–2144. doi: 10.2135/cropsci2011.02.0060

Cunningham, S. M., Volenec, J. J., and Teuber, L. R. (1998). Plant survival and root and bud composition of alfalfa populations selected for contrasting fall dormancy. *Crop Sci.* 38, 962–969. doi: 10.2135/cropsci1998.0011183X003800040014x

da Silva, W. L., Ingram, J., Hackett, C. A., Coombs, J. J., Douches, D., Bryan, G. J., et al. (2017). Mapping loci that control tuber and foliar symptoms caused

by PVY in autotetraploid Potato (*Solanum tuberosum L.*). *G3 (Bethesda)* 7, 3587–3595. doi: 10.1534/g3.117.300264

Dhanaraj, A. L., Alkharouf, N. W., Beard, H. S., Chouikha, I. B., Matthews, B. F., Wei, H., et al. (2007). Major differences observed in transcript profiles of blueberry during cold acclimation under field and cold room conditions. *Planta* 225, 735–751. doi: 10.1007/s00425-006-0382-1

Doyle, J., and Doyle, J. (1987). Genomic plant DNA preparation from fresh tissue-CTAB method. *Phytochem. Bull* 19, 11–15.

Duke JA. (1981). *Handbook of Legumes of World Economic Importance*. New York, NY: Plenum Press.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* 6:e19379. doi: 10.1371/journal.pone.0019379

El-Soda, M., Malosetti, M., Zwaan, B. J., Koornneef, M., and Aarts, G. M. (2019). Genotype × environment interaction QTL mapping in plants: lessons from Arabidopsis. *Trends Plant Sci.* 19, 390–398. doi: 10.1016/j.tplants.2014.01.001

Eujayl, I., Sledge, M. K., Wang, L., May, G. D., Chekhovskiy, K., Zwonitzer, J. C., et al. (2004). Medicago truncatula EST-SSRs reveal cross-species genetic markers for *Medicago* spp. *Theor. Appl. Genet.* 108, 414–422. doi: 10.1007/s00122-003-1450-6

Gar, O., Sargent, D. J., C.-,Tsai, J., Pleban, T., Shalev, G., et al. Zamir, D. (2011). An autotetraploid linkage map of Rose (*Rosa hybrida*) validated using the Strawberry (*Fragaria vesca*) genome sequence. *PLoS ONE* 6:e20463. doi: 10.1371/journal.pone.0020463

Gorjanc, G., Cleveland, M. A., Houston, R. D., and Hickey, J. M. (2015). Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genet. Select. Evolut.* 47, 12. doi: 10.1186/s12711-015-0102-z

Grattapaglia, D., and Sederoff, R. (1994). Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* 137, 1121–1137.

Gusta, L. V., and Wisniewski, M. (2013). Understanding plant cold hardiness: an opinion. *Physiol. Plant.* 147, 4–14. doi: 10.1111/j.1399-3054.2012.01611.x

Hackett, C. A., Boskamp, B., Vogogias, A., Preedy, K. F., and Milne, I. (2017). TetraploidSNPMap: software for linkage analysis and QTL mapping in autotetraploid populations using SNP dosage data. *J. Heredity* 108, 438–442. doi: 10.1093/jhered/esx022

Hackett, C. A., Milne, I., Bradshaw, J. E., and Luo, Z. (2007). TetraploidMap for Windows: linkage map construction and QTL mapping in autotetraploid species. *J. Hered.* 98, 727–729. doi: 10.1093/jhered/esm086

Haggard, J. E., Johnson, E. B., and St. Clair, D. A. (2015). Multiple QTL for horticultural traits and quantitative resistance to *Phytophthora infestans* linked on *Solanum habrochaites* chromosome 11. *G3 (Bethesda)* 5, 219–233. doi: 10.1534/g3.114.014654

Ibáñez, C., Kozarewa, I., Johansson, M., Ögren, E., Rohde, A., and Eriksson, M. E. (2010). Circadian clock components regulate entry and affect exit of seasonal dormancy as well as winter hardiness in populus trees. *Plant Physiol.* 153, 1823–1833. doi: 10.1104/pp.110.158220

Julier, B., Flajoulot, S., Barre, P., Cardinet, G., Santoni, S., Huguet, T., et al. (2003). Construction of two genetic linkage maps in cultivated tetraploid alfalfa (*Medicago sativa*) using microsatellite and AFLP markers. *BMC Plant Biol.* 3:9. doi: 10.1186/1471-2229-3-9

Kearsey, M. J., Ramsay, L. D., Jennings, D. E., Lydiate, D. J., Bohuon, E. J., Marshall, D. F., et al. (1996). Higher recombination frequencies in female compared to male meisoses in Brassica oleracea. *Theor. Appl. Genet.* 92, 363–367. doi: 10.1007/BF00223680

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, X., Alarcón-Zú-iga, B., Kang, J., Nadeem Tahir, M. H., Jiang, Q., Wei, Y., et al. (2015). Mapping fall dormancy and winter injury in tetraploid alfalfa. *Crop Sci.* 55, 1995–2011. doi: 10.2135/cropsci2014.12.0834

Li, X., and Brummer, E. C. (2012). Applied genetics and genomics in alfalfa breeding. *Agronomy* 2:40. doi: 10.3390/agronomy2010040

Li, X., Han, Y., Wei, Y., Acharya, A., Farmer, A. D., Ho, J., et al. (2014a). Development of an Alfalfa SNP array and its use to evaluate patterns of population structure and linkage disequilibrium. *PLoS ONE* 9:e84329. doi: 10.1371/journal.pone.0084329

Li, X., Wei, Y., Acharya, A., Jiang, Q., Kang, J., and Brummer, E. C. (2014b). A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa L.*) developed using genotyping-by-sequencing is highly syntenous with the *Medicago truncatula* genome. *G3 (Bethesda)* 4, 1971–1979. doi: 10.1534/g3.114.012245

Li, X., Wei, Y., Moore, K. J., Michaud, R., Viands, D. R., Hansen, J. L., et al. (2011). Association mapping of biomass yield and stem composition in a tetraploid alfalfa breeding population. *Plant Genome* 4, 24–35. doi: 10.3835/plantgenome2010.09.0022

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., et al. (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet.* 9:e1003215. doi: 10.1371/journal.pgen.1003215

McCaslin, M., Woodward, T., and Undersander, D. (2003). *Winter Survival, Standard Tests to Characterize Alfalfa Cultivars*. Beltsville, MD: North America Alfalfa Improvement Conference. Available online at: http://www.naaic.org/stdtests/wintersurvivalnew.htm (Accessed May 20, 2013).

McKenzie, J. S., Paquin, R., and Duke, S. H. (1988). "Cold and Heat Tolerance," in Alfalfa and Alfalfa Improvement, American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, eds A. A. Hanson, D. K. Barnes, and R. R. Hill (Madison, WI), 259–302.

Melo, A. T. O., Bartaula, R., and Hale, I. (2016). GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinformatics* 17:29. doi: 10.1186/s12859-016-0879-y

Moriguchi, Y., Ujino-Ihara, T., Uchiyama, K., Futamura, N., Saito, M., Ueno, S., et al. (2012). The construction of a high-density linkage map for identifying SNP markers that are tightly linked to a nuclear-recessive major gene for male sterility in *Cryptomeria japonica* Don, D. *BMC Genomics* 13:95. doi: 10.1186/1471-2164-13-95

Musial, J. M., Mackie, J. M., Armour, D. J., Phan, H. T., Ellwood, S. E., Aitken, K. S., et al. (2007). Identification of QTL for resistance and susceptibility to Stagonospora meliloti in autotetraploid lucerne. *Theor. Appl. Genet.* 114, 1427–1435. doi: 10.1007/s00122-007-0528-y

Olsen, J. E. (2010). Light and temperature sensing and signaling in induction of bud dormancy in woody plants. *Plant Mol. Biol.* 73, 37–47. doi: 10.1007/s11103-010-9620-9

Pastina, M., Malosetti, M., Gazaffi, R., Mollinari, M., Margarido, G., Oliveira, K., et al. (2012). A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. *Theor. Appl. Genet.* 124, 835–849. doi: 10.1007/s00122-011-1748-8

Perry, M. C., McIntosh, M. S., Wiebold, W. J., and Welterlen, M. (1987). Genetic analysis of cold hardiness and dormancy in alfalfa. *Genome* 29, 144–149. doi: 10.1139/g87-024

Robins, J., Luth, D., Campbell, T., Bauchan, G., He, C., Viands, D., et al. (2007). Genetic mapping of biomass production in tetraploid alfalfa. *Crop Sci.* 47, 1–10. doi: 10.2135/cropsci2005.11.0401

SAS Institute Inc. (2004). Cary, NC: SAS Institute Inc.

Santhosh, K. (1989). http://bio-bwa.sourceforge.net/.

Scott, L. J. (2004). *Implications of evolutionary history and population structure for the analysis of quantitative trait loci in the ancient conifer Araucaria cunninghamii*. PhD thesis, Southern Cross University, Lismore, NSW.

Semagn, K., Bjørnstad, Å., and Xu, Y. (2010). The genetic dissection of quantitative traits in crops. *Electr. J. Biotechnol.* 13, 16–17. doi: 10.2225/vol13-issue5-fulltext-14

Shepherd, M., Cross, M., Dieters, M. J., and Henry, R. (2003). Genetic maps for *Pinus elliottii var.* elliottii and P. caribaea var. hondurensis using AFLP and microsatellite markers. *Theor. Appl. Genet.* 106, 1409–1419. doi: 10.1007/s00122-002-1185-9

Shu, Y., Li, W., Zhao, J., Zhang, S., Xu, H., Liu, Y., et al. (2017). Transcriptome sequencing analysis of alfalfa reveals CBF genes potentially playing important roles in response to freezing stress. *Genet. Mol. Biol.* 40, 824–833. doi: 10.1590/1678-4685-gmb-2017-0053

Stout, D. G., and Hall, J. W. (1989). Fall growth and winter survival of alfalfa in interior British Columbia. *Can. J. Plant Sci.* 69, 491–499. doi: 10.4141/cjps89-060

Tanino, K. K., Kalcsits, L., Silim, S., Kendall, E., and Gray, G. R. (2010). Temperature-driven plasticity in growth cessation and dormancy development in deciduous woody plants: a working hypothesis suggesting how molecular and cellular function is affected by temperature during dormancy induction. *Plant Mol. Biol.* 73, 49–65. doi: 10.1007/s11103-010-9610-y

Teuber, L., Taggard, K., Gibbs, L., McCaslin, M., Peterson, M., and Barnes, D. (1998). "Fall dormancy, Standard tests to characterize alfalfa cultivars," in 36th *international Conference the North American Alfalfa Improvement* (Bozeman, MT), 2–6.

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

Wan, S. M., Liu, H., Zhao, B. W., Nie, C. H., Wang, W. M., et al. (2017). Construction of a high-density linkage map and fine mapping of QTLs for growth and gonad related traits in blunt snout bream. *Sci. Rep.* 7:46509. doi: 10.1038/srep46509

Weishaar, M. A., Brummer, E. C., Volenec, J. J., Moore, K. J., and Cunningham, S. (2005). Improving Winter Hardiness in Nondormant Alfalfa Germplasm This journal paper of the Iowa Agric. Home Econ. Exp. Stn., Ames, IA, Project No. 6631 and 6525, was supported by the Hatch Act and State of Iowa funds. *Crop Sci.* 45, 60–65. doi: 10.2135/cropsci2005.0060

Weller, J. I. (1992). "Statistical methodologies for mapping and analysis of quantitative trait loci," in *Plant Genomes: Methods for Genetic and Physical Mapping*. eds J. S. Beckmann and T. C. Osborn (Dordrecht: Springer), 181–207.

Wu, K., Burnquist, W., Sorrells, M., Tew, T., Moore, P., and Tanksley, S. (1992). The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theor. Appl. Genet.* 83, 294–300. doi: 10.1007/BF00224274

Wu, S., Yang, J., Huang, Y., Li, Y., Yin, T., Wullschleger, S. D., et al. (2010). An improved approach for mapping quantitative trait loci in a pseudo-testcross: revisiting a poplar mapping study. *Bioinform. Biol. Insights* 4, 1–8. doi: 10.4137/BBI.S4153

Xie, W., Zhang, X., Cai, H., Huang, L., Peng, Y., and Ma, X. (2010). Genetic maps of SSR and SRAP markers in diploid orchardgrass (*Dactylis glomerata* L.) using the pseudo-testcross strategy. *Genome* 54, 212–221. doi: 10.1139/G10-111

Xiong, Y., Fei, S.-Z., Arora, R., Brummer, E. C., Barker, R. E., Jung, G., et al. (2007). Identification of quantitative trait loci controlling winter hardiness in an annual × perennial ryegrass interspecific hybrid population. *Mol. Breed.* 19, 125–136. doi: 10.1007/s11032-006-9050-1

Zhang, S., Shi, Y., Cheng, N., Du, H., Fan, W., and Wang, C. (2015). *De novo* characterization of fall dormant and nondormant Alfalfa (*Medicago sativa L.*) leaf transcriptome and identification of candidate genes related to fall dormancy. *PLoS ONE* 10:e0122170. doi: 10.1371/journal.pone.0122170

Zúñiga, B. A., Scott, P., Moore, K., Luth, D., and Brummer, E. (2004). Quantitative Trait Locus Mapping of Winter Hardiness Metabolites in Autotetraploid Alfalfa (*M. sativa). Molecular Breeding of Forage and Turf. Developments in Plant Breeding*, Vol. 11, eds A. Hopkins, Z.Y. Wang, R. Mian, M. Sledge, and R. E.Barker (Dordrecht: Springer), 97–104.

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: 26 June 2018
doi: 10.3389/fpls.2018.00845

Genetic Architecture of Nitrogen-Deficiency Tolerance in Wheat Seedlings Based on a Nested Association Mapping (NAM) Population

Deqiang Ren[1†], Xiaojian Fang[1†], Peng Jiang[1], Guangxu Zhang[2], Junmei Hu[1], Xiaoqian Wang[1], Qing Meng[1], Weian Cui[1], Shengjie Lan[1], Xin Ma[1], Hongwei Wang[1] and Lingrang Kong[1]**

[1] State Key Laboratory of Crop Biology, Shandong Key Laboratory of Crop Biology, College of Agronomy, Shandong Agricultural University, Tai'an, China, [2] Lianyungang Academy of Agricultural Sciences, Lianyungang, China

OPEN ACCESS

Edited by:
Yiwei Jiang,
Purdue University, United States

Reviewed by:
Shuanghe Cao,
Institute of Crop Sciences (CAAS), China
Xingwang Yu,
North Carolina State University, United States

***Correspondence:**
Hongwei Wang
wanghongwei@sdau.edu.cn
Lingrang Kong
lkong@sdau.edu.cn

† These authors have contributed equally to this work.

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *05 March 2018*
Accepted: *31 May 2018*
Published: *26 June 2018*

Citation:
Ren D, Fang X, Jiang P, Zhang G, Hu J, Wang X, Meng Q, Cui W, Lan S, Ma X, Wang H and Kong L (2018) Genetic Architecture of Nitrogen-Deficiency Tolerance in Wheat Seedlings Based on a Nested Association Mapping (NAM) Population. Front. Plant Sci. 9:845. doi: 10.3389/fpls.2018.00845

Genetic divergence for nitrogen utilization in germplasms is important in wheat breeding programs, especially for low nitrogen input management. In this study, a nested association mapping (NAM) population, derived from "Yanzhan 1" (a Chinese domesticated cultivar) crossed with "Hussar" (a British domesticated cultivar) and another three semi-wild wheat varieties, namely, "Cayazheda 29" (*Triticum aestivum* ssp. *tibetanum* Shao), "Yunnan" (*T. aestivum* ssp. *yunnanense* King), and "Yutian" (*T. aestivum petropavloski* Udats et Migusch), was used to detect quantitative trait loci (QTLs) for nitrogen utilization at the seedling stage. An integrated genetic map was constructed using 2,059 single nucleotide polymorphism (SNP) markers from a 90 K SNP chip, with a total coverage of 2,355.75 cM and an average marker spacing of 1.13 cM. A total of 67 QTLs for RDW (root dry weight), SDW (shoot dry weight), TDW (total dry weight), and RSDW (root to shoot ratio) were identified under normal nitrogen conditions (N^+) and nitrogen deficient conditions (N^-). Twenty-three of these QTLs were only detected under N^- conditions. Moreover, 23 favorable QTLs were identified in the domesticated cultivar Yanzhan 1, 15 of which were detected under N^+ conditions, while only four were detected under N^- conditions. In contrast, the semi-wild cultivars contributed more favorable N^--specific QTLs (eight from Cayazheda 29; nine from Yunnan), which could be further explored for breeding cultivars adapted to nitrogen-deficient conditions. In particular, *QRSDW-5A.1* from YN should be further evaluated using high-resolution mapping.

Keywords: nitrogen utilization, nested association mapping (NAM), semi-wild wheat, quantitative trait loci (QTLs), wheat breeding

INTRODUCTION

Nitrogen (N), an essential plant nutrient, is vital for various aspects of crop growth and development, including seed germination, root architecture regulation, shoot development, flowering, and grain production (Lea and Morot-Gaudry, 2001; Alboresi et al., 2005; Kiba and Krapp, 2016; Yuan et al., 2016). Wheat production mainly depends on fertilizer input, particularly

N fertilizer (Peng et al., 2009). From 2008 to 2015, the total global N consumption increased annually by 3.5%. In 2015, the total global N consumption was 223 million tons, and the average N application to wheat was 71–370 kg/hm^2 (FAOSTAT, 2015), which is far higher than the safety threshold of 260 kg/hm^2 (Liu, 2017) in many areas. This excessive N input not only raises the cost of production, but it also causes various soil and environmental issues (Peng et al., 2009). Thus, it is necessary that N use in agriculture is reduced without decreasing grain yields. Wheat varieties are typically developed for maximum production with high N fertilizer input, which results in a decrease in N use efficiency (Dong et al., 2014). Therefore, in order to increase production without further damage to the environment, high-yield crop varieties tolerant of N deficient conditions, or those that can efficiently utilize limited N, are desirable.

A thorough understanding of the mechanisms of N-deficiency tolerance and N-use efficiency in crop plants is required for the development of wheat varieties that are less dependent on N fertilizer. Plants have developed complex adaptive response pathways to cope with N fluctuations (Sandrine et al., 2014). N, as a metabolite, has been well studied. For instance, the nitrate transporter NRT1.1 not only transports NO_3^-, but also enhances the movement of basipetal auxin out of the roots, leading to the repression of lateral root development (Krouk et al., 2006; Wang et al., 2017). Many other proteins, such as the transcription factors ANR1 (nitrite regulator 1, Gan et al., 2005), SPL9 (squamosa promoter binding protein-like 9, Krouk et al., 2010), and NLP7 (NIN-like protein 7, Marchive et al., 2013), as well as the RING-type ubiquitin ligase NLA (Peng et al., 2007), respond to N metabolites. Transcriptome studies have shown that a wide range of physiological and developmental processes are controlled by N signals (Sandrine et al., 2014). Most N-responsive genes are also regulated by hormone and carbon signaling, indicating that N signaling mechanisms are highly integrated with other regulatory pathways (Sandrine et al., 2014). Despite this thorough understanding of N metabolites, the genetic mechanisms by which wheat tolerates or efficiently uses limited N are largely unclear.

Quantitative trait loci (QTL) mapping is a powerful tool for dissecting and understanding the genetic regulation of complex quantitative traits (Cui et al., 2014). Previous QTL studies have focused on morphological traits and crop yields in plants with low N tolerance or with efficient N uptake in hydroponic culture experiments (An et al., 2006; Laperche et al., 2007) and in field experiments (Quarrie et al., 2005; An et al., 2006; Laperche et al., 2007, 2008; Fontaine et al., 2009; Cui et al., 2014), leading to the identification of important QTLs on chromosomes 2A, 2B, 4A, 5A, 7A, and 7B. For instance, Quarrie et al. (2005) reported that major QTLs for grain yield components (ears per plant, grains per ear, and 1000s grain weight) under nitrogen deficiency condition were mapped on chromosomes 4AS, 7AL, 7BL, and around centromeres of chromosomes 4B and 6A using a spring wheat doubled haploid (DH) population derived from the cross Chinese Spring × SQ1. Laperche et al. (2007) detected 233 QTL for traits measured in each combination of environment and clustered into 82 genome regions, the dwarfing gene (*Rht-B1*), the photoperiod sensitivity gene (*Ppd-D1*) and the awns inhibitor gene (*B1*) coincided with regions that contained the highest numbers of QTL. Cui et al. (2016) reported that the *Rht-B1* affected not only plant height but also grain quality and its adaptability to N-deficient environments. Several other co-localizations between QTLs related to yield, physiological traits and enzyme activities involved in the control of N assimilation and recycling were detected for nitrate reductase (NR) and glutamate dehydrogenase (GDH) in maize (Hirel et al., 2001; Gallais and Hirel, 2004), glutamine synthetase (GS) in wheat (Habash et al., 2007; Fontaine et al., 2009). It is important to understand the identified specific QTLs associated with the adaptation of the plant to different N supply conditions. QTLs controlling high levels of N uptake and utilization can be detected specifically under high N conditions, and QTLs specifically detected under N limited conditions are involved in N-deficiency tolerance and adaption processes (Gaju et al., 2011). Direct selection for QTLs specifically detected under low N supply would be effective for the genetic improvement of N-deficiency tolerance traits (Dong et al., 2014).

Most studies were conducted on single, bi-parental population, thus the genetic polymorphisms are limited between two parents. Joint-multiple family analyses, such as "NAM," potentially detect more QTLs, more accurately estimate QTL effects, better resolve QTL positions, and directly assess the distribution of functional allelic variation across multiple families, as compared to QTL analysis by bi-parental population (Yu et al., 2008; McMullen et al., 2009; Li et al., 2013; Ogut et al., 2015; Vatter et al., 2017). NAM population is a joint-multiple family comprising multiple bi-parental mapping families all sharing one common reference parent (Yu et al., 2008). For example, Sophie et al. (2015) developed a sorghum NAM population comprised of 2,400 recombinant inbred lines (RILs) from 10 families with the sorghum hybrid RTx430 as the common parent. The recombination rate of the NAM population was 4 cM/Mb, estimated based on 96,000 SNPs generated with a genotyping-by-sequencing approach, and 57,500 recombination events were observed (Sophie et al., 2015). Using this NAM population, Sophie et al. (2015) detected 41 QTLs and reduced the QTL mapping region to between 63 Kb and 1.9 Mb.

Here, we studied the genetic architecture of N-deficiency tolerance in wheat seedlings using a NAM dataset comprised of four related RIL populations. We constructed an integrative genetic map using high-density SNP markers genotyped with a 90K SNP chip. We aimed to detect QTLs involved in N-deficiency tolerance and to identify favorable alleles.

MATERIALS AND METHODS

Plant Materials

The NAM population that we constucted was comprised of four RIL populations derived from crosses between a single female parent "Yanzhan 1" (YZ) and four different male parents. YZ is a good-quality, high-yield, disease-resistant variety of winter wheat cultivated in Henan Province of the Huanghuai region, China (in 2003). The male parents were "Hussar" (HR, a British dwarf cultivar), and three semi-wild wheat varieties from China: "Chayazheda" (CY, *Triticum aestivum* ssp. *tibetanum* Shao) from

Tibet, "Yunnanxiaomai" (YN, *T. aestivum* ssp. *yunnanense* King) from Yunnan, and Yutiandaomai (YT, *T. aestivum petropavloski* Udats et Migusch) from Xinjiang. We crossed YZ with HR, CY, YN, and YT to develop separate RILs using a single seed descent approach. The final population sizes for each cross were 97, 82, 98, and 93, respectively.

Experimental Design

All of the plants were grown in hydroponic culture (following Sun et al., 2013) in a greenhouse at Shandong Agricultural University, Shandong, China. We used Hoagland's solution (Hoagland and Arnon, 1950) to optimize plant growth (**Supplementary Table S1**). To inhibit any potential nitrification of the nutrient solution, we added 2 mg/L dicyandiamide (a nitrification inhibitor). We tested two levels of N: normal (N^+; 5.0 mmol/L N) and low (N^-; 0.5 mmol/L N). We used a randomized complete block design with three replicates for each treatment. Wheat seeds were sterilized for 10 min in 10% sodium hypochlorite, washed with distilled water, and then germinated in a germination tray. After 7 days, one healthy seedling from each line and each treatment was transferred to a 200-cell bottomless tray. The tray was placed in an opaque plastic tank containing 20 L nutrient solution. The tank was opaque in order to encourage healthy root growth and to restrict the growth of algae. The nutrient solution was renewed every 3 days, and the pH was adjusted to 6.0 every day. We repeated this entire procedure six times in 2017: February 26–March 31 (E1); March 2–April 7 (E2); March 9–April 14 (E3); March 16–April 21 (E4); March 23–April 28 (E5); and March 31–May 5 (E6). Our experiment thus comprised 12 environmental combinations: $E1N^+$, $E1N^-$, $E2N^+$, $E2N^-$, $E3N^+$, $E3N^-$, $E4N^+$, $E4N^-$, $E5N^+$, $E5N^-$, $E6N^+$, and $E6N^-$.

Trait Measurements

All of the plants were harvested after 30 days in the nutrient solution. The roots were cleaned with distilled water, and excess water was blotted with absorbent paper. Plants were then dried for 24 h at 56°C in a drying oven before measuring dry root weight (RDW, in mg) and dry shoot weight (SDW, in mg). The total dry weight (TDW, in mg) was calculated as RDW + SDW and the ratio of dry root weight to dry shoot weight (RSDW) was calculated as RDW/SDW. To estimate the plant response to N deficiency, we calculated a "global" interaction variable (G × N) as $(N_y{}^- - N_y{}^+)/N_y{}^+$, where $N_y{}^-$ and $N_y{}^+$ represent the trait values in the N^- and N^+ treatments, respectively.

Genotyping and Genetic Map Construction

Genomic DNA was extracted from the seedling leaves of all five parents and all of the RILs (Doyle and Doyle, 1987). DNA samples were genotyped with an Illumina 90K assay (Wang et al., 2014). All of the SNP markers for each line were converted based on the alleles of the parents: 'A' for the common parent YZ, 'B' for the other parental lines, 'H' for the heterozygous genotype, and '-' for missing information. Individual genetic maps for each RIL population were constructed using Kosambi mapping (Kosambi, 1944) and individual maps were combined with Joinmap v4.0 (Van Ooijen and Jansen, 2013)[1]. The integrative map was drawn using MapChart v2.2 (Voorrips, 2002)[2].

Data Analysis and QTL Mapping

We tested the significance of the phenotypic differences between each pair of parents using Student's *t*-tests. To estimate the variance across genotypes (G), environment (E), genotype/environment interactions (GEI), and replicates, we used analysis of variance (ANOVA) with generalized linear models (GLMs) in SPSS v20 (Bryman and Cramer, 2012)[3]. Heritability (h^2) was computed using the estimated variance components $V_G/(V_G + V_{GEI}/s + V_e/sr)$, where V_G, V_{GEI}, and V_e are the variances of G, GEI, and the residuals, respectively; s is the number of environments; and r is the number of replicates. The best linear unbiased estimates (BLUE) for each line with respect to each trait across all traits were used to analyze pairwise correlations. We mapped QTLs with the ICIM-ADD method (Li et al., 2007) using stepwise regression, and considered all of the marker information simultaneously in Ici-Mapping v4.1 (Li et al., 2007)[4]. We used a walking speed of 1.0 cM for all of the QTL calculations, and a stepwise regression probability (*P*-value inclusion threshold) of 0.001. We considered a QTL to be present if the limit of detection (LOD) was >2.5 in the NAM population, and >2.0 in at least one RIL population.

RESULTS

Phenotypic Variation

The traits of the five parents differed substantially under both N^+ and N^- conditions across all of the treatments and environments, and exhibited distinctly different responses to N deficiency (**Figures 1A–D** and **Supplementary Table S2**). For instance, under N^- conditions, both the RDW and SDW of the common parent YZ were reduced, but RDW increased in YT and SDW increased in YN, suggesting that these parental species possessed different levels of N-deficiency tolerance (**Figures 1A–D** and **Supplementary Table S2**). Strong transgressive segregation was observed in all of the RIL populations, indicating that favorable alleles were distributed among the parents. Considerable continuous variation was observed in all of the measured traits across all of the populations (**Figures 1E–H** and **Supplementary Tables S2**, **S3**). For each parent and RIL population, the differences in SDW between the N^+ and N^- conditions were significantly greater than the differences in RDW (one-way ANOVA, $P < 0.001$, **Figure 1**; **Supplementary Table S3**). The estimated h^2 of all of the traits ranged from 18.5% for RSDW to 74.0% for TDW. The h^2 of TDW, RDW, and SDW was high (mostly > 40%) under both N^+ and N^- conditions. The h^2 of RSDW was lower, however, ranging from 18.5 to 47.8%. For each trait, h^2 varied between populations. Our results suggested that all of the measured traits were affected not only by

[1]http://www.kyazma.nl/

[2]http://www.biometris.nl/uk/Software/MapChart/

[3]http://en.wikipedia.org/wiki/SPSS

[4]http://www.isbreeding.net/

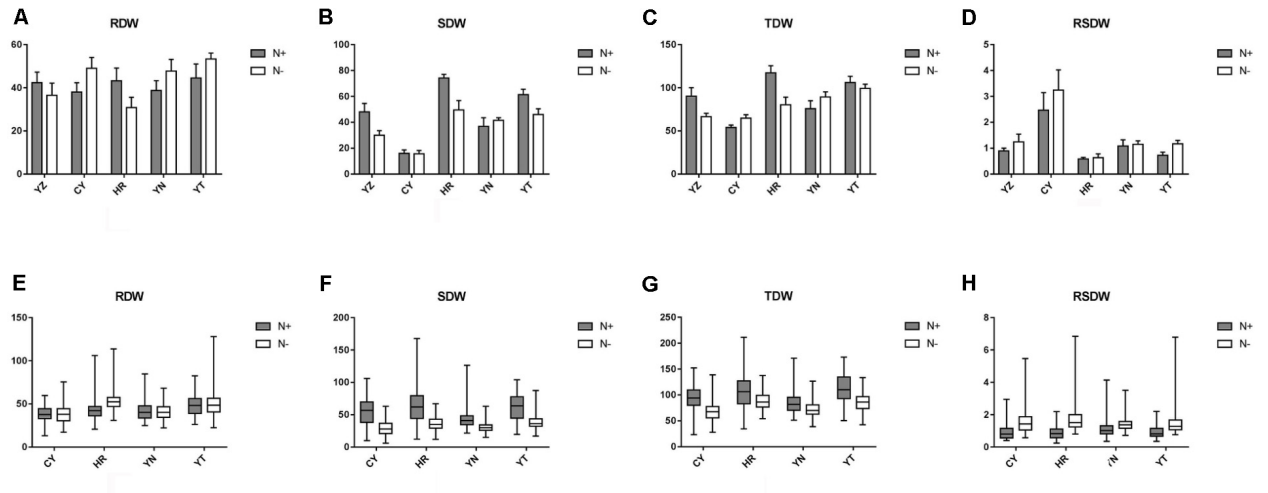

FIGURE 1 | Performance of the five parental varieties and the phenotypic variations of each RIL population with respect to the measured seedling growth traits. **(A–D)** The performance of the five parental varieties is shown using a histogram. **(E–H)** The phenotypic variation of the RIL populations is shown using a boxplot. RDW indicates root dry weight; SDW indicates shoot dry weight; TDW indicates total dry weight; RSDW indicates root to shoot ratio. Performance under N^+ (normal N supply) and N^- (low N supply) conditions are shown in a gray and white pattern. The *x*-axis represents five parental lines: YZ (Yanzhan 1), CY (Cayazheda), HR (Hussar), YN (Yunnanxiaomai), and YT (Yutiandaomai).

genotype, environment, and GEI, but also by genetic background (**Supplementary Table S3**). The phenotypic pairwise correlations between the measured traits were similar under both N^+ and N^-. RDW and SDW were positively correlated with each other and with TDW. RSDW was positively correlated with RDW (as expected), but negatively correlated with SDW (**Table 1**).

The Novel Genetic Map

We selected several polymorphic markers distributed across all 21 chromosomes for linkage analysis: 548 for CY, 1,127 for YN, 1,514 for YT, and 1,595 for HR. We mapped 2,059 loci, including 34 linkage groups, to our integrated genetic map (**Figure 2**, **Table 2**, and **Supplementary Table S4**), with a total coverage of 2,355.75 cM and an average marker spacing of 1.13 cM. Our integrated map included three genomes: the A genome was 887.67 cM (38.0%), and contained 946 loci (45.89%); the B genome was 955.34 cM (40.90%), and contained 979 loci (47.55%); and the D genome was 492.74 cM (21.1%), and contained 135 loci (6.06%). The chromosome sizes ranged from 0.61 cM (chromosome 4D) to 186.86 cM (chromosome 1B). Chromosome 2B had the most loci (202), while chromosome 3D had the least (4). We obtained good coverage for the A and B genomes, but few polymorphic loci were identified for the D genome. Our integrated linkage map had greater genome coverage, more markers, and lower average maker distance than the individual maps (**Table 2**).

QTL Mapping for Seedling Growth Traits

We detected 67 QTLs affecting seedling growth traits, including 31 QTLs identified only under N^+ treatment, 22 only under N^- treatment, and 14 detected under both (**Table 3**, **Supplementary Tables S5**, **S6**, and **Supplementary Figure S1**). The 67 QTLs were distributed across all 21 chromosomes, except 2D, 3D, 4B, 4D, 5D, and 7D. The phenotypic variance (PVE) explained by these QTLs ranged from 2.3% (SDW in E5N^+) to 38.0% (SDW in E3N^+). The 44 QTLs with PVEs greater than 10% (identified as "primary QTLs") were mainly concentrated on chromosomes 1B, 2A, 2B, and 3A. Statistics of the favorable QTLs donated by parents are shown in **Figure 3** and **Supplementary Table S4**. Twenty-three favorable QTLs were donated by the domesticated cultivar of YZ, in which 15 were detected only under N^+ conditions, and four were detected only under N^- conditions. The semi-wild cultivars CY and YN contributed eight and nine favorable QTLs detected only under N^- conditions.

In multiple environments, we repeatedly detected 18 QTLs for TDW on chromosomes 1A, 2A, 2B, 3A, 3B, 4A, 5A, 5B, 6D, 7A, and 7B. These 18 QTLs included eight identified only under the N^+ treatment, seven only under the N^- treatment, and three under both treatments. Most of the favorable alleles (those that increased the value of a given trait) were donated by parent CY. Four QTLs (*QTDW-2A.3*, *QTDW-3A.1 QTDW-7A.1*, and *QTDW-7B.1*) were regarded as primary QTLs, as they explained 10.5–31.7% of the phenotypic variation.

We identified 26 QTLs for RDW on chromosomes 1A, 1B, 1D, 2A, 2B, 3A, 3B, 4A, 6A, 6B, and 7A. Of these 26, 16 were only detected under N^+ treatment, six were only detected under N^- treatment, and four were detected under both N^+ and N^- treatments. Most of these QTLs explained more than 10% of the phenotypic variation. Two QTLs detected in the HR population, *QRDW-2A.2* and *QRDW-7A.2*, accounted for 18.0–34.6% and 23.2–31.9% of the phenotype variance, respectively. Most of the favorable alleles were donated by parent YZ.

The QTLs for SDW were detected on chromosomes 1B, 2A, 2B, 3A, 4A, 6B, and 7B. These QTLs included six under N^+ treatment only, four under N^- treatment only, and four under both treatments. The favorable alleles were mainly donated

TABLE 1 | Pairwise phenotypic correlations among the measured seedling traits.

	RDW-N^+	RSDW-N^+	SDW-N^+	TDW-N^+	RDW-N^-	RSDW-N^-	SDW-N^-	TDW-N^-
RDW-N^+		0.184**	0.296**	0.651**	0.282**	0.038	0.226**	0.309**
RSDW-N^+			−0.686**	−0.470**	−0.083	0.250**	−0.329**	−0.243**
SDW-N^+				0.918**	0.318**	−0.221**	0.563**	0.528**
TDW-N^+					0.366**	−0.156**	0.533**	0.542**
RDW-N^-						0.440**	0.376**	0.844**
RSDW-N^-							−0.518**	−0.028
SDW-N^-								0.810**

***Indicates significance at $P \leq 0.01$.*

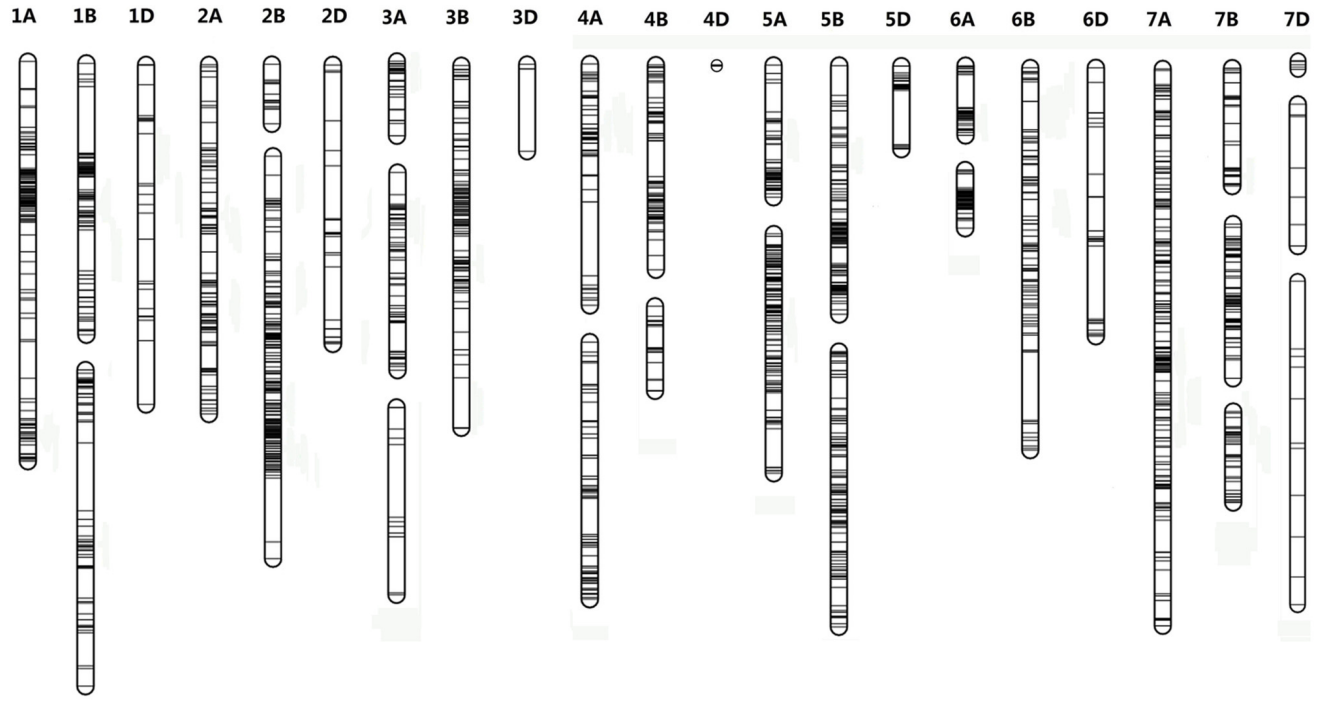

FIGURE 2 | The integrated genetic linkage map of the NAM population. The integrated genetic map of wheat was developed using a NAM population derived from the cross of Yanzhan 1 with four other parents. The positions of the marker loci are indicated using black lines on the chromosome.

by parent YZ. We identified four primary QTLs: *QSDW-2A.1*, *QSDW-2A.3*, *QSDW-2B.1,* and *QSDW-3A.1*.

We identified nine QTLs for RSDW on chromosomes 1A, 1B, 3A, 5A, and 6B. These included two QTLs under N^+ treatment only, five under N^- treatment only, and two under both treatments. Four primary QTLs (*QRSDW-1A.1*, *QRSDW-1B.1*, *QRSDW-6B.1*, *QRSDW-6B.2*) were identified with PVEs ranging from 10.3 to 31.0%.

DISCUSSION

NAM Population and the Novel Integrated Genetic Map

The five parents of the NAM population in this study constitute local adaptable varieties of different origins, and exhibit high phenotypic and genetic diversity. The common parent of YZ was domesticated with the features of a short lifecycle and high yield (Yao et al., 2010). HR (Squadron/Rendezvous) is a British dwarf cultivar developed by Cambridge Plant Breeders (Cambridge, United Kingdom) and Syngenta (formerly Imperial Chemical Industries) that is resistant to many wheat diseases, but was domesticated with a longer lifecycle (Wilde et al., 2008). The other three semi-wild parents are wheat germplasm resources unique to Western China, and possess many morphological characteristics that differ significantly from common wheat, such as late-flowering, brittle rachis when naturally mature, hard glumes, high protein content, and barren tolerance (Sun et al., 1998; Chen et al., 2007; Zeng et al., 2010; Guo and Han, 2014). The use of this NAM population increased the number of QTLs identified and enhanced the mapping resolution in comparison to the bi-parental population analyses. We constructed an integrated map of the NAM population using 2,059 SNP markers with an average marker spacing distance of 1.13 cM. The novel integrated

TABLE 2 | The novel integrated genetic linkage map of the NAM population.

Chromosome	CY			HR			YN			YT			Integrated map		
	Coverage (cM)	Markers No.	Average spacing (cM)	Coverage (cM)	Markers No.	Average spacing (cM)	Coverage (cM)	Markers No.	Average spacing (cM)	Coverage (cM)	Markers No.	Average spacing (cM)	Coverage (cM)	Markers No.	Average spacing (cM)
1A	65.7	22	2.99	91.77	31	2.96	77.8	29	2.68	103.24	105	0.98	126.41	145	0.87
1B	52.54	25	2.1	64.99	25	2.6	95.52	54	1.77	78.51	46	1.71	186.86	139	1.34
1D	38.93	4	9.73	81.28	18	4.52	12.34	6	2.06	8.81	6	1.47	107.51	24	4.48
2A	87.5	24	3.65	135.49	48	2.82	78.6	32	2.46	70.76	30	2.36	110.61	96	1.15
2B	6.54	5	1.31	92.17	77	1.2	132.27	72	1.84	77.51	84	0.92	147.90	212	0.70
2D	3.51	2	1.75	53.48	11	4.86	11.08	5	2.22	38.67	13	2.97	86.92	27	3.22
3A	30.12	13	2.32	69.6	45	1.55	158.83	51	3.11	19.43	16	1.21	150.92	96	1.57
3B	163.64	39	4.2	76.43	32	2.39	91.67	53	1.73	49.97	6	8.33	115.48	107	1.08
3D	–	–	–	28.73	3	9.58	1.68	3	0.56	–	–	–	30.32	4	7.58
4A	6.73	3	2.24	86.63	38	2.28	98.37	35	2.81	76.52	45	1.7	158.20	107	1.48
4B	0.65	2	0.32	90.37	53	1.71	78.52	35	2.24	19.9	16	1.24	95.38	89	1.07
4D	–	–	–	0.6	5	0.12	0.62	2	0.31	–	–	–	0.61	5	0.12
5A	112.87	39	2.89	108.23	57	1.9	93.64	47	1.99	65.89	80	0.82	121.59	186	0.65
5B	57.73	8	7.22	130.48	48	2.72	136.46	80	1.71	189.31	104	1.82	167.19	212	0.79
5D	4.7	6	0.78	28.41	15	1.89	5.78	6	0.96	1.62	9	0.18	28.58	25	1.14
6A	36.21	10	3.62	9.83	29	0.34	15.68	21	0.75	41.4	68	0.61	45.13	113	0.40
6B	8.51	6	1.42	103.48	28	3.7	99	37	2.68	58.97	27	2.18	121.05	86	1.41
6D	29.8	9	3.31	21.23	5	4.25	44.55	8	5.57	41.3	9	4.59	85.35	28	3.05
7A	61.87	10	6.19	141.9	80	1.77	202.47	68	2.98	128.69	82	1.57	174.82	202	0.87
7B	22.5	15	1.5	118.61	67	1.77	74.79	33	2.27	55.25	47	1.18	121.47	134	0.91
7D	5.81	4	1.45	114.25	15	7.62	47.04	7	6.72	4.58	29	0.16	153.45	22	6.98
Total	795.85	246	3.24	1647.92	730	2.26	1556.74	684	2.28	1130.31	822	1.38	2355.75	2059	1.13

TABLE 3 | Stable additive QTLs for seedling traits in the NAM population.

Chromosome	Position	Right maker	Left maker	QTL detected for N^+, N^- treatments			Comparison with previous studies	
				N^+	N^-	N^+ and N^-	QTL	References
1A	115	*BobWhite_c12305_959*	*wsnp_Ex_c11939_19147790*		***QRDW-1A.1***		*QTdw-1A.1*	Sun et al., 2013
					QTDW-1A.1			
							QTdw.1, QRdw.1	Guo et al., 2012
1A	120	*BS00062759_51*	*BS00063847_51*			***QRSDW-1A.1***		
1B	37	*Tdurum_contig20299_368*	*BobWhite_c48550_198*		*QRSDW-1B.1* ***QSDW-1B.1***			
1B	42	*wsnp_Ex_c5098_9047611*	*BS00063092_51*		*QRSDW-1B.2*		*SDW-H, NUP-H, RDW-H,*	An et al., 2006
1B	44	*tplb0048b10_1365*	*Ku_c28580_432*	*QRDW-1B.1*	*QRSDW-1B.3*			
1B	50	*RAC875_c1785_366*	*BS00082071_51*	***QSDW-1B.2***				
1B	52	*wsnp_Ex_c1440_2764269*	*BS00072289_51*	***QRDW-1B.2***				
1B	58	*wsnp_Ex_c21559_30710510*	*Kukri_c44191_452*			***QSDW-1B.3***	*TN-H*	An et al., 2006
							cTRL	Laperche et al., 2006
1B	1	*Kukri_c8390_1102*	*BS00080212_51*			***QSDW-1B.4***		
1B	55	*Ex_c29452_302*	*BobWhite_c20073_382*			*QRSDW-1B.4*		
1D	26	*BS00042197_51*	*RAC875_c62_1514*		***QRDW-1D.1***		*Qsnp-1D*	Xu et al., 2014
							RFW, RDW	Zhang et al., 2013
2A	31	*IAAV7468*	*Tdurum_contig66015_346*		*QRDW-2A.1*		*R-GS*	Fontaine et al., 2009
							QTkw-2A.1, QKl-2A.1	Cui et al., 2014
2A	48	*RAC875_c20700_853*	*RAC875_c13116_943*	***QTDW-2A.1***		*QSDW-2A.1*	*QTkw-2A.2*	Cui et al., 2014
2A	52	*Tdurum_contig47258_1039*	*wsnp_Ex_c22645_31845564*	***QRDW-2A.2***	*QSDW-2A.2* *QTDW-2A.2*		*QKnps-2A.2*	Cui et al., 2014
2A	72	*Ra_c26702_797*	*Excalibur_c96_619*	***QRDW-2A.3***			*NS%, GPC*	Laperche et al., 2007
2A	75	*BobWhite_c16923_64*	*RAC875_c104160_61*			***QRDW-2A.4***		
2A	76	*Tdurum_contig60205_806*	*Excalibur_rep_c66618_87*	***QSDW-2A.3***				
2A	81	*Ku_c13700_1196*	*BS00107804_51*	*QRDW-2A.5*		***QTDW-2A.3***		
2A	84	*IAAV2861*	*wsnp_Ex_rep_c70299_69243401*	***QRDW-2A.6***	*QTDW-2A.4*		*SDW*	Zhang et al., 2013
2A	94	*BobWhite_c17783_174*	*CAP7_c1527_136*	***QRDW-2A.7***				
2A	97	*Tdurum_contig47508_250*	*BS00053834_51*	***QRDW-2A.8***				
2B	10	*BS00099658_51*	*Excalibur_c17250_592*	***QSDW-2B.1***			*QGY-2B*	Xu et al., 2014
							R-GS, A-GS, A-DTH	Fontaine et al., 2009
							LA, cTDM, tADM	Laperche et al., 2006

Continued

TABLE 3 | Continued

Chromosome	Position	Right maker	Left maker	QTL detected for N^+, N^- treatments			Comparison with previous studies	
				N^+	N^-	N^+ and N^-	QTL	References
2B	40	*Ex_c12051_875*	*BS00041585_51*		*QSDW-2B.2*			
2B	73	*RFL_Contig4849_702*	*BS00066545_51*	***QRDW-2B.1***				
2B	77	*RFL_Contig1987_3440*	*Tdurum_contig26542_457*	***QTDW-2B.1***				
2B	78	*Excalibur_rep_c66577_159*	*RFL_Contig4718_1269*	***QRDW-2B.2***				
2B	80	*GENE-4029_80*	*wsnp_BE490763B_Ta_2_1*			***QRDW-2B.3***	*R-GDH, R-NH4*	Fontaine et al., 2009
							TDM, tRDM, TRM, TRL, RUE	Laperche et al., 2006
2B	91	*Tdurum_contig28795_322*	*wsnp_Ex_c45468_51254832*		***QTDW-2B.2***			
2B	93	*RAC875_c35778_201*	*BS00094578_51*		***QTDW-2B.3***			
2B	94	*BS00022374_51*	*BS00009060_51*	***QRDW-2B.4***				
2B	98	*Excalibur_c42146_266*	*BS00026432_51*		***QTDW-2B.4***			
2B	108	*BS00060618_51*	*wsnp_RFL_Contig2324_1803878*	***QRDW-2B.5***				
3A	4	*Excalibur_c34889_526*	*Jagger_c6722_104*	***QRDW-3A.1***				
3A	7	*RAC875_c787_431*	*Ku_c17560_91*	***QRDW-3A.2***				
3A	10	*BS00036492_51*	*BS00022129_51*		*QRSDW-3A.1*	***QSDW-3A.1*** ***QRDW-3A.3*** ***QTDW-3A.1***	*TN-H, RDW-H*	An et al., 2006
							GPA-3A	Laperche et al., 2007
3B	63	*wsnp_Ku_c29102_39008953*	*wsnp_Ex_c64005_62986957*	*QTDW-3B.1*				
3B	115	*IAAV8892*	*BS00071041_51*	***QRDW-3B.1***			*QTKW.3B*	Xu et al., 2014
							QRFW.WL.3B	Zhang et al., 2013
4A	17	*BobWhite_rep_c66057_98*	*wsnp_Ex_c5690_9994305*			***QRDW-4A.1***	*NUP-LN2*	An et al., 2006
4A	21	*Ra_c7973_1185*	*IAAV4609*	*QTDW-4A.1*				
4A	22	*Ku_c11865_406*	*wsnp_Ex_c31508_40288653*	***QSDW-4A.1***				
5A	24	*BS00047242_51*	*Excalibur_rep_c95828_165*	*QTDW-5A.1*				
5A	30	*wsnp_Ex_c18941_27840714*	*Tdurum_contig10601_289*		*QRSDW-5A.1*		*NUP-HN1*	An et al., 2006
5B	30	*RAC875_c29907_115*	*BS00015136_51*	***QTDW-5B.1***				
6B	10	*wsnp_Ku_c24391_34351602*	*Kukri_c23491_274*	*QSDW-6B.1*			*NUP-LN1*	An et al., 2006
6B	19	*CAP11_c7959_386*	*wsnp_Ex_c7191_12352173*	***QRSDW-6B.1***	*QRDW-6B.1*			
6B	48	*wsnp_Ra_c20409_29673950*	*BS00033629_51*	***QRSDW-6B.2***				
6D	85	*BS00022787_51*	*IACX5958*	***QTDW-6D.1***				
7A	7	*Tdurum_contig5646_929*	*Excalibur_c20311_240*		*QRDW-7A.1*		*NS%-7A1*	Laperche et al., 2007
							SFW	Zhang et al., 2013
							QTkw-7A.2	Cui et al., 2014

Continued

TABLE 3 | Continued

Chromosome	Position	Right maker	Left maker	QTL detected for N^+, N^- treatments			Comparison with previous studies	
				N^+	N^-	N^+ and N^-	QTL	References
7A	90	*RFL_Contig5285_365*	*BS00108184_51*			***QTDW-7A.1***	*QSLFW.WY.7A*	Zhang et al., 2013
7A	105	*Excalibur_c47990_159*	*wsnp_Ku_c10202_16937059*	***QRDW-7A.2***			*NUP-H*	An et al., 2006
7A	113	*CAP7_c7296_88*	*BobWhite_c5125_258*			*QTDW-7A.2*		
7B	0	*wsnp_Ex_c7934_13467460*	*Tdurum_contig55961_526*		*QSDW-7B.2QTDW-7B.1*		*NUR*	Laperche et al., 2006
							NUP	Laperche et al., 2006
7B	35	*BobWhite_c10448_80*	*Jagger_c9314_100*	***QSDW-7B.1***			*NS%-7B1*	Laperche et al., 2007
							cLRN_PRL	Laperche et al., 2006

[a]*QTLs marked with a bold typeface represent primary QTLs.*

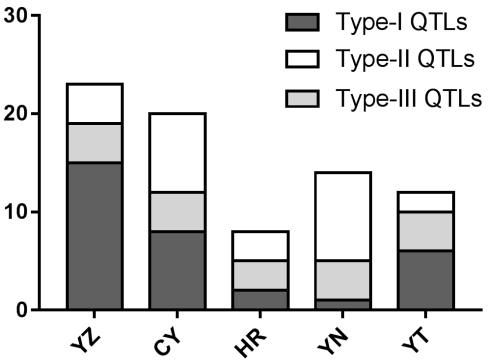

FIGURE 3 | Statistics of favorable QTLs/allele donated by the parents in this study. The identified QTLs were divided into three types: Type-I (QTLs detected only under N^+ conditions), Type-II (QTLs detected only under N^- conditions), and Type-III (QTLs detected under both conditions). Statistics of the Type-I, Type-II, and Type-III QTLs are indicated with gray, white, and light gray columns, respectively. The *x*-axis represents the five parental lines: YZ (Yanzhan 1), CY (Cayazheda), HR (Hussar), YN (Yunnanxiaomai), and YT (Yutiandaomai).

genetic map shows good genome coverage, high density, and good collinearity with physical maps, and is thus more suitable for genetic research than the four individual genetic maps.

Mapping of QTLs Involved in N Deficiency Tolerance

N uptake and utilization at the seedling stage are important for accumulating a N reservoir, which then fulfills the N requirements during plant growth until the maturity stage (Lian et al., 2005). Genotype selection based on comprehensive performance under N^+ and N^- conditions would be valuable for evaluating N deficiency tolerance (Fontaine et al., 2009; Wang et al., 2017). In this study, we tested two different N supply levels, namely, N^+ (normal nitrogen supplement) and N^- (low nitrogen supplement), under hydroponic culture conditions. The identified QTLs could be divided into three types: Type-I (QTLs detected only under N^+ conditions), Type-II (QTLs detected only under N^- conditions), and Type-III (QTLs detected under both conditions). We identified 14 Type-III QTLs that were indispensable for constitutive processes, with polymorphisms existing between their parents (Laperche et al., 2007). Thirty-one Type-I QTLs identified in our study were assumed to be associated with high levels of N uptake or utilization. Fifteen favorable alleles of Type-I QTLs were donated by parent YZ. *QRDW-2A.2* was mapped to the chromosomal region associated with NS% (straw nitrogen content) and GPC (grain protein content) reported by Laperche et al. (2007). *QSDW-6B.1* and *QRDW-7A.2* have been reported to affect NUP (root N content; An et al., 2006). We detected 23 Type-II QTLs involved in N-deficiency tolerance. These favorable alleles were mainly donated by parents CY and YN. Type-II QTLs, namely *QRSDW-1B.2*, *QRDW-1D.1*, *QRDW-2A.1*, *QSDW-2A.2*, *QTDW-2A.2*, *QRSDW-5A.1*, *QSDW-7B.2,* and *TDW-7B.1,* have been reported to influence N deficiency tolerance and related

traits in previous studies (An et al., 2006; Laperche et al., 2006, 2007; Habash et al., 2007; Fontaine et al., 2009; Guo et al., 2012; Sun et al., 2013; Zhang et al., 2013; Cui et al., 2014; Xu et al., 2014). The remaining common QTLs/genome regions are listed in **Table 3**. The coincidence of the QTLs across different genetic backgrounds not only implies the reliability of the QTLs detected in this study, but also highlights the importance of the chromosomal region.

The "global" interaction variable has previously been used to characterize plant responses to stress (Yan et al., 1999; Yadav et al., 2003; Lian et al., 2005; Laperche et al., 2006, 2007). In this study, we compared two QTL sets detected under the two N levels, from which 23 Type-II QTLs were discovered to be involved in N deficiency tolerance. To further distinguish the QTLs specifically involved in the adaptation of wheat to N deficiency, the "global" interaction variable of $(N_y^- - N_y^+)/N_y^+$ was alternatively used for QTL detection. We hypothesized that the QTLs identified both by the "global" interaction variable and by the N^- treatments constituted high confidence QTLs involved in N deficiency tolerance. Four QTLs (*QRSDW-1B.2*, *QRDW-2A.1*, *QRSDW-5A.1,* and *QRDW-7A.1*) were identified that met these criteria, and have previously been reported to influence several traits (**Table 3**), including SDW, NUP (root N content), RDW (An et al., 2006), NUP, and NS% (Laperche et al., 2007), R-GS (glutamine synthetize activity, Fontaine et al., 2009), and TKW (thousand kernel weight, Cui et al., 2014).

Implications for Breeding

The size and topology of the root system determines the N uptake ability of the plant (Lea and Morot-Gaudry, 2001). When N is limited or deficient, wheat responds by increasing root growth and proliferation at the expense of the shoots, leading to high root/shoot ratios (Ericsson, 1995; Ameziane et al., 1997). As N uptake during the vegetative stage plays an important role in plant growth even into maturity, breeders select wheat genotypes that perform well under both N^+ and N^- conditions (Fontaine et al., 2009). In the NAM population, the phenotypic variation of the parents resulted in a rich allelic variation in response to N fluctuation. RDW, SDW, and TDW were high in the parents YZ, HR, and YT under N^+ conditions, but these traits decreased substantially under N^- conditions. In comparison, RDW, SDW, and TDW were lower in parents CY and YN under N^+ conditions, but increased under N^- conditions. Many RILs had greater RDW, SDW, and TDW under N^- conditions than under N^+ conditions; for instance, RDW, SDW, and TDW were greater in many RILs under N^- conditions in comparison to their parents. This can be explained as the pyramiding of favorable alleles from both parents, which is valuable for the breeding of wheat varieties tolerant of low N levels.

To develop wheat varieties adapted to limited or deficient N conditions, direct selection for favorable QTLs specifically detected under N^- condition is effective. We identified eight primary QTLs (*QRDW-1A.1*, *QTDW-1A.1*, *QSDW-1B.1*, *QRDW-1D.1*, *QTDW-2B.2*, *QTDW-2B.3*, *QTDW-2B.4,* and *QRDW-6A.1*), all of which are probably involved in N-deficiency tolerance. These QTLs are of value in wheat breeding programs designed to increase N deficiency tolerance. Moreover, N uptake or utilization traits have been considered as indirect selection criteria for the improvement of N-deficiency tolerance (Lian et al., 2005; Fontaine et al., 2009; Wang et al., 2017). In this study, we also identified 25 primary QTLs implicated in N uptake and utilization, and 11 primary QTLs associated with constitutive process (**Table 3**). For instance, *QTDW-3A.1* showed multiple effects on biomass, grain number and yield in the mature periods; *QTDW-5A.1* was also mapped to chromosomal region affecting thousand kernel weight in the mature periods (**Supplementary Table S7**). The QTLs also could be used in breeding programs by pyramiding the different types of QTLs or by using pleiotropic QTLs through MAS. Thus, the mapped QTL interval markers could be used in MAS after being converted into high-throughput KASP (Kompetitive Allele Specific PCR) markers.

In this study, 23 favorable QTLs were donated by the domesticated cultivar of Yanzhan 1, in which 15 were Type-I (detected only in N^+ conditions) and only four were Type-II (detected only in N^- conditions). In contrast, the semi-wild cultivars contributed more favorable Type-II QTLs, including eight favorable QTLs from CY and nine from YN. Seven Type-II favorable QTLs (*QRSDW-1B.1*, *QSDW-1B.1*, *QRSDW-1B.3*, *QTDW-2B.2*, *QTDW-2B.3*, *QTDW-2B.4,* and *QRDW-6B.1*), donated by CY and YN, are novel QTLs that have not been reported in previous studies. The modern variety (YZ) possessed more favorable QTLs/genes for N uptake and utilization under N^+ conditions, while the semi-wild wheat varieties were more likely to have favorable QTLs/genes for N-deficiency tolerance (**Figure 3** and **Supplementary Table S8**). This indicates that a domesticated selection might have occurred in the breeding process. Modern domesticated varieties are supplied with adequate N during yield experiments, and therefore lines that use more N to increase yield are more likely to be selected for cultivation. The semi-wild wheat varieties were from wilderness areas with limited N, and are thus subject to strong evolutionary pressure to maintain N-deficiency tolerance. Semi-wild wheat varieties are therefore an important genetic resource that can be used to improve the N-deficiency tolerance of modern varieties.

The "global" interaction variable identified *QRSDW-1B.2*, *QRSDW-5A.1*, and *QRDW-7A.1* as high confidence QTLs involved in N stress adaption, with favorable alleles donated by the semi-wild wheat YN. *QRSDW-5A.1* with positive alleles increased the RSDW value from 18.9% to 22.7%, indicating tremendous potential for its application in wheat breeding programs designed to increase N-deficiency tolerance. We predicted that the candidate genes for *QRSDW-5A.1* might be within the 0.7 cM confidence interval of *wsnp_Ex_c18941_27840714–Tdurum_contig10601_289*. Based on our integrated genetic map, which had high density and good collinearity with the physical map, we further compared the overlapping intervals of the collocated QTL peaks with the IWGSC RefSeq Annotations database v 1.0.[5] The confidence intervals of *wsnp_Ex_c18941_27840714–Tdurum_contig10601_289* spanned 0.8 Mb (5A: 547647367–548503773). This region harbors 12

[5]https://wheat-urgi.versailles.inra.fr/Seq-Repository/Annotations

annotated genes in wheat (**Supplementary Table S9**), most notably an auxin responsive gene (*ARF*) cluster (including eight genes), which might be a candidate for *QRSDW-5A.1*. This information provides a reference for the future high-resolution mapping and map-based cloning of *QRSDW-5A.1*.

CONCLUSION

A NAM population comprised of four RIL populations was used for QTL mapping. An integrated genetic map of wheat, with high density and good collinearity with the physical maps, was developed. The NAM population was highly variable for all of the measured traits. Many RILs tolerant of N deficiency exhibited high RDW, SDW, and TDW under the N^- treatment. We detected 31 QTLs under N^+ conditions that are possibly involved in N uptake or utilization, with favorable alleles mainly donated by the modern parent YZ. We detected 23 QTLs under N^- conditions, possibly associated with N-deficiency tolerance, with most of the favorable being alleles donated by the semi-wild parents CY and YN. Four QTLs detected under N^- conditions were identified as high confidence QTLs involved in N-deficiency tolerance. A domesticated selection might have occurred during the breeding process. Semi-wild wheat varieties constitute an important genetic resource that can be used to improve the N-deficiency tolerance of modern varieties.

AUTHOR CONTRIBUTIONS

DR and XF designed the experiments. GZ and JH created the mapping population. DR, QM, WC, and SL carried out phenotypic experiments. DR analyzed experimental results. XF and PJ analyzed Illumina 90K assay sequencing data. XM assisted with Illumina sequencing. DR, XW, HW, and LK wrote the manuscript.

FUNDING

This work was supported by the 973 Program of China (2014CB138100) and National Natural Science Foundation of China (31171553, 31520103911 and 31471488).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00845/full#supplementary-material

FIGURE S1 | Overlap of QTLs detected for all of the traits under N^+ and N^- treatment. **(A–E)** Statistical overlap of the QTLs detected for all of the traits under N^+ and N^- conditions are shown by Venn diagrams. The gray circle indicates the QTLs detected only under N^+ conditions; the white circle indicates QTLs detected only under N^- conditions; and the light gray circle indicates the QTLs detected for both conditions.

TABLE S1 | Nutrient solution ingredients for wheat seedling growth.

TABLE S2 | Phenotype distribution of the parents and individual population.

TABLE S3 | ANOVA analysis.

TABLE S4 | The integrative genetic map.

TABLE S5 | QTL detected for N^+, N^-, and G × N.

TABLE S6 | Distribution of QTLs detected for growth traits in our study.

TABLE S7 | QTL detected for yield traits in mature periods.

TABLE S8 | Favorable QTLS/alleles donated by parents in this study.

TABLE S9 | Candidate genes.

REFERENCES

Alboresi, A., Gestin, C., Leydecker, M. T., Bedu, M., Meyer, C., and Truong, H. N. (2005). Nitrate, a signal relieving seed dormancy in Arabidopsis. *Plant Cell Environ.* 28, 500–512. doi: 10.1111/j.1365-3040.2005.01292.x

Ameziane, R., Deleens, E., Noctor, G., Morot-Gaudry, J. F., and Limami, M. A. (1997). Stage of development is an important determinant in the effect of nitrate on photoassimilate (l3C) partitioning in chicory (*Cichorium intybus*). *J. Exp. Bot.* 48, 25–33. doi: 10.1093/jxb/48.1.25

An, D. G., Su, J. Y., Liu, Q. Y., Zhu, Y. G., Tong, Y. P., Li, J. M., et al. (2006). Mapping QTLs for nitrogen uptake in relation to the early growth of wheat (*Triticum aestivum* L.). *Plant Soil* 284, 73–84. doi: 10.1007/s11104-006-0030-3

Bryman, A., and Cramer, D. (2012). Quantitative data analysis with IBM SPSS 17, 18 and 19: a guide for social scientists. *Int. Stat. Rev.* 80, 334–335. doi: 10.1111/j.1751-5823.2012.00187_14.x

Chen, F., Yu, Y. X., Xia, X. C., and He, Z. H. (2007). Prevalence of a novel puroindoline b allele in Yunnan endemic wheats (*Triticum aestivum* ssp. yunnanense King). *Euphytica* 156, 39–46. doi: 10.1007/s10681-006-9347-5

Cui, F., Fan, X. L., Chen, M., Zhang, N., Zhao, C. H., Zhang, W., et al. (2016). QTL detection for wheat kernel size and quality and the responses of these traits to low nitrogen stress. *Theor. Appl. Genet.* 129, 469–484. doi: 10.1007/s00122-015-2641-7

Cui, F., Zhao, C. H., Ding, A. M., Li, J., Wang, L., Li, X. F., et al. (2014). Construction of an integrative linkage map and QTL mapping of grain yield-related traits using three related wheat RIL populations. *Theor. Appl. Genet.* 127, 659–675. doi: 10.1007/s00122-013-2249-8

Dong, W., Cui, K. H., Ye, G. Y., Pan, J. F., Xiang, J., Huang, J. L., et al. (2014). QTL mapping for nitrogen use efficiency and nitrogen deficiency tolerance traits in rice. *Plant Soil* 359, 281–295.

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Ericsson, T. (1995). Growth and shoot : root ratio of seedlings in relation to nutrient availability. *Plant Soil* 168, 205–214. doi: 10.1007/BF00029330

FAOSTAT (2015). *FAO Statistical Databases. Food and Agriculture Organization (FAO) of the United Nations*. Rome: FAOSTAT.

Fontaine, J. X., Ravel, C., Pageau, K., Heumez, E., Dubois, F., Hirel, B., et al. (2009). A quantitative genetic study for elucidating the contribution of glutamine synthetase, glutamate dehydrogenase and other nitrogen-related physiological traits to the agronomic performance of common wheat. *Theor. Appl. Genet.* 119, 645–662. doi: 10.1007/s00122-009-1076-4

Gaju, O., Allard, V., Martre, P., Snape, J. W., Heumez, E., LeGouis, J., et al. (2011). Identification of traits to improve the nitrogen-use efficiency of wheat genotypes. *Field Crop Res.* 123, 139–152. doi: 10.1016/j.fcr.2011.05.010

Gallais, A., and Hirel, B. (2004). An approach to the genetics of nitrogen use efficiency in maize. *J. Exp. Bot.* 55, 295–306. doi: 10.1093/jxb/erh006

Gan, Y., Filleur, S., Rahman, A., Gotensparre, S., and Forde, B. G. (2005). Nutritional regulation of *ANR1* and other root-expressed MADS-box genes in *Arabidopsis thaliana*. *Planta* 222, 730–742. doi: 10.1007/s00425-005-0020-3

Guo, X., and Han, F. P. (2014). Asymmetric epigenetic modification and elimination of rDNA sequences by polyploidization in Wheat. *Plant Cell* 26, 4311–4327. doi: 10.1105/tpc.114.129841

Guo, Y., Kong, F. M., Xu, Y. F., Zhao, Y., Liang, X., Wang, Y. Y., et al. (2012). QTL mapping for seedling traits in wheat grown under varying concentrations of N, P and K nutrients. *Theor. Appl. Genet.* 124, 851–865. doi: 10.1007/s00122-011-1749-7

Habash, D. Z., Bernard, S., Schondelmaier, J., Weyen, J., and Quarrie, S. A. (2007). The genetics of nitrogen use in hexaploid wheat: N utilisation, development and yield. *Theor. Appl. Genet.* 114, 403–419. doi: 10.1007/s00122-006-0429-5

Hirel, B., Bertin, P., Quillere, I., Bourdoncle, W., Attagnant, C., Dellay, C., et al. (2001). Towards a better understanding of the genetic and physiological basis for nitrogen use efficiency in maize. *Plant Physiol.* 125, 1258–1270. doi: 10.1104/pp.125.3.1258

Hoagland, D. R., and Arnon, D. I. (1950). The water culture method for growing plants without soil. *Calif. Agric. Exp. Stn. Circ.* 347, 357–359.

Kiba, T., and Krapp, A. (2016). Plant nitrogen acquisition under low availability: regulation of uptake and root architecture. *Plant Cell Physiol.* 57, 707–714. doi: 10.1093/pcp/pcw052

Kosambi, D. D. (1944). The estimation of map distances from recombination values. *Ann. Eugen.* 12, 172–175. doi: 10.1111/j.1469-1809.1943.tb02321.x

Krouk, G., Lacombe, B., and Bielach, A. (2010). Nitrate-regulated auxin transport by NRT1.1 defines a mechanism for nutrient sensing in plants. *Dev. Cell* 18, 927–937. doi: 10.1016/j.devcel.2010.05.008

Krouk, G., Tillard, P., and Gojon, A. (2006). Regulation of the high-affinity NO3– uptake system by NRT1.1-mediated NO3– demand signaling in Arabidopsis. *Plant Physiol.* 142, 1075–1086. doi: 10.1104/pp.106.087510

Laperche, A., Brancourt, H. M., Heumez, E., Gardet, O., Hanocq, E., Devienne, B. F., et al. (2007). Using genotype × nitrogen interaction variables to evaluate the QTL involved in wheat tolerance to nitrogen constraints. *Theor. Appl. Genet.* 115, 399–415. doi: 10.1007/s00122-007-0575-4

Laperche, A., Devienne, B. F., Maury, O., Le, G. J., and Ney, B. (2006). A simplified conceptual model of carbon/nitrogen functioning for QTL analysis of winter wheat adaptation to nitrogen deficiency. *Theor. Appl. Genet.* 113, 1131–1146. doi: 10.1007/s00122-006-0373-4

Laperche, A., Le, G. J., Hanocq, E., and Brancourt, H. M. (2008). Modelling nitrogen stress with probe genotypes to assess genetic parameters and genetic determinism of winter wheat tolerance to nitrogen constraint. *Euphytica* 161, 259–271. doi: 10.1007/s10681-007-9433-3

Lea, P. J., and Morot-Gaudry, J. F. (2001). *Plant nitrogen*, ed. J. Lea (Berlin: Springer Verlag), 362–374. doi: 10.1007/978-3-662-04064-5

Li, C., Li, Y., Sun, B., Peng, B., Liu, C., and Liu, Z. (2013). Quantitative trait loci mapping for yield components and kernel-related traits in multiple connected RIL populations in maize. *Euphytica* 193, 303–316. doi: 10.1007/s10681-013-0901-7

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. Breed.* 20, 167–178. doi: 10.1007/s11032-007-9080-3

Lian, X., Xing, Y., Yan, H., Xu, C., Li, X., and Zhang, Q. (2005). QTLs for low nitrogen tolerance at seedling stage identified using a recombinant inbred line population derived from an elite rice hybrid. *Theor. Appl. Genet.* 112, 85–96. doi: 10.1007/s00122-005-0108-y

Liu, Q. P. (2017). Spatio-temporal changes of fertilization intensity and environmental safety threshold in China. *Trans. Chin. Soc. Agric. Eng.* 33, 214–221.

Marchive, C., Roudier, F., Castaings, L., Brehaut, V., Blondet, E., Colot, V., et al. (2013). Nuclear retention of the transcription factor *NLP7* orchestrates the early response to nitrate in plants. *Nat. Commun.* 4:1713. doi: 10.1038/ncomms2650

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., and Sun, Q. (2009). Genetic properties of the maize nested association mapping population. *Science* 325, 737–740. doi: 10.1126/science.1174320

Ogut, F., Bian, Y., Bradbury, P. J., and Holland, J. B. (2015). Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. *Heredity* 114, 552–563. doi: 10.1038/hdy.2014.123

Peng, M., Hannam, C., Gu, H., Bi, Y. M., and Rothstein, S. J. (2007). A mutation in NLA, which encodes a RING-type ubiquitin ligase, disrupts the adaptability of Arabidopsis to nitrogen limitation. *Plant J.* 50, 320–337. doi: 10.1111/j.1365-313X.2007.03050.x

Peng, S. B., Tang, Q. Y., and Zou, Y. B. (2009). Current status and challenges of rice production in China. *Plant Prod. Sci.* 12, 3–8. doi: 10.1626/pps.12.3

Quarrie, S. A., Steed, A., Calestani, C., Semikhodskii, A., Lebreton, C., Chinoy, C., et al. (2005). A high-density genetic map of hexaploid wheat *Triticum aestivum* L. from the cross Chinese Spring 9 × SQ1 and its use to compare QTLs for grain yield across a range of environments. *Theor. Appl. Genet.* 110, 865–880. doi: 10.1007/s00122-004-1902-7

Sandrine, R., Alain, G., and Laurence, L. (2014). Signal interactions in the regulation of root nitrate uptake. *J. Exp. Bot.* 65, 5509–5517. doi: 10.1093/jxb/eru321

Sophie, B., Marcus, O., Sandeep, M., Brian, W., Ram, P., Tesfaye, T., et al. (2015). Power and resolution of QTL mapping in sorghum using a nested association mapping population and diversity panels. *Genet* 206, 573–585.

Sun, J. J., Guo, Y., Zhang, G. Z., Gao, M. G., Zhang, G. H., Kong, F. M., et al. (2013). QTL mapping for seedling traits under different nitrogen forms in wheat. *Euphytica* 191, 317–331. doi: 10.1007/s10681-012-0834-6

Sun, Q. X., Ni, Z. F., Liu, Z. Y., Gao, J. W., and Huang, T. C. (1998). Genetic relationships and diversity among Tibetan wheat, common wheat and European spelt wheat revealed by RAPD markers. *Euphytica* 99, 205–211. doi: 10.1023/A:1018316129246

Van Ooijen, J. W., and Jansen, J. (2013). Genetic mapping in experimental populations. *Camb. Univ. Press* 49, 701–707. doi: 10.1017/CBO9781139003889

Vatter, T., Maurer, A., Kopahnke, D., Perovic, D., Ordon, F., and Pillen, K. (2017). A nested association mapping population identifies multiple small effect QTL conferring resistance against net blotch (*Pyrenophora teres* f. teres) in wild barley. *PLoS One* 12:0186803. doi: 10.1371/journal.pone.0186803

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, S. H., Debbie, W. G., Kerrie, F., Alexandra, A., Shaoman, C., Bevan, E. H., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90000 single nucleotide polymorphism array. *Plant Biotechnol.* 12, 787–796. doi: 10.1111/pbi.12183

Wang, T., Li, C. X., Wu, Z. H., Jia, Y. C., Wang, H., Sun, S. Y., et al. (2017). Abscisic acid regulates auxin homeostasis in rice root tips to promote root hair elongation. *Front. Plant Sci.* 8:1121. doi: 10.3389/fpls.2017.01121

Wilde, F., Schön, C. C., Korzun, V., Ebmeyer, E., Schmolke, M., Hartl, L., et al. (2008). Marker-based introduction of three quantitative-trait loci conferring resistance to Fusarium head blight into an independent elite winter wheat breeding population. *Theor. Appl. Genet.* 117, 29–35. doi: 10.1007/s00122-008-0749-8

Xu, Y. F., Wang, R. F., Tong, Y. P., Zhao, H. T., Xie, Q. G., Liu, D. G., et al. (2014). Mapping QTLs for yield and nitrogen related traits in wheat: influence of nitrogen and phosphorus fertilization on QTL expression. *Theor. Appl. Genet.* 127, 59–72. doi: 10.1007/s00122-013-2201-y

Yadav, R., Bidinger, R. F., Hash, C., Yadav, Y., Yadav, O., Bhatnagar, S., et al. (2003). Mapping and characterisation of QTL × E interactions for traits determining grain and stover yield in pearl millet. *Theor. Appl. Genet.* 106, 512–520. doi: 10.1007/s00122-002-1081-3

Yan, J., Zhu, J., He, C., Benmoussa, M., and Wu, P. (1999). Molecular markerassisted dissection of genotype × environment interaction for plant type traits in rice (*Oryza sativa* L.). *Crop Sci.* 39, 538–544. doi: 10.2135/cropsci1999.0011183X003900020039x

Yao, Q., Zhou, R. H., Pan, Y. M., Fu, T. H., and Jia, J. Z. (2010). Construction of genetic linkage map and QTL analysis of agronomic important traits based on a RIL population derived from common wheat variety Yanzhan 1 and Zaosui 30. *Sci. Agric. Sin.* 43, 4130–4139.

Yu, J., Holland, J. B., McMullen, M. D., and Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genet* 178, 539–551. doi: 10.1534/genetics.107.074245

Yuan, S., Zhang, Z. W., Zheng, C., Zhao, Z. Y., Wang, Y., Feng, L. Y., et al. (2016). Arabidopsis cryptochrome 1 functions in nitrogen regulation of flowering. *Proc. Natl. Acad. Sci. U.S.A.* 113, 7661–7666. doi: 10.1073/pnas.1602004113

Zeng, X. Q., Wang, Y. G., Li, W. Y., Wang, C. Y., Liu, X. L., and Ji, W. Q. (2010). Comparison of the genetic diversity between *Triticum aestivum* ssp. tibetanum Shao and Tibetan wheat landraces (*Triticum aestivum* L.) by using intron-splice junction primers. *Genet. Resour. Crop Evol.* 57, 1141–1150. doi: 10.1007/s10722-010-9553-9

Zhang, H., Cui, F., Wang, L., Li, J., Ding, A. M., Zhao, C. H., et al. (2013). Conditional and unconditional QTL mapping of drought-tolerance-related traits of wheat seedling using two related RIL populations. *Genet* 92, 213–231. doi: 10.1007/s12041-013-0253-z

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: 03 May 2018
doi: 10.3389/fpls.2018.00604

Genome Sequencing and Analysis of the Peanut B-Genome Progenitor (*Arachis ipaensis*)

Qing Lu[1†], Haifen Li[1†], Yanbin Hong[1†], Guoqiang Zhang[2], Shijie Wen[1], Xingyu Li[1], Guiyuan Zhou[1], Shaoxiong Li[1], Hao Liu[1], Haiyan Liu[1], Zhongjian Liu[2], Rajeev K. Varshney[3,4], Xiaoping Chen[1] and Xuanqiang Liang[1*]*

[1] South China Peanut Sub-Center of National Center of Oilseed Crops Improvement, Guangdong Provincial Key Laboratory of Crop Genetic Improvement,Crops Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou, China, [2] Shenzhen Key Laboratory for Orchid Conservation and Utilization, National Orchid Conservation Center of China and Orchid Conservation and Research Center of Shenzhen, Shenzhen, China, [3] International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, India, [4] School of Plant Biology, The Institute of Agriculture, University of Western Australia, University of Western Australia, Crawley, WA, Australia

OPEN ACCESS

Edited by:
Jun Yang,
Shanghai Chenshan Plant Science Research Center (CAS), China

Reviewed by:
Xuehui Huang,
Shanghai Normal University, China
Junjie Fu,
Institute of Crop Sciences (CAAS), China

***Correspondence:**
Xiaoping Chen
chenxiaoping@gdaas.cn
Xuanqiang Liang
liangxuanqiang@gdaas.cn

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

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: 06 February 2018
Accepted: 16 April 2018
Published: 03 May 2018

Citation:
Lu Q, Li H, Hong Y, Zhang G, Wen S, Li X, Zhou G, Li S, Liu H, Liu H, Liu Z, Varshney RK, Chen X and Liang X (2018) Genome Sequencing and Analysis of the Peanut B-Genome Progenitor (Arachis ipaensis). Front. Plant Sci. 9:604. doi: 10.3389/fpls.2018.00604

Peanut (*Arachis hypogaea* L.), an important leguminous crop, is widely cultivated in tropical and subtropical regions. Peanut is an allotetraploid, having A and B subgenomes that maybe have originated in its diploid progenitors *Arachis duranensis* (A-genome) and *Arachis ipaensis* (B-genome), respectively. We previously sequenced the former and here present the draft genome of the latter, expanding our knowledge of the unique biology of *Arachis*. The assembled genome of *A. ipaensis* is ∼1.39 Gb with 39,704 predicted protein-encoding genes. A gene family analysis revealed that the FAR1 family may be involved in regulating peanut special fruit development. Genomic evolutionary analyses estimated that the two progenitors diverged ∼3.3 million years ago and suggested that *A. ipaensis* experienced a whole-genome duplication event after the divergence of *Glycine max*. We identified a set of disease resistance-related genes and candidate genes for biological nitrogen fixation. In particular, two and four homologous genes that may be involved in the regulation of nodule development were obtained from *A. ipaensis* and *A. duranensis*, respectively. We outline a comprehensive network involved in drought adaptation. Additionally, we analyzed the metabolic pathways involved in oil biosynthesis and found genes related to fatty acid and triacylglycerol synthesis. Importantly, three new *FAD2* homologous genes were identified from *A. ipaensis* and one was completely homologous at the amino acid level with FAD2 from *A. hypogaea*. The availability of the *A. ipaensis* and *A. duranensis* genomic assemblies will advance our knowledge of the peanut genome.

Keywords: *Arachis ipaensis*, genome sequence, genome evolution, polyploidizations, whole genome duplication

INRODUCTION

Peanut (*Arachis hypogaea* L.) is a grain legume and oilseed crop that is an important source of vegetable oil and protein. It is widely cultivated in tropical and subtropical regions. In Africa and some Asia countries, peanut is more prevalent than any other leguminous crop, including soybean. With an annual production of ∼46 million tons and a remarkable 45–56% oil content, it plays a key role in daily human nutrition. Moreover, peanut oil is important to human health owing to its rich

nutritional elements, such as oleic acid, linoleic acid, resveratrol, fiber, and vitamins (Parthasarathy et al., 1990).

The *Arachis* genus originated in South America and is composed of about 80 diploid species that have been divided taxonomically into nine sections (Krapovickas and Gregory, 1994). *Arachis* species have an unusual reproductive biology in that all members have a geocarpic reproductive habit, with unique growth characteristics of aerial flowers and subterranean fruit (Smith, 1950), that allows them to adapt to particularly harsh environments (Tan et al., 2010). *A. hypogaea*, cultivated peanut or groundnut, is an allotetraploid (2n = 4x = 40), with an AABB genomic constitution (Temsch and Greilhuber, 2000), which was probably derived from a single recent hybridization of two diploid progenitors (Kochert et al., 1991, 1996; Moretzsohn et al., 2013). Molecular evidence indicates that *Arachis duranensis* and *Arachis ipaensis* are the two most likely progenitors that donated the A and B subgenomes, respectively (Kochert et al., 1996; Ramos et al., 2006; Grabiele et al., 2012; Moretzsohn et al., 2013). The genome sizes of the two species are ∼1.25 and ∼1.56 Gb, respectively (Samoluk et al., 2015), and their sum is close to the total genome size of *A. hypogaea* (∼2.8 Gb) (Temsch and Greilhuber, 2000), indicating that no large changes that affected genome size have taken place since polyploidy. Moreover, researches indicated that the genomes of *A. duranensis* and *A. ipaensis* are similar to cultivated peanut's A and B subgenomes (Kochert et al., 1996; Seijo et al., 2007; Robledo et al., 2009; Robledo and Seijo, 2010; Moretzsohn et al., 2013). The high-DNA identity between the *A. ipaensis* genome and the B subgenome of cultivated peanut, along with biogeographic evidence, indicates that *A. ipaensis* may be the direct descendant of *A. hypogaea* that contributed the B subgenome (Bertioli et al., 2016).

The large genome size of *A. hypogaea* (∼2.8 Gb) and highly repetitive content (64%) makes the assembly of the peanut genome sequence very challenging (Dhillon et al., 1980; Temsch and Greilhuber, 2000; Bertioli et al., 2016). Therefore, sequencing and analyzing the genomes of the two diploid ancestors to uncover the genome of cultivated peanut was considered a sensible initial strategy. Our previous sequencing of the peanut A-genome progenitor, *A. duranensis*, provided new insights into *Arachis* biology, evolution and genomic changes (Chen et al., 2016). To gain insights into the genomic evolution, as well as the divergence, of the peanut B subgenome and to provide candidate genes to enable a better understanding of the biology of leguminous species, we sequenced the suspected peanut B-genome progenitor, *A. ipaensis*, and re-sequenced two A-genome and three B-genome genotypes (Chen et al., 2016). The *A. ipaensis* genome sequencing will facilitate future research on the genome assembly of cultivated peanut and, has the potential to accelerate the molecular breeding of peanut varieties.

RESULTS AND DISCUSSION

Genome Sequencing, Assembly, and Annotation

The genome of the peanut B-genome progenitor, *A. ipaensis* (ICG_8206), was sequenced using a shotgun approach on the Illumina HiSeq2500 platform (**Supplementary File 1**: Figure S1). We generated 250.40 Gb of high-quality reads, representing 149.53 × genome coverage, with fragment lengths ranging from 250 to 20 Kb (**Supplementary File 1**: Table S1). A total of ∼1,391.70 Mb of the *A. ipaensis* genome sequence was assembled using SOAPdenovo2 (Luo et al., 2012) with a contig N50 of 8,067 bp and a scaffold N50 of 170,050 bp (**Table 1**; **Supplementary File 1**: Tables S2, S3). An assessment of the draft genome assembly using the core eukaryotic gene mapping approach method (Parra et al., 2007) revealed that >98% of conserved genes were present in the assembly (**Supplementary File 1**: Table S4). Over 98% of transcript sequences (>500 bp) were mapped to the assembled genome (**Supplementary File 1**: Table S5). Based on *k*-mer statistics, the *A. ipaensis* genome is estimated to be ∼1,475.83 Mb, which is consistent with the total scaffold length (**Supplementary File 1**: Table S6 and Figure S2). The average GC content is 36.70% (**Table 1**; **Supplementary File 1**: Figure S3), which is equivalent to that of the *A. duranensis* genome (Chen et al., 2016), and its distribution is highly similar to previously reported *Arachis* genomes (Bertioli et al., 2016; Chen et al., 2016) but different from those of *Glycine max*, *Arabidopsis thaliana,* and *Oryza sativa* (**Supplementary File 1**: Figure S4).

We predicted 39,704 genes with average transcript and coding sequence lengths of 3,741 and 1,246 bp, respectively (**Table 1**). The whole-genome's gene density is one gene per 35.05 Kb (**Figure 1** and **Table 1**), and the mean exon and intron lengths per gene are 250 and 625 bp (**Table 1**), respectively, which were relatively longer than those in other leguminous species, such as *Cicerarietinum* (Varshney et al., 2013) and *G. max* (Schmutz et al., 2010). Compared with the gene sets of legumes, oilseeds, and other plant species (**Supplementary File 1**: Table S7), the distribution of the *A. ipaensis* gene features is most similar to those of *A. duranensis* and legumes, such as *C. arietinum* and *G. max*, but different from those of non-leguminous species, such as *A. thaliana* and *O. sativa* (**Supplementary File 1**: Table S8 and Figure S5). Moreover, the *A. ipaensis* gene number is comparable to those of *Lotus japonicus* (39,366) and *Zea mays* (39,498), greater than that of *C. arietinum* (24,819), and substantially lower than those of *G. max* (54,174) and *Medicago truncatula* (50,444) (**Supplementary File 1**: Table S9). Functions were tentatively assigned to 39,645 genes but not to 59 genes that may be peanut-specific (**Table 1**). Most of the *A. ipaensis* genes have homologous gene models in the TrEMBL (99.82%) and Interpro (71.29%) databases (Bairoch and Apweiler, 2000; Zdobnov and Apweiler, 2001), and ∼99.85% of the gene models matched entries in publically available databases (**Supplementary File 1**: Table S10). Conservative analyses indicated that the predicted proteins of *A. ipaensis* were most similar to those of *A. duranensis* (88.10%), followed by *Cajanus cajan* (67.4%), and least similar to those of gramineous crops, such as *Sorghum italica* (33.53%) and *S. bicolor* (34.51%) (**Supplementary File 1**: Table S8).

A total of 2,530 putative *A. ipaensis* transcription factor (TF) genes were identified in 58 families, which was equal to or slightly higher than of the numbers found in *O. sativa* and *A. thaliana*, much higher than in *L. japonicus* but lower than in *G. max* and

TABLE 1 | Genome assembly and annotation of the *A. ipaensis*.

Genome features	Measures
ASSEMBLY FEATURES	
Number of scaffolds	79,408
Total span	1,391,700,926 bp (∼1.39 G)
N50 (scaffolds)	170,050 bp
Longest scaffold	1,172,168 bp
Number of contigs	1,008,989
N50 (contigs)	8,067 bp
Longest contig	81,804 bp
GC content	36.70%
GENE MODELS	
Number of gene models	39,704
Mean transcript length	3,741 bp
Mean coding sequence length	1,246 bp
Mean number of exons per gene	4.99
Mean exon length	250 bp
Mean intron length	625 bp
Mean gene density	35.05 Kb
Number of genes annotated	39,645
Number of genes unannotated	59
NON-PROTEIN CODING GENES/ELEMENTS	
Number of pre-miRNA genes	71
Mean length of pre-miRNA genes	123 bp
Pre-miRNA genes share in genome	0.000590%
Number of pre-rRNA fragments	313
Mean length of pre-rRNA fragments	186 bp
Pre-rRNA fragments share in genome	0.003928%
Number of pre-tRNA genes	2,914
Mean length of pre-tRNA genes	75 bp
Pre-tRNA genes share in genome	0.014836%
Number of pre-snRNA genes	152
Mean length of pre-snRNA genes	111 bp
Pre-snRNA genes share in genome	0.001139%
Total transposable elements, bp (TEs)	1,125,924,736
Transposable element percent in genome	75.97%

Glycine soja (**Supplementary File 1**: Figure S6). The distribution of the *A. ipaensis* TF genes among the families was highly similar to those of *A. duranensis* and *G. max* (**Supplementary File 1**: Figure S7). FAR1 was dominant in *A. ipaensis* (**Figure 2A**), as in the A-genome progenitor, *A. duranensis* (Chen et al., 2016). More importantly, the FAR1 TF families play pivotal roles in modulating phyA-signaling homeostasis (Lin et al., 2007), and phyA, together with phyB, regulate skotomorphogenesis and photomorphogenesis in higher plants (Medzihradszky et al., 2013). The FAR1 TF families identified in *A. thaliana* contained several conservative motifs (**Supplementary File 1**: Figure S8), and *phyA* and *phyB* were highly expressed in different tissues (shoot, seed, leaf, flower, and root) at different growth stages in *A. thaliana* (**Supplementary File 1**: Figure S9). In addition, previous non-synonymous substitutions per non-synonymous site (*Ka*)/synonymous substitutions per synonymous site (*Ks*) analyses of *phyB* in *A. duranensis* and *G. max* showed evidence of positive selection (Chen et al., 2016). These findings may enhance our understanding of peanut's unique fructification, having aerial flowers but subterranean fruit, as well as providing evidence for different regulators of biological functions in *Arachis* and other plants.

We identified 71 *Arachis* pre-microRNAs (pre-miRNAs) (**Supplementary File 2**: **Data S1**) with an average length of 123 bp, 2,914 pre-transfer RNAs (pre-tRNAs) with an average length of 75 bp, 313 pre-ribosomal RNAs (pre-rRNAs) with an average length of 186 bp including 5S (108), 5.8S (55), 18S (82), and 28S (68), and 152 pre-small nuclear RNAs (pre-snRNAs) with an average length of 111 bp. These genes represent 0.000590, 0.014836, 0.003928, and 0.001139% of the *A. ipaensis* genome, respectively (**Table 1**; **Supplementary File 1**: Table S11).

Approximately 75.97% of the *A. ipaensis* genome is composed of transposable elements (**Figure 1**; **Tables 1**, **2**), which was higher than other legumes, such as *G. max* (59.00%) (Schmutz et al., 2010), *C. cajan* (51.60%) (Varshney et al., 2011) and *M. truncatula* (30.50%) (Young et al., 2011). Long-terminal repeat (LTR) retrotransposons are the dominant transposable elements, covering 64.15% of the nuclear genome (**Table 2**). Sequence divergence analyses indicated that most of *A. ipaensis* transposable elements had a ∼30% divergence rate (**Supplementary File 1**: Figure S10).

The *A. ipaensis* genome contains 188,075 simple sequence repeats (SSRs), for which 80,218 SSR primers were designed (**Supplementary File 1**: Table S12; **Supplementary File 3**: **Data S2**). Of these SSRs, the di-nucleotide repeats are the most abundant, accounting for 48.38% of the total SSRs, followed by tri- nucleotide repeats (28.06%) (**Supplementary File 1**: Table S12). Among the di-nucleotide type, the AT/AT motif type had the greatest frequency (∼21.9%). Among the tri-nucleotide type, the AAT/ATT is dominant (∼11.4%) (**Supplementary File 1**: Figure S11). Using two A-genome genotypes (ICG_8123 and ICG_8138) and three B-genome genotypes (ICG_8960, ICG_8209, and ICG_13160) that were re-sequenced in our earlier study (Chen et al., 2016), we identified 26,050,150 variations, including 24,688,277 single nucleotide polymorphisms (SNPs) and 1,361,873 insertion-deletions (InDels) (**Supplementary File 1**: Table S13 and Figure S12). Among these variations, ∼4 million SNPs were present in the two diploid A species (ICG_8123 and ICG_8138). By contrast, ∼5 million SNPs were identified in the comparison of the three diploid B species (ICG_8960, ICG_8209, and ICG_13160) (**Supplementary File 1**: Table S13 and Figure S12). Thus, the diploid B species *Arachis magna* and *Arachis batizocoi* may have more abundant genetic diversity than the diploid A species *A. duranensis* when compared with the reference *A. ipaensis* (ICG_8206) genome assembly. The geographical origin of *Arachis* indicated that the distribution of *A. duranensis* is more extensive and also closer to that of *A. ipaensis* which has only one known location of origin, than *A. magna* (Bertioli et al., 2016). Another source of confusion among the variations may result from the two A-genome genotypes having fewer mapped reads than the three B-genome genotypes.

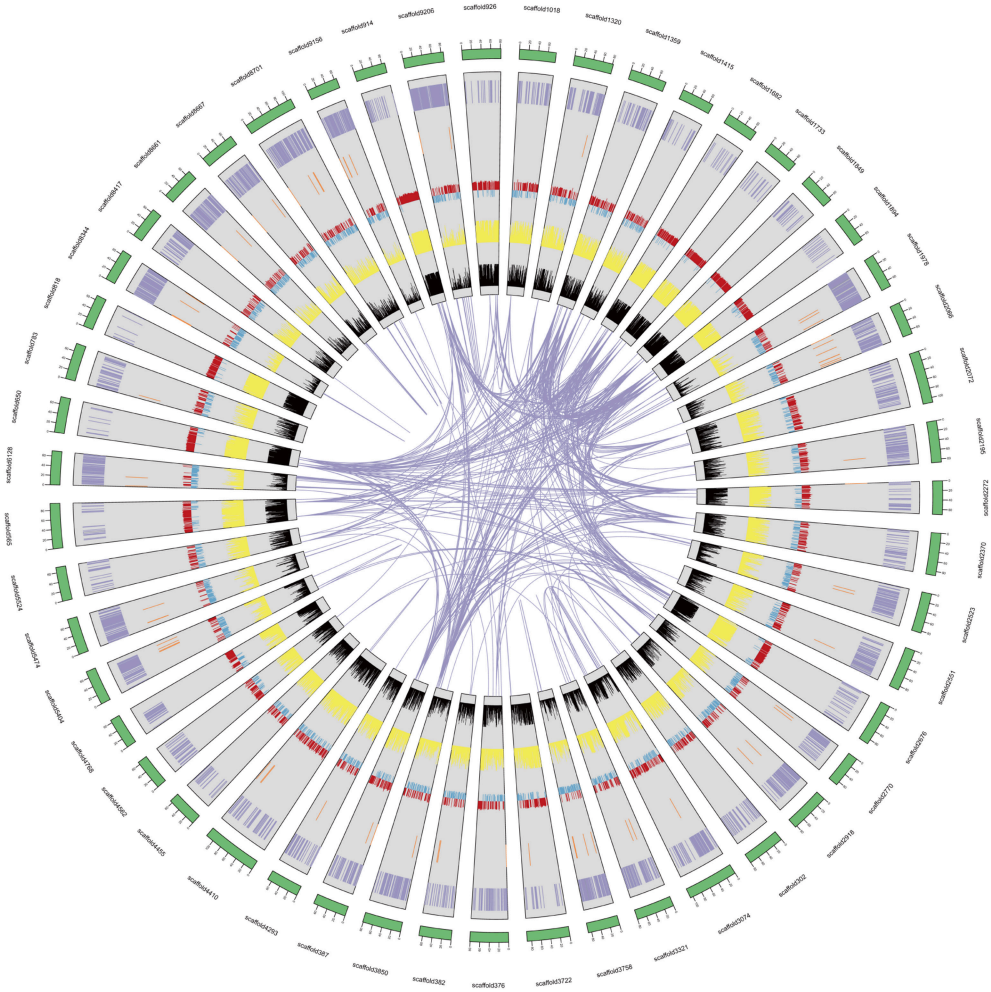

FIGURE 1 | *A. ipaensis* genome overview. From the outer edge inward, circles represent the 50 largest DNA sequence scaffolds (green), the genes on each scaffold (purple), the non-coding RNA on each scaffolds (brown), GC content (red and blue), repeat density at 10 Kb (yellow), and transposable element density at 10 Kb (black).

Gene Family and Phylogenetic Analysis

A total of 16,791 orthologous gene groups were identified among 18 species using OrthoMCL (Li et al., 2003; **Supplementary File 4**: **Data S3**), including 959 *A. ipaensis*-specific families containing 6,443 genes (**Supplementary File 1**: Table S14). A gene ontology (GO) annotation suggested differentially enriched functional categories in the peanut-specific families, indicating that these gene families may be closely related to the unique *Arachis* growth characteristics, such as aerial flowers but subterranean fruit, and lipid biosynthesis (**Supplementary File 1**: Figures S13–S15). Moreover, 1,624 of these orthologous groups were single-copy orthologs (**Supplementary File 1**: Table S15 and Figure S16). In addition, 6,443 unique paralogs of *A. ipaensis* genes occurred in species-specific homolog groups, indicating that these groups could arise from genomic structural rearrangements that obscured simple orthology, such as nonallelic recombination or gene conversion, followed by duplication (**Supplementary File 1**: Table S15 and Figure S16; Varshney et al., 2013). We identified 12,017 orthologous groups common to all five leguminous species (**Figure 2B**), 11,985 groups between *A. ipaensis* and *Ricinus communis* (oilseed crop) (**Supplementary File 1**: Figure S17), 9,099 groups between *A. ipaensis* and Gramineae/Poaceae crops (**Supplementary File 1**: Figure S18), and 10,501 orthologous groups are common to *A. ipaensis* and other distantly related plant species (**Supplementary File 1**: Figure S19). These results provide an important molecular foundation for comparative biology and for functional mechanistic inferences in *A. ipaensis*, as well as other species, because simple orthologous family genes often exhibit conserved molecular functions that were maintained during evolution process.

A polygenetic tree based on single-copy orthologous genes showed *A. ipaensis* and *A. duranensis* in the same clade, which did not include any other leguminous species, indicating their closer genetic distance and divergence time (**Figure 2C**). Furthermore, a special phylogenetic tree estimated that the

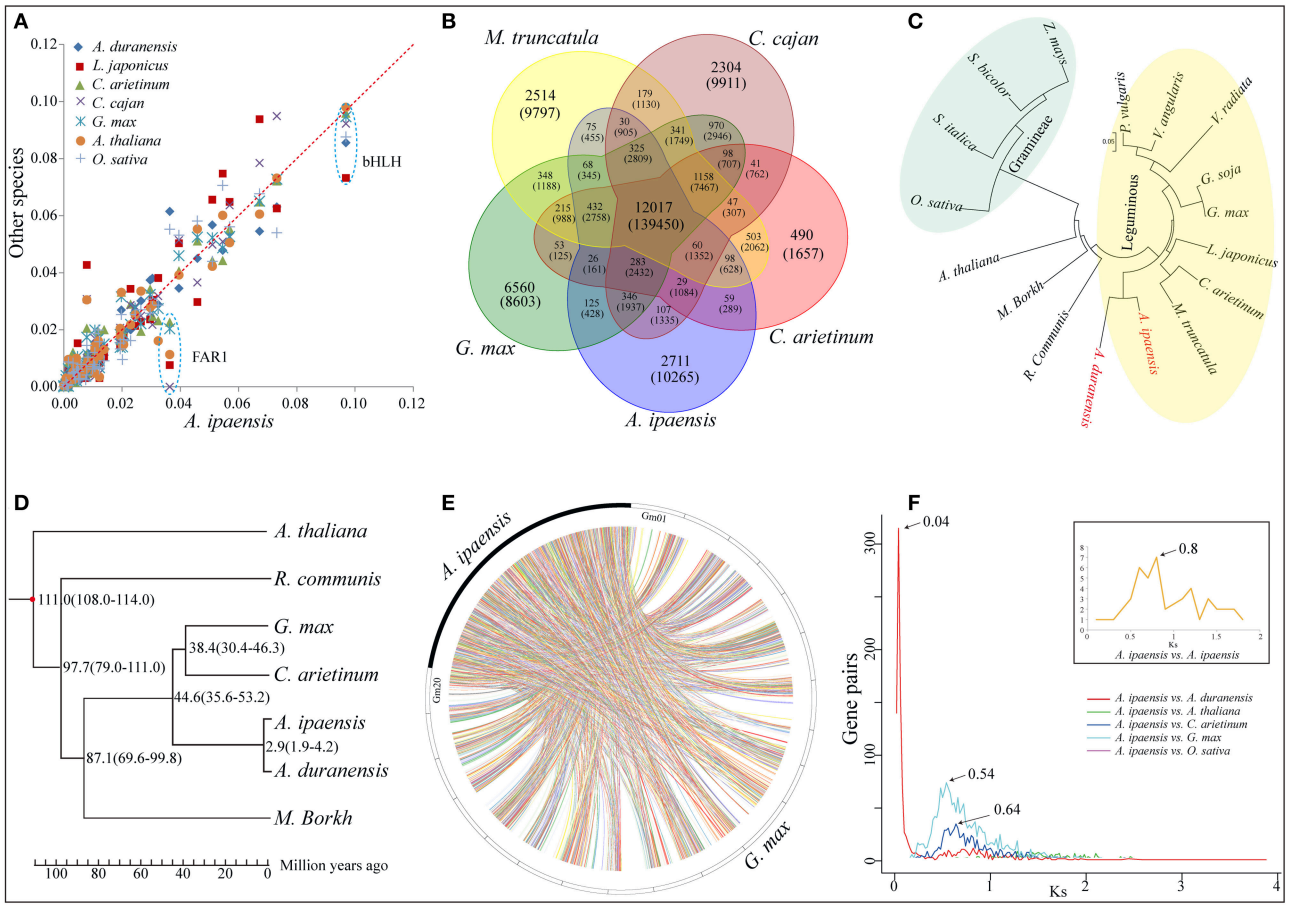

FIGURE 2 | Comparative genomic and evolutionary analysis. **(A)** Scatter plot of percentage of *A. ipaensis* transcription factors in relation to *L. japonicas*, *C. arietinum*, *C. cajan*, *G. max*, *A. thaliana* and *O. sativa*. **(B)** Venn diagram showing distribution of gene families among *A. ipaensis*, *G. max*, *M. truncatula*, *C. cajan* and *C. arietinum*. **(C)** Cluster tree for 17 plant species including common leguminous and gramineous crops based on single copy orthologous genes. **(D)** Phylogenetic tree for 7 representative plant species. The numerical on each node represents the estimated differentiation time using the evolutionary time between *A. thaliana* and *G. max* (~108–114 Mya) as a correction. **(E)** Syntenic relationship between *A. ipaensis* scaffolds and *G. max* chromosomes. **(F)** Synonymous substitution rate (*Ks*) dating of duplication blocks in *A. ipaensis* and different combinations of orthologs of *A. duranensis*, *A. thaliana*, *C. arietinum*, *G. max,* and *O. sativa*. Different colored lines represent the distribution of *Ks* against orthologs gene pairs among different plant species. Inset shows the distribution of *Ks* between the gene pairs present in the duplicated blocks within the *A. ipaensis* genome.

divergence of the two species occurred ~2.9 million years ago (Mya) (**Figure 2D**), which was fairly consistent with a previous report (~2.16 Mya) (Bertioli et al., 2016). Syntenic blocks identified between *A. ipaensis* and other species was found to be extensively conserved (**Supplementary File 1**: Table S16). The largest number of syntenic blocks was identified between *A. ipaensis* and *G. max* (**Figure 2E**). The longest syntenic block (>10 Kb) was observed between *A. ipaensis* and *A. duranensis* (**Supplementary File 1**: Table S16). The numbers of syntenic blocks identified within the respective *A. ipaensis* and *A. duranensis* genomes were extremely lower than the number between the two genomes (**Supplementary File 1**: Figure S20) as well as the number between the *A. ipaensis* and *G. max* genomes (**Supplementary File 1**: Figure S21; Bertioli et al., 2016), indicating that few large-scale genome duplication events occurred in the *A. ipaensis* genome's evolution or that syntenic blocks were lost after genome duplication events.

The *Ks* values between paralogous or orthologous genes reveals a mechanism of molecular evolution (Lna, 1996). Distributions of *Ks* distances between paralogs within *A. ipaensis* and orthologs among *A. ipaensis*, leguminous crops and other species were plotted (**Figure 2F**). The *A. ipaensis* paralogs showed a peak at ~ 0.80, which is similar to those of *M. truncatula* (~0.80) and *L. japonicus* (~0.73) (Cannon et al., 2006) but lower than those of *A. duranensis* (~0.9) and *A. ipaensis* (~0.95) (Chen et al., 2016). Thus, the whole-genome duplication events of *A. duranensis* and *A. ipaensis* occurred around the time that corresponds to a *Ks* value range of 0.8–0.95. In addition, *A. duranensis* and *A. ipaensis* orthologs showed a prominent peak at ~0.04, which is consistent with a previous study (Bertioli et al., 2016). Assuming a synonymous substitution rate per synonymous site of 6.1×10^{-9} per year for eudicots (Lynch and Conery, 2000), the two species were estimated to have diverged ~3.28 Mya, which is close to the estimation based on the

TABLE 2 | Organization of repetitive sequences in *A. ipaensis* genome.

Repetitive elements	Repeat number	Length (bp)	In total repeat (%)	In genome (%)
Total retrotransposons	2,444,183	9,88,193,900	87.77	66.68
LINE retrotransposons	163,947	43,942,874	3.9	2.97
SINE retrotransposons	2,859	726,676	0.06	0.05
LTR retrotransposons	2,277,377	950,690,158	84.44	64.15
Gypsy	1,727,232	796,763,491	70.77	53.76
Copia	343,066	91,500,532	8.13	6.17
LTR	23,529	1,543,961	0.14	0.10
Other	183,550	98,476,493	8.75	6.64
Other retrotransposons	668	47,680	0	0.00
Total DNA transposons	364,250	98,441,246	8.74	6.64
Total unclassified elements	311,209	84,709,729	7.52	5.72
Total transposable elements	3,120,310	1,125,924,736	–	75.97
Redundant		1,171,344,875		
Nonredundant		1,125,924,736		

phylogenetic tree (**Figure 2D**). Furthermore, *Ks* dating suggested the divergence of *A. ipaensis* and *G. max* ($Ks = \sim 0.54$) at 44.3 Mya and that of *A. ipaensis* and *C. arietinum* ($Ks = \sim 0.64$) at 52.5 Mya.

The graphic trend of the *Ka*/*Ks* (ω) and *Ks* between the orthologs of *A. duranensis* and *A. ipaensis* formed three clusters, such as $Ks = 0$–0.3, 0.5–1.5, and >1.5, and the ω values decreased as the *Ks* values increased (**Supplementary File 1**: Figure S22). The genes with $Ks \geq 1.5$ are attributed to pan-eudicot palaeoploidization, and the genes with lower ω ratios are considered to be under neutral selection. Here, the 45 *A. ipaensis* genes with ω ratios > 1 may be under positive selection pressure (**Supplementary File 1**: Figure S23).

Peanut is an allotetraploid species that may have originated from a single resent hybridization event between two diploid species, followed by polyploidization. Cytogenetic, phylogeographic and molecular evidence indicates that *A. duranensis* and *A. ipaensis* are the most likely donors of the A and B subgenomes, respectively (Kochert et al., 1996; Seijo et al., 2007; Robledo et al., 2009; Robledo and Seijo, 2010; Moretzsohn et al., 2013). A previous study estimated the divergence of the two species at ∼2.88 Mya (Moretzsohn et al., 2013). The estimation using a comparative genomics analyses between them was ∼2.9 Mya, which was fairly consistent with our report. Moreover, sequence comparisons with tetraploid cultivated peanut estimated the divergence times of *A. duranensis* and *A. ipaensis* from the A and B subgenomes of *A. hypogaea* as ∼247,000 and ∼9,400 years, respectively (Bertioli et al., 2016).

Comparative genomics analyses of chromosomal structure and synteny between *A. duranensis* and *A. ipaensis* indicated that some chromosomes shared a conservative linear structure that was partially in accordance with our other analyses (**Supplementary File 1**: Figure S20). Other analyses showed a large inversion in one or both arms of a chromosome (Bertioli et al., 2016). In contrast, chromosomes 07 and 08 have undergone complex rearrangements that were consistent with cytogenetic observations (Seijo et al., 2007; Nielen et al., 2010). Importantly, a genomic comparison showed a fundamentally one-to-one correspondence between the diploid chromosomes and cultivated peanut linkage groups. However, the *A. duranensis* chromosomes were less similar to *A. hypogaea* sequences than those of *A. ipaensis* (Bertioli et al., 2016). These results may help to uncover potential mechanisms of hybridization events in the future.

Disease Resistances and Nucleotide-Binding Site (NBS)-Leucine-Rich Repeat (LRR) Encoding Genes

Plant NBS-LRR proteins encoded by resistance genes (*R* genes) play key roles in the responses of plants to various pathogens. The *R* genes can be classified into various subfamilies based on the present of different domain, such as CC-NB-LRR, TIR-NB-LRR, ser/thr-LRR, Kin-LRR, and others (e.g., *Mol* and *Asc-1*; Sanseverino et al., 2010). The *A. ipaensis* genomic assembly contains 1,437 putative disease *R* genes as assessed by a screening of the PRG database (**Supplementary File 1**: Table S17; **Supplementary File 5**: **Data S4**; Sanseverino et al., 2010). Compared with other legumes, the *A. ipaensis* genome possesses more *R* genes than the *G. max* and *M. truncatula* genomes but less than the *A. duranensis* and *C. cajan* genomes. Moreover, these *R* genes tend to cluster on different scaffolds. For example, several large clusters containing 6–10 *R* genes are located on six different scaffolds (**Supplementary File 1**: Figure S24). The NL subfamily of genes, which confers resistance against pests and diseases, is the second largest *R* gene-containing family, and these genes can be clustered reasonably into different individual clades in *A. ipaensis*, *A. duranensis,* and *A. thaliana*, indicating that gene divergence occurred during the evolution of the three species (**Supplementary File 1**: Figure S25). In addition, we analyzed protein motifs in the most homology of the top 20 *R* genes found in PRG database using MEME (Bailey et al., 2009), and the results showed highly conserved LRR

motifs (**Supplementary File 1**: Figure S26). However, further investigation is required to determine the biological functions of these *R* genes.

Identification of Genes Related to Biological Nitrogen Fixation

Nitrogen is one of the most important plants require nutrients, and in agriculture nitrogen availability influences both crop yield and quality. Leguminous plants, such as peanut, soybean, and *Medicago*, can transform molecular nitrogen into available ammonia nitrogen through the leguminous-root-nodule bacteria nitrogen-fixing system that results from the symbiotic interactions between leguminous plants and rhizobia (**Figure 3A**). In the *A. ipaensis* and *A. duranensis* genomic assemblies, 16 and 38 root-nodule developmental genes respectively, have been identified (**Supplementary File 1**: Table S18; **Supplementary File 6**: **Data S5**). As expected, there are greater numbers of nodulation-related genes present in leguminous plants than in non-leguminous plants, such as *A. thaliana*, *O. sativa,* and *Z. mays* (**Supplementary File 1**: Figure S27).

Nitrogen-fixing root nodules are important symbiotic organs that provide an epiphytic site for rhizobia to convert atmospheric nitrogen to ammonia, and supply its host plant with fixed nitrogen. In return, the rhizobia gain photosynthates from the plant (**Figure 3A**). In leguminous plants, multiple genes are involved in the formation and development of root nodules, as well as in the autoregulation of the nodulation (AON) process, which is a systemic feedback loop used to avoid an excessive energy expenditure from "unwanted" nodulation (**Figure 3A**; **Supplementary File 7**: **Data S6**). Here, four homologous LRR receptor kinase genes were identified in *A. ipaensis* (XP_004512550.1-D2 and XP_007158329.1-D2) and *A. duranensis* (XP_015956675.1 and XP_015963325.1) (**Figure 3B**; **Supplementary File 1**: Figure S28; **Supplementary File 7**: **Data S6**). A phylogenetic tree showed that the four homologous genes were clustered into an independent clade, together with other LRR receptor kinase genes (**Figure 3C**). Interestingly, these four genes contain multiple common motifs, including a conserved LRR motif, indicating a similar biological function (**Figure 3D**). The GO analyses suggested that the four homologous genes are involved in ion binding and signal transducer activity (**Supplementary File 1**: Figures S29–S32). More importantly, the proteins encoded by the four genes showed similar three-dimensional structures and localized on the cell membrane (**Supplementary File 1**: Figures S29–S32).

We also identified two other nodule development-related genes (XP_015934647.1 and XP_015939255.1) that are homologous to the TF genes of the GRAS family in *A. duranensis*. One gene is homologous with *MtNSP2* and *PsSYM7* (Kaló et al., 2005), while the other is an ortholog of *MtNSP1* (Imaizumi-Anraku et al., 2005) (**Figures 3A,B**; **Supplementary File 1**: Figure S33; **Supplementary File 7**: **Data S6**). The phylogenetic tree indicated that the two homologs were classified into the TF category but appeared in different branches (**Figure 3C**). In addition, the GO enrichment indicated that the two genes participate in the regulation of multiple biological processes, such as nucleic acid-binding TF and signal transducer activities (**Supplementary File 1**: Figures S34, S35). The three-dimensional structures of the two proteins were completely dissimilarity, and the two proteins localized in the nucleus (**Supplementary File 1**: Figures S34, S35). These results could provide candidate genes and basic bioinformation for further functional studies of nodule formation in leguminous crops.

Genetic Mechanism of Drought Adaptation

Peanut (*A. hypogaea* L.) is a typical upland crop in tropical, subtropical, and warm temperate climates. Drought adaptation plays a central role in their growth and development. During drought stress, TFs, such as MYB, MYB-related, NAC, WRKY, bZIP, and ERF, are involved in numerous physiological responses (Shinozaki and Yamaguchi-Shinozaki, 2000) (**Supplementary File 1**: Figure S36). Here, the total number of TF genes identified in upland crops was greater than that found in hygrophilous plants (**Supplementary File 1**: Table S19 and Figure S37). Notably, in *A. ipaensis* we identified 185 MYB and 129 MYB-related TFs (**Supplementary File 1**: Table S19), most of which contain a highly conserved DNA-binding domain, and they are key factors in regulatory networks controlling development, metabolism and responses to biotic and abiotic stresses (Dubos et al., 2010). The second large number of drought tolerance-related TFs, with 170 members, is the ERF family (**Supplementary File 1**: Table S19). The ERF proteins, sharing a conserved 58–59 amino-acid domain, are key regulators linked to responses to plant stresses, such as cold, drought and pathogen attack (**Supplementary File 1**: Figure S38; Singh et al., 2002). In *A. duranensis*, *A. ipaensis* and *A. hypogaea* species, sets of 51, 57 and 53 ERF TF family proteins, respectively, were obtained from the Plant Transcription Factor Database (Jin et al., 2015, 2017; **Supplementary File 1**: Figures S39–S41). These TF proteins contained different DNA-binding domains and can be categorized into different branches based on different motif permutation structures, indicating the distinct functional and evolutionary features of ERF TFs in different *Arachis* species (**Supplementary File 1**: Figures S39–S41).

Heat-shock proteins (Hsps)/chaperones are important defense mechanism members against abiotic stresses, such as drought, salinity and extreme temperatures (Wang et al., 2004; **Supplementary File 1**: Figure S42). Drought stress is a common factors that induces Hsp expression (Kimpel et al., 1990; Sun et al., 2002). To elucidate the cause of drought tolerance, five major families of Hsps/chaperones were identified in upland crops and hygrophilous plants (**Supplementary File 1**: Table S20 and Figure S43). As expected, the total number of Hsps/chaperones obtained in upland crops was much great than in hygrophilous plants (**Supplementary File 1**: Figure S43).In particular, *A. ipaensis* and *G. soja* had 118 and 34 Hsp70 subfamilies, respectively, compared with only 1 in rice (**Supplementary File 1**: Table S20). The great number of Hsps/chaperones detected in *A. ipaensis* and *G. soja* indicates the nature of drought adaptation in upland crops.

The subtilisin-like protease (SDD1) gene family is involved in the regulation of stomatal density and distribution to adjust for

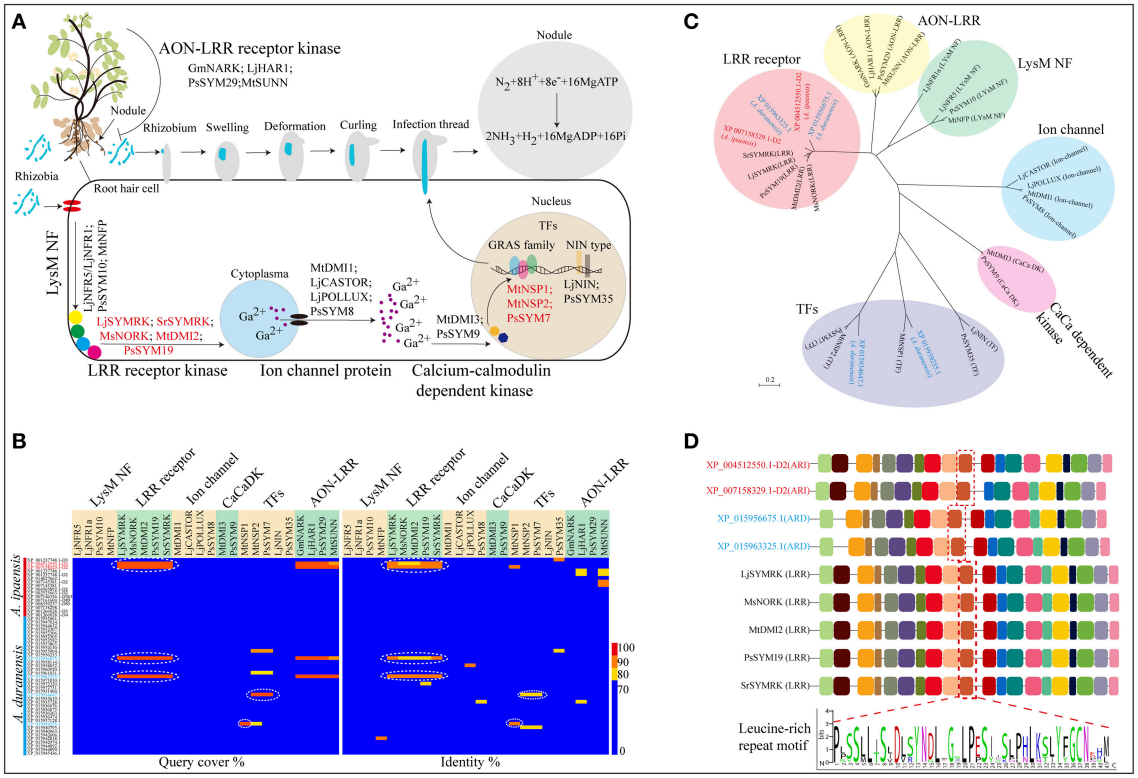

FIGURE 3 | Biological nitrogen fixation in leguminous plants. **(A)** Genes involved in nodule initiation, development and signal recognition pathway. **(B)** Protein sequence alignment of Nod related genes identified in *A. ipaensis* and *A. duranensis*. **(C)** Phylogenetic tree of nodule development genes and their homologs from *A. ipaensis* and *A. duranensis*. **(D)** Identification of high conserved domains of leucine-rich repeat (LRR) receptor kinases. Red dashed boxes represent LRR conserved motif. **(A)** The rhizobium (blue) attach to the surface of root hair cell. After swelling, deformation, curling and infection thread, the bacteria are released into cells via endocytosis then a vacuole-like structures (symbiosomes), in which the bacteria convert N_2 to NH_3, formed. But how is the Nod signal transmitted? Initially, the rhizobia-derived signal is perceived by LysM-type protein receptor kinases, such as NRF1 and 5 (Radutoiu et al., 2003) and SYM10 (Schneider et al., 2002) identified in *L. japonicus* and *P. sativuml*, followed by a downstream leucine-rich receptor kinase, for example SYMRK (Stracke et al., 2002 and Capoen et al., 2005), NORK (Endre et al., 2002), DMI2 (Catoira et al., 2000), and SYM19 (Stracke et al., 2004) from *L. japonicus*, *Sesbania rostrata, M. sativa*, *M. truncatula,* and *P. sativuml*, respectively. Then, the Nod factor (NF) signal is processed through a signal transduction cascade involving proteins including ion channels [MDI1(Ané et al., 2004), CASTOR (Imaizumi-Anraku et al., 2005), POLLUX (Imaizumi-Anraku et al., 2005), and SYM8 (Edwards et al., 2007)], calcium-calmodulin-dependent kinase (CaCaDK) (MDI3 and SYM9) (Lévy et al., 2004) and transcription factors [NSP1 (Smit et al., 2005), NSP2 (Kaló et al., 2005), SYM7 (Kaló et al., 2005), NIN (Schauser et al., 1999), and SYM35 (Borisov et al., 2003)]. Finally, rhizobia infection occurred primarily through uncharacterized target genes that may be activated by these TFs.

drought stress by modulating the apertures of these pores flanked by two guard cells (Berger and Altmann, 2000). In the expanded gene families, 39 and 40 SDD1 genes were identified in *A. ipaensis* and *A. duranensis*, respectively (**Supplementary File 8**: **Data S7**). These gene families were divided into different clusters according to their related functions but showed a pattern of cross-distribution in each cluster based on their different genetic relationships (**Supplementary File 1**: Figure S44).

Oil Synthesis

Because peanut is an important oilseed crop, 1,613 *A. ipaensis* genes related to the biosynthesis of fatty acids and triacylglycerols were identified, which was more than were identified in the nonoilseed plant *Arabidopsis* (1,380) and rice (1,419) (**Supplementary File 1**: Table S21). In addition, fatty acids and triacylglycerols synthesis involves many key enzymes, such as ACCase (Slabas and Fawcett, 1992), acyl-ACP thioesterase (A and B) (Dörmann et al., 2000; Bonaventure et al., 2003; Serrano-Vega et al., 2005), LACS (Zhao et al., 2010), DGAT (Yen et al., 2008), and FAD (Pham et al., 2012) (**Supplementary File 1**: Figure S45). When we manually investigated the homologous genes in the storage lipid biosynthesis pathway using the Arabidopsis Lipid Gene Database (Mekhedov) (http://lipids.plantbiology.msu.edu/), 116 nonredundant homologs potentially involve in lipid biosynthesis were obtained in *A. ipaensis* (**Supplementary File 9**: **Data S8**). Consistent with the lipids produced in peanut seeds, one, and nine homologous genes encoding acyl-ACP thioesterase A and B (*FATA* and *FATB*), respectively, the two key enzymes leading to the synthesis of fatty acid, were identified. Moreover, multiple copies or isoforms of some key genes, such as *FAD2*, *LACS,* and *KAS*, involved in triacylglycerol synthesis were also detected in the *A. ipaensis* genome (**Supplementary File 9**: **Data S8**).

FAD2, encoding δ-12 oleic acid desaturase, is the essential gene that controls linoleic acid biosynthesis (López et al., 2000). It converts oleic acid to linoleic acid by desaturating the δ-12 carbon and determines the multi-polyunsaturated fatty acid content and proportion in most oil seed plants (**Figure 4A**). In this study, three new *FAD2* homologous genes (XP_004497897.1-D3, XP_007162321.1, and XP_007162321.1-D2) were identified in *A. ipaensis* (**Supplementary File 1**: Figure S46). The proteins of *FAD2* and its homologs contain the highly conserved feature of three enzyme-specific histidine boxes (**Figure 4B**), which are considered to be essential for desaturase activity because they act as potential ligands for iron atoms (Sakai and Kajiwara, 2005). A phylogenetic tree showed that *FAD2* clustered into five groups based on its genus, and the three homologous genes were more closely related to the evolutionary kinship of oil seed plants, especially *A. hypogaea* (**Figure 4C**). This result indicated that *FAD2* is an extremely conserved gene in the fatty acid biosynthesis pathway. In addition, the GO terms revealed that the three homologous genes having δ-12 oleic acid dehydrogenase activities (ω-6 fatty acid desaturase activities) were involved in the fatty acid biosynthesis process and that the proteins encoded by the three genes were subcomponents of the endoplasmic reticulum membrane. They had similarity three-dimensional structures, which was supported by the predicted protein subcellular localization (**Supplementary File 1**: Figures S47–S49).

Pairwise comparisons of the amino acid sequences of XP_007162321.1-D2 from *A. ipaensis* with FAD2 from *A. hypogaea* revealed 100% sequence identities with no gaps (**Supplementary File 1**: Figure S50), which confirmed the ancestral origin of *FAD2* as being the *A. ipaensis* genome. A signal peptide analysis showed a low level S-score, indicating a typical non-secretory protein with no leading peptide (**Figure 4D**). This was supported by the predicted protein subcellular localization (**Supplementary File 1**: Figure S47E). Moreover, four transmembrane domains were predicted in their amino acid sequence (**Figure 4E**). Importantly, the protein hydrophobicity/hydrophilicity prediction revealed four strong hydrophobic regions, which completely overlapped with the transmembrane regions (**Figure 4F**). These results provide information for exploring the origin of *FAD2*, and the homologous gene will be of service to peanut improvement for high oleic acid.

Among the key enzyme-encoding genes, 82 nonredundant homologous genes had high distributions of non-synonymous substitutions ($Ka/Ks > 1.0$) between *A. ipaensis* and *A. thaliana* as assessed by the branch-site likelihood ratio test, indicating positive selection during domestication (**Supplementary File 1**: Figure S51; **Supplementary File 10**: **Data S9**). Coincidentally, 21 fatty acid biosynthesis genes located in multiple improvement-selective sweeps regions were obtained through combined genome selective sweeps and GWAS analyses in soybean (Zhou et al., 2015). Thus, we hypothesize that these 82 genes, including *FAD2* (2), *KASIII* (2), and *FATB* (6) homologs with high *Ka*/*Ks* values (**Supplementary File 1**: Figure S52) may also have undergone domestication.

TFs that regulate seed development play crucial roles in seed lipid accumulation. To date, the TFs regulating lipid metabolism mainly belong to the following 6 super gene families, AP2/EREBP, B3, NF-Y, Dof, MYB, and MYC (Song et al., 2016). The number of the TF families identified in oilseed crops is much great than in non-oilseed plants (**Supplementary File 1**: Figure S53). Information related to these genes involved in fatty acid and triacylglycerol metabolic pathways is important for modifying the oil quality of peanut as well as other oilseed crops.

CONCLUSIONS

The draft genome sequence of *A. ipaensis*, together with those of *L. japonicus*, *M. truncatula*, *C. cajan*, *C. arietinum*, and *G. max*, will provide new biological information for an important branch of the legume clade. The *A. ipaensis* genome sequence presented here, combined with our previous sequence of *A. duranensis*, will shed light on the genomic evolution and polyploidization mechanisms of polyploid species. In addition, the biological information of the *A. ipaensis* genome provides a fundamental resource for understanding disease resistance, symbiotic nitrogen fixation, environmental adaptation and oil biosynthesis in peanut. Moreover, high-density molecular markers, such as SSRs and SNPs, identified in the *A. ipaensis* draft genome can be used to investigate the genetic diversity and make genetic changes to improve important agronomic traits in peanut.

MATERIALS AND METHODS

Plant Material

The *Arachis* genus is composed mostly of diploid species ($2n = 2x = 20$). Peanut (*A. hypogaea* L.) is an allotetraploid (AABB-type genome; $2n = 4x = 40$), probably derived from a single recent hybridization event between *A. duranensis* (AA subgenome, $2n = 2x = 20$) and *A. ipaensis* (BB subgenome, $2n = 2x = 20$) (**Supplementary File 1**: Figure S1; Koppolu et al., 2010; Chen et al., 2016). In 2016, an accession of *A. ipaensis* K30076 has already been sequenced (Bertioli et al., 2016). The accession collected by A. Krapovickas, W.C. Gregory, D.J. Banks, J.R. Pietrarelli, A. Schinini, and C.E. Simpson in 1977 was maintained at Embrapa Genetic Resources and Biotechnology (Brasília, Brazil), which probably originated form Villa Montes near Camatindi or Tigüipa, Bolivia (https://www.peanutbase.org/; Bertioli et al., 2016). In this study, the accession of *A. ipaensis* ICG_8206 maintained at International Centre for Research in the Semi-Arid Tropics (India) then introduced to Crops Research Institute-Guangdong Academy of Agricultural Sciences (China) was used. Although cytogenetic, phylogeographic and molecular evidence showed that the accession of *A. ipaensis* K30076 was the most probable B-genome donor for *A. hypogaea* (Seijo et al., 2007; Robledo and Seijo, 2010; Bertioli et al., 2016), genetic relationship analyses indicated that the B-genome accession ICG 8206 (*A. ipaensis*) was the most closely related to *A. hypogaea* (Koppolu et al., 2010).

Here, the *A. ipaensis* (ICG_8206) was sequenced by Illumina HiSeq2500 platform. Total genomic DNA was isolated from the etiolated unopened young leaves of 20-day-old plants cultivated

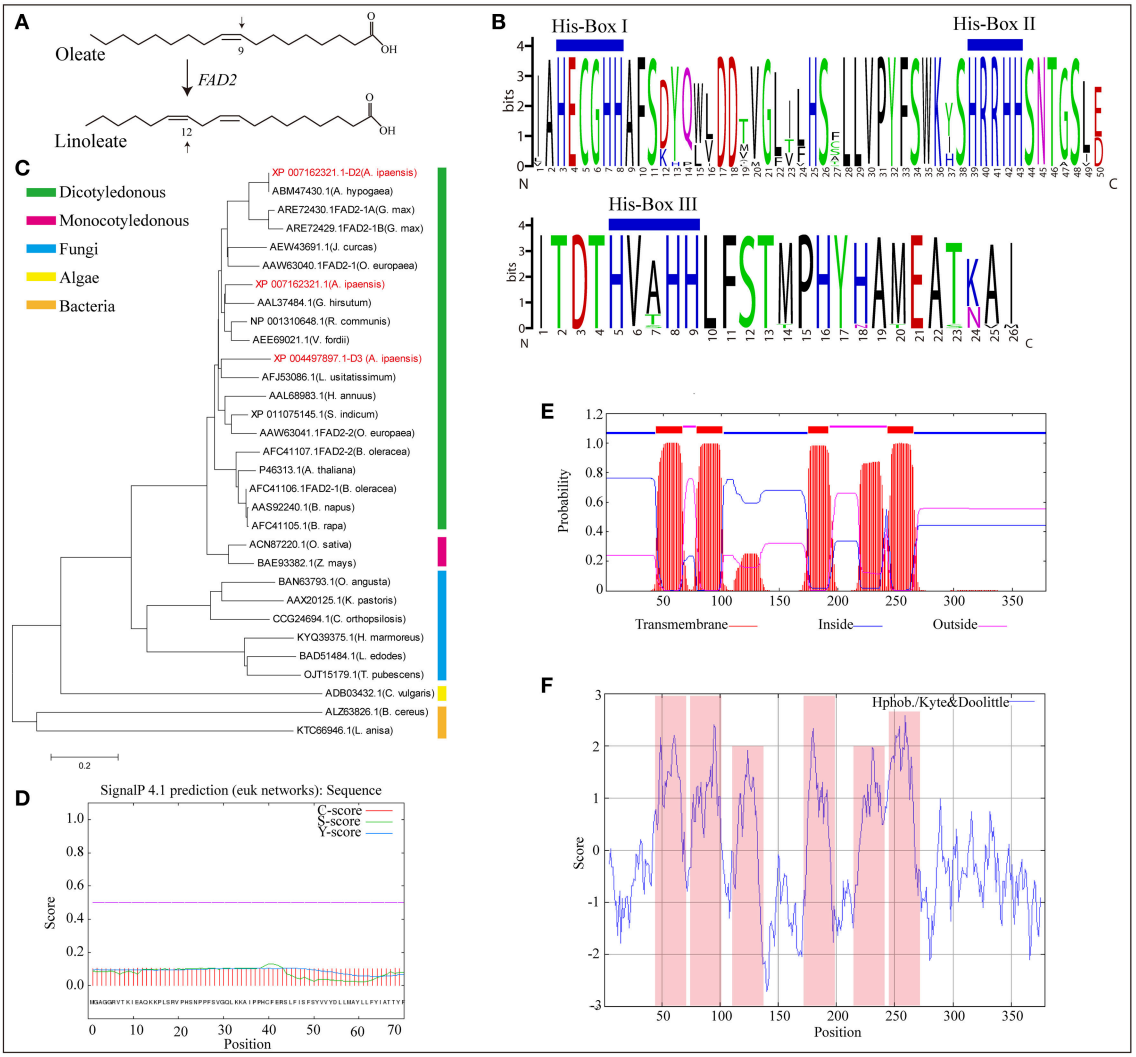

FIGURE 4 | Homologous genes of δ-12 oleic acid desaturase (*FAD2*). **(A)** *FAD2* catalyze oleate into linoleate. **(B)** Multiple alignment of amino acid sequence of substrate binding motif of FAD2 in oil seed plants and its homologous genes in *A. ipaensis*. **(C)** Phylogenetic tree of FAD2 and its homologous genes from different species. **(D)** Signal peptides analysis of FAD2 homologous gene (XP_007162321.1-D2) from *A. ipaensis*. **(E)** Tansmembrane region prediction of FAD2 homologous gene, XP_007162321.1-D2. Red, blue, and pink boxes represent transmembrane, inside, and outside domains. **(F)** Hydrophobicity and hydrophilicity prediction for the homologous gene XP_007162321.1-D2. Pink box represent protein hydrophobic region.

in dark chamber according to a modified CTAB procedure (Doyle and Doyle, 1990). This work will also be of great importance to guide cultivated peanut's genome assembly as a necessary complement in future.

Whole-Genome Shotgun Sequencing and *de Novo* Assembly

Whole-genome shotgun sequencing was performed under the HiSeq2500 Sequencing System with 11 paired-end sequencing libraries, including 250, 500, 800 bp, 2, 5, 10, and 20 Kb using the standard protocol provided by Illumina (San Diego, USA).

SOAPdenovo2 (version 2.04.4) (Luo et al., 2012) was employed with optimized parameters to construct contigs and original scaffolds as previous described (Chen et al., 2016). Subsequently, SSPACE (version 2.0) (Boetzer et al., 2011) was used to link the scaffolds constructed by the SOAPdenovo2 as previous described (Chen et al., 2016).

The genome size was estimated based on the 17 *K*-mer distribution using the total length of sequence reads divided by sequencing depth, and the frequency of each 17-mer were calculated from the whole genome sequenced reads to evaluate the sequencing depth. Subsequently, the *A. ipaensis* genome size was calculated by following the algorithm: Genome size = *K*-mer number/Peak depth (Bertioli et al., 2016).

The gene coverage of the assembled genome was comprehensively evaluated using available public transcript sequence tags or expressed sequence tags. Core eukaryotic genes identified by CEGMA v.2.3 (Parra et al., 2007) were remapped

to the *A. ipaensis* genome assembly by BLAT (Kent, 2002) to evaluate the quality of the assembly. CEGMA data were downloaded from the Korf Lab research group at the Genome Center, UC Davis (http://korflab.ucdavis.edu/datasets/cegma/#SCT6).

Gene Prediction and Function Annotation

To annotate the *A. ipaensis* genome, an automated genome annotation pipeline MAKER was performed to produce *de novo* gene prediction, infer 5′ and 3′ UTR, and integrate these data to generate final downstream gene models with quality control statistics (Cantarel et al., 2008). All predicted genes were functionally annotated as previous described (Chen et al., 2016). The annotation was conducted using the BLAST+ (version 2.2.27) with 1e-5 as the E-value threshold to against the SwissProt and TrEMBL databases (Bairoch and Apweiler, 2000). To infer functions for the predicted genes, InterProScan (version 4.7) (Zdobnov and Apweiler, 2001) was used to search the predicted genes against the protein signature from InterPro with default parameters. All genes were also aligned against to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Kanehisa et al., 2004).

In order to evaluate the conservation of *A. ipaensis* ICG_8206 gene model, the BLASTP was used to query the *A. ipaensis* ICG_8206 proteome against the proteomes of other plant species (**Supplementary File 1**: Table S7) with an E value of 1e-10 as cut-off (**Supplementary File 1**: Table S8).

Gene Family Analysis

All the predicted gene models were analyzed using OrthoMCL (Li et al., 2003) to identify shared and specific gene families among 17 species (**Supplementary File 1**: Table S7). In the first step, inter and intra species BLASTP with an E-value cutoff of 1e-10 was performed to detect reciprocal best hit pairs between species (putative orthologs), as well as sets of genes within species (putative co-orthologs or in-paralogs). The reciprocal best hit matrix served as the basis for ortholog definition using OrthoMCL. Subsequently, orthologous groups were organized into species-specific and higher taxonomic level groups by requiring that at least one sequence from each clade under comparison be present in the intersecting set. Finally, based on fourfold degenerate sites of single-copy ortholog genes in all species, a phylogenetic tree was constructed using MEGA v6.0 (Tamura et al., 2013) and PhyML v3.0 (Guindon et al., 2010).

To identify TFs in *A. ipaensis*, the PlantTFDB database was used to search TFs in other plant species (http://planttfdb.cbi.pku.edu.cn/). The predicted genes were used to BLAST search against the PlantTFDB (E-value: 1e-10). The FAR1 motif was predicted using the Multiple Expectation Maximization for Motif Elicitation (MEME)/Motif Alignment and Search Tool (MAST) system (http://meme-suite.org/) (Bailey et al., 2009) and visualized using the TBtools (version 0.4999) (https://github.com/CJ-Chen/TBtools).

Non-coding RNAs and Repetitive Sequence Annotation

Non-coding RNAs were predicted by aligned *A. ipaensis* genome assembly to against the Rfam database (version 12.1) (Nawrocki et al., 2015). The pre-tRNAs were identified using tRNAscan-SE (Lowe and Eddy, 1997), pre-rRNAs were predicted using RNAmmer (Lagesen et al., 2007), pre-snRNAs were annotated using INFERNAL (Nawrocki et al., 2009) and others were also obtained by BLAST search against the Rfam database.

The RepeatMasker (Chen, 2004), RepeatProteinMask (http://repeatmasker.org/), Tendem Repeats Finder (TRF) (Benson, 1999) and RepeatModeler (Smith and Hubley, 2014) were performed to identify repetitive sequences through homolog and *de novo* prediction. The RepeatMasker and RepeatProteinMask were used to screen the *A. ipaensis* genome against the RepBase database (http://www.girinst.org/). The transposable elements (TEs) were classified as described without consideration of the gaps in the genome assembly (Wicker et al., 2007).

Identification of SSRs and SNPs

MIcroSAtellite (http://pgrc.ipk-gatersleben.de/misa/) was used to mine SSRs in *A. ipaensis* genome, and primer 3 v3.0 was used for primer design (Thiel et al., 2003; Untergasser et al., 2012). A SSR was defined with at least 6 repeats for di-nucleotide motifs or 4 repeats for tri-, tetra-, penta-, and hexa-nucleotide motifs. The maximum number of interrupting nucleotides in a compound SSR was set as 100.

Reads from five re-sequenced genotypes including two A-genome genotypes (ICG_8123 and ICG_8138) and three B-genome genotypes (ICG_8960, ICG_8209, and ICG_13160) were used to identify genome SNP and InDel variations (Chen et al., 2016). Total of these sequenced reads were aligned to the reference genome (ICG_8026) using the Burrows Wheeler Aligner program (BWA) (Li and Durbin, 2009). Subsequently, SNPs and InDels were identified using GATK v3.5 (http://www.broadinstitute.org/gatk) with default parameters, respectively.

Evolutionary and Syntenic Block Analyses

The phylogenetic tree was constructed based on single-copy orthologous genes shared in *A. ipaensis* and other 17 plants (**Supplementary File 1**: Table S7) using MEGA v6.0 with the maximum-liklihood algorithm (Tamura et al., 2013).

Syntenic blocks between the genomes of *A. ipaensis* and other plants were identified using the MCScanX with default parameters (Wang et al., 2012) and visualized on the genome using Circos (Krzywinski et al., 2009). Genomic sequences were first aligned annotated genes based on amino acid sequence using Promer package of Mummer (version 3.22) (Delcher et al., 2002). Whole genome dot plots were generated using Mummerplot (Delcher et al., 2002) and Gunplot 5.0 (www.gnuplot.info/). *Ks* values of the homologs within collinearity blocks were calculated using the perl script, add_ka_and_ks_to_collinearity.pl included in MCScanX package, and the median of *Ks* values was considered to be a representative of the collinearity blocks.

Genes Involved in Disease Resistance, Symbiotic Nitrogen Fixation, Environmental Adaptation, and Oil Synthesis

All the disease R genes were identified using the genome assembly of *A. ipaensis* and other plant species as a TBLASTN query to against the PRG datebase with an E-value of 1e-10 as cut-off. Amino acid sequences of all NBS-LRR genes from *A. ipaensis*, *A. duranensis*, and *A. thaliana* were aligned to construct phylogenetic tree using MEGA v6.0 with automatic bootstrap criteria (Maximum Likelihood) (Tamura et al., 2013). The conserved motifs of top 20 homologies NBS-LRR were identified using MEME suite (Bailey et al., 2009; **Supplementary File 1**: Figure S26).

Nodulation regulatory and nodulin genes were identified based on GO analyses. The GO IDs for each gene were obtained through BLAST search against KEGG proteins (E-value: 1e-5). Genes involved in symbiotic nitrogen fixation associated with nodule development and AON process were obtained by comparison with orthologous genes in other legumes using multiple protein sequence alignment in COBALT (https://www.ncbi.nlm.nih.gov/tools/cobalt/). The PredictProtein was used to perform GO terms, protein-protein and protein-DNA binding sites and sub-cellular localization (Yachdav et al., 2014). The SWISS-MODEL was used to predict protein tertiary structure (Biasini et al., 2014).

Genes involved in oil biosynthesis for *Arabidopsis* were obtained from the Arabidopsis Lipid Gene Database (Mekhedov) (http://lipids.plantbiology.msu.edu/). All the *Arabidopsis* lipid genes (81) in the database were used to TBLASTN search against the *A. ipaensis* genome with a cutoff E-value of 1e-50. Finally, a total of 116 non-redundant oil biosynthesis genes were obtained in *A. ipaensis*. Multiple amino acid sequence alignment of *FAD2* homologs was performed using the COBALT (https://www.ncbi.nlm.nih.gov/tools/cobalt/). The PredictProtein and SWISS-MODEL was used to integrate GO terms, protein binding sites, sub-cellular localization and protein tertiary structure, respectively (Biasini et al., 2014; Yachdav et al., 2014).

Signal peptide analysis of the XP_007162321.1-D2 was predicted using SignalP 4.1 Server with default parameter (Petersen et al., 2011). Prediction of transmembrane helices was performed using TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/). Hydrophobicity and hydrophilicity regions were predicted using ProtScale (Gasteiger et al., 2005).

AUTHOR CONTRIBUTIONS

XQL, and XC designed the experiments and managed the project. QL, HFL, and YH performed the research. QL, SW, XYL, GYZ, SL, HL, and HYL analyzed the data. QL wrote the manuscript with the help of GQZ, ZL, and RV. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31771841 and 31501246); the Science and Technology Planning Project of Guangdong Province (2013B050800021, 2015A030313565, 2015B020231006, 2016B020201003, 2013B020301014, 2017A030311007); the Modern Agro-industry Technology Research System (CARS-13); the Research and Demonstration of Agricultural Technology Demand in Guangdong (2016LM3161, 2016LM3164); the Key Discipline Construction of the Guangdong Academy of Agricultural Sciences (201609); and the Special Foundation of President of the Guangdong Academy of Agricultural Sciences (201831).

ACKNOWLEDGMENTS

We sincerely thank all the participants. In particular, we thank Jianan Zhang and Haofa Lan for helping us to analysis and upload the sequencing data. We thank Lesley Benyon, Ph.D., from Liwen Bianji, Edanz Group 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.2018.00604/full#supplementary-material

The genome assembly and annotation data were deposited in the DRYAD digital repository (https://datadryad.org/). All the data can be downloaded from DRYAD by searching for doi: 10.5061/dryad.hm5vs13 or by directly accessing the link: https://doi.org/10.5061/dryad.hm5vs13.

Supplementary File 1 | Supplementary **Tables S1–S21** and **Figures S1–S53**.

Supplementary File 2: Data S1 | Details on the 71 miRNA identified in *A. ipaensis*.

Supplementary File 3: Data S2 | Summary of primer sequences for SSR markers.

Supplementary File 4: Data S3 | Summary of orthologous gene groups.

Supplementary File 5: Data S4 | Detils on the 1,437 putative disease resistance genes in *A. ipaensis*.

Supplementary File 6: Data S5 | Summary of 16 nodulin and nodulation associated genes in *A. ipaensis*.

Supplementary File 7: Data S6 | Summary of multiple genes associated with nodulation development and nodule autoregulation (AON) signal pathway.

Supplementary File 8: Data S7 | Summary of 39 subtilisin-like protease (SDD1) genes in *A. ipaensis*.

Supplementary File 9: Data S8 | List of *A. ipaensis* genes orthologous to encoding key enzymes in the lipid biosynthesis pathways.

Supplementary File 10: Data S9 | Details on the 82 oil sythesis genes with high *Ka/Ks* values (>1) in *A. ipaensis*.

REFERENCES

Ané, J. M., Kiss, G. B., Riely, B. K., Penmetsa, R. V., Oldroyd, G. E., Ayax, C., et al. (2004). *Medicago truncatula* DMI1 required for bacterial and fungal symbioses in legumes. *Science* 303, 1364–1367. doi: 10.1126/science.1092986

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME suite: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48. doi: 10.1093/nar/28.1.45

Benson, G. (1999). Tandem repeats finder, a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Berger, D., and Altmann, T. (2000). A subtilisin-like serine protease involved in the regulation of stomatal density and distribution in *Arabidopsis thaliana*. *Genes Dev.* 14, 1119–1131.

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K., et al. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat Genet.* 48, 438–446. doi: 10.1038/ng.3517

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 42, W252–W258. doi: 10.1093/nar/gku340

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579. doi: 10.1093/bioinformatics/btq683

Bonaventure, G., Salas, J. J., Pollard, M. R., and Ohlrogge, J. B. (2003). Disruption of the FATB gene in Arabidopsis demonstrates an essential role of saturated fatty acids in plant growth. *Plant Cell.* 15, 1020–1033. doi: 10.1105/tpc.008946

Borisov, A. Y., Madsen, L. H., Tsyganov, V. E., Umehara, Y., Voroshilova, V. A., Batagov, A. O., et al. (2003). The Sym35 gene required for root nodule development in pea is an ortholog of Nin from *Lotus japonicus*. *Plant Physiol.* 131, 1009–1017. doi: 10.1104/pp.102.016071

Cannon, S. B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., et al. (2006). Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14959–14964. doi: 10.1073/pnas.0603228103

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., et al. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18, 188–196. doi: 10.1101/gr.6743907

Capoen, W., Goormachtig, S., de Rycke, R., Schroeyers, K., and Holsters, M. (2005). SrSymRK, a plant receptor essential for symbiosome formation. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10369–10374. doi: 10.1073/pnas.0504250102

Catoira, R., Galera, C., de Billy, F., Penmetsa, R. V., Journet, E. P., Maillet, F., et al. (2000). Four genes of *Medicago truncatula* controlling components of a nod factor transduction pathway. *Plant Cell.* 12, 1647–1666. doi: 10.1105/tpc.12.9.1647

Chen, N. (2004). Using repeatmasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* Chapter 4: Unit 4.10. doi: 10.1002/0471250953.bi0410s05

Chen, X., Li, H., Pandey, M. K., Yang, Q., Wang, X., Garg, V., et al. (2016). Draft genome of the peanut A-genome progenitor (*Arachis duranensis*) provides insights into geocarpy, oil biosynthesis, and allergens. *Proc. Natl. Acad. Sci. U.S.A.* 113, 6785–6790. doi: 10.1073/pnas.1600899113

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* 30, 2478–2483. doi: 10.1093/nar/30.11.2478

Dhillon, S. S., Rake, A. V., and Miksche, J. P. (1980). Reassociation kinetics and cytophotometric characterization of peanut (*Arachis hypogaea* L.) DNA. *Plant Physiol.* 65, 1121–1127. doi: 10.1104/pp.65.6.1121

Dörmann, P., Voelker, T. A., and Ohlrogge, J. B. (2000). Accumulation of palmitate in Arabidopsis mediated by the acyl-acyl carrier protein thioesterase FATB1. *Plant Physiol.* 123, 637–644. doi: 10.1104/pp.123.2.637

Doyle, J., and Doyle L. (1990). Isolation of plant DNA from fresh tissue. *Focus* 12, 13–15.

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., and Lepiniec, L. (2010). MYB transcription factors in Arabidopsis. *Trends Plant Sci.* 15, 573–581. doi: 10.1016/j.tplants.2010.06.005

Edwards, A., Heckmann, A. B., Yousafzai, F., Duc, G., and Downie, J. A. (2007). Structural implications of mutations in the pea SYM8 symbiosis gene, the DMI1 ortholog, encoding a predicted ion channel. *Mol. Plant Microbe Interact.* 20, 1183–1191. doi: 10.1094/MPMI-20-10-1183

Endre, G., Kereszt, A., Kevei, Z., Mihacea, S., Kaló, P., and Kiss, G. B. (2002). A receptor kinase gene regulating symbiotic nodule development. *Nature* 417, 962–966. doi: 10.1038/nature00842

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005). "Protein identification and analysis tools on the ExPASy server," in *The Proteomics Protocols Handbook*, ed J. M. Walker (Totowa, NJ: Humana press), 571–607. doi: 10.1385/1-59259-890-0:571

Grabiele, M., Chalup, L., Robledo, G., and Seijo, G. (2012). Genetic and geographic origin of domesticated peanut as evidenced by 5S rDNA and chloroplast DNA sequences. *Plant Syst. Evol.* 298, 1151–1165. doi: 10.1007/s00606-012-0627-3

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Imaizumi-Anraku, H., Takeda, N., Charpentier, M., Perry, J., Miwa, H., Umehara, Y., et al. (2005). Plastid proteins crucial for symbiotic fungal and bacterial entry into plant roots. *Nature* 433, 527–531. doi: 10.1038/nature03237

Jin, J., He, K., Tang, X., Li, Z., Lv, L., Zhao, Y., et al. (2015). An Arabidopsis transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors. *Mol. Biol. Evol.* 32, 1767–1773. doi: 10.1093/molbev/msv058

Jin, J., Tian, F., Yang, D. C., Meng, Y.Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D104–D105. doi: 10.1093/nar/gkw982

Kaló, P., Gleason, C., Edwards, A., Marsh, J., Mitra, R. M., Hirsch, S., et al. (2005). Nodulation signaling in legumes requires NSP2, a member of the GRAS family of transcriptional regulators. *Science* 308, 1786–1789. doi: 10.1126/science.1110951

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280. doi: 10.1093/nar/gkh063

Kent, W. J. (2002). BLAT-The BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Kimpel, J. A., Nagao, R. T., Goekjian, V., and Key, J. L. (1990). Regulation of the heat shock response in soybean seedlings. *Plant Physiol.* 94, 988–995. doi: 10.1104/pp.94.3.988

Kochert, G., Halward, T., Branch, W. D., and Simpson, C. E. (1991). RFLP variability in peanut (*Arachis hypogaea* L.) cultivars and wild species. *Theor. Appl. Genet.* 81, 565–570. doi: 10.1007/BF00226719

Kochert, G., Stalker, H. T., Gimenes, M., Galgaro, L., Lopes, C. R., and Moore, K. (1996). RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, *Arachis hypogaea* (leguminosae). *Am. J. Bot.* 83, 1282–1291. doi: 10.1002/j.1537-2197.1996.tb13912.x

Koppolu, R., Upadhyaya, H. D., Dwivedi, S. L., Hoisington, D. A., and Varshney, R. K. (2010). Genetic relationships among seven sections of genus Arachis studied by using SSR markers. *BMC Plant Biol.* 10:15. doi: 10.1186/1471-2229-10-15

Krapovickas, A., and Gregory, W. C. (1994). Taxonomia del genero "Arachis (leguminosae)". *Bonplandia* 8, 1–186.

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

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H. H., Rognes, T., and Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108. doi: 10.1093/nar/gkm160

Lévy, J., Bres, C., Geurts, R., Chalhoub, B., Kulikova, O., Duc, G., et al. (2004). A putative Ca^{2+} and calmodulin-dependent protein kinase required for bacterial and fungal symbioses. *Science* 303, 1361–1364. doi: 10.1126/science.1093038

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, L., Stoeckert, C. J. Jr., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Lin, R., Ding, L., Casola, C., Ripoll, D. R., Feschotte, C., and Wang, H. (2007). Transposase-derived transcription factors regulate light signaling in Arabidopsis. *Science* 18, 1302–1305. doi: 10.1126/science.1146281

Lna, Y. (1996). Pattern of synonymous and nonsynonymous substitutions: an indicator of mechanisms of molecular evolution. *J. Genet.* 75, 91. doi: 10.1007/BF02931754

López, Y., Nadaf, H. L., Smith, O. D., Connell, J. P., Reddy, A. S., and Fritz, A. K. (2000). Isolation and characterization of the Δ12-fatty acid desaturase in peanut (*Arachis hypogaea* L.) and search for polymorphisms for the high oleate trait in Spanish market-type lines. *Theor. Appl. Genet.* 101, 1131–1138. doi: 10.1007/s001220051589

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.0955

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*. 1:18. doi: 10.1186/2047-217X-1-18

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*. 290, 1151–1155. doi: 10.1126/science.290.5494.1151

Medzihradszky, M., Bindics, J., Ádám, É., Viczián, A., Klement, É., Lorrain, S., et al. (2013). Phosphorylation of phytochrome B inhibits light-induced signaling via accelerated dark reversion in Arabidopsis. *Plant Cell*. 25, 535–544. doi: 10.1105/tpc.112.106898

Moretzsohn, M. C., Gouvea, E. G., Inglis, P. W., Leal-Bertioli, S. C., Valls, J. F., and Bertioli, D. J. (2013). A study of the relationships of cultivated peanut (*Arachis hypogaea*) and its most closely related wild species using intron sequences and microsatellite markers. *Ann. Bot.* 111, 113–126. doi: 10.1093/aob/mcs237

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., et al. (2015). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* 43, D130–D137. doi: 10.1093/nar/gku1063

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25, 1335–1337. doi: 10.1093/bioinformatics/btp157

Nielen, S., Campos-Fonseca, F., Leal-Bertioli, S., Guimarães, P., Seijo, G., Town, C., et al. (2010). FIDEL–a retrovirus-like retrotransposon and its distinct evolutionary histories in the A- and B-genome components of cultivated peanut. *Chromosome Res.* 18, 227–246. doi: 10.1007/s10577-009-9109-z

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067. doi: 10.1093/bioinformatics/btm071

Parthasarathy, S., Parthasarathy, S., Khoo, J. C., Miller, E., Barnett, J., Witztum, J. L., et al. (1990). Low density lipoprotein rich in oleic acid is protected against oxidative modification: implication for dietary prevention of atherosclerosis. *Proc. Natl. Acad. Sci. U.S.A.* 87, 3894–3898. doi: 10.1073/pnas.87.10.3894

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Pham, A. T., Shannon, J. G., and Bilyeu, K. D. (2012). Combinations of mutant FAD2 and FAD3 genes to produce high oleic acid and low linolenic acid soybean oil. *Theor. Appl. Genet.* 125, 503–515. doi: 10.1007/s00122-012-1849-z

Radutoiu, S., Madsen, L. H., Madsen, E. B., Felle, H. H., Umehara, Y., Grønlund, M., et al. (2003). Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature* 425, 585–592. doi: 10.1038/nature02039

Ramos, M. L., Fleming, G., Chu, Y., Akiyama, Y., Gallo, M., and Ozias-Akins, P. (2006). Chromosomal and phylogenetic context for conglutin genes in Arachis based on genomic sequence. *Mol. Gen. Genomics.* 275, 578–592. doi: 10.1007/s00438-006-0114-z

Robledo, G. and Seijo, G. (2010). Species relationships among the wild B genome of Arachis species (section Arachis) based on FISH mapping of rDNA loci and heterochromatin detection: a new proposal for genome arrangement. *Theor. Appl. Genet.* 121, 1033–1046. doi: 10.1007/s00122-010-1369-7

Robledo, G., Lavia, G. I., and Seijo, G. (2009). Species relations among wild Arachis species with the A genome as revealed by FISH mapping of rDNA loci and heterochromatin detection. *Theor. Appl. Genet.* 118, 1295–1307. doi: 10.1007/s00122-009-0981-x

Sakai, K., and Kajiwara, S. (2005). Cloning and functional characterization of a Δ12 fatty acid desaturase gene from the basidiomycete *Lentinula edodes*. *Mol. Gen. Genomics* 273, 336–341. doi: 10.1007/s00438-005-1138-5

Samoluk, S. S., Robledo, G., Podio, M., Chalup, L., Ortiz, J. P., Pessino, S. C., et al. (2015). First insight into divergence, representation and chromosome distribution of reverse transcriptase fragments from L1 retrotransposons in peanut and wild relative species. *Genetica* 143, 113–125. doi: 10.1007/s10709-015-9820-y

Sanseverino, W., Roma, G., De Simone, M., Faino, L., Melito, S., Stupka, E., et al. (2010). PRGdb, a bioinformatics platform for plant resistance gene analysis. *Nucleic Acids Res.* 38, D814–D821. doi: 10.1093/nar/gkp978

Schauser, L., Roussis, A., Stiller, J., and Stougaard, J. (1999). A plant regulator controlling development of symbiotic root nodules. *Nature* 402, 191–195. doi: 10.1038/46058

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Schneider, A., Walker, A., Sagan, M., Duc, G., Ellis, N., and Downie, A. (2002). Mapping of the nodulation loci sym9 and sym10 of pea (*Pisum sativum* L.). *Theor. Appl. Genet.* 104, 1312–1316. doi: 10.1007/s00122-002-0896-2

Seijo, G., Lavia, G. I., Fernández, A., Krapovickas, A., Ducasse, D. A., and Bertioli, D. J. (2007). Genomic relationships between the cultivated peanut (*Arachis hypogaea*, Leguminosae) and its close relatives revealed by double GISH. *Am. J. Bot.* 94, 1963–1971. doi: 10.3732/ajb.94.12.1963

Serrano-Vega, M. J., Garcés, R., and Martínez-Force, E. (2005). Cloning, characterization and structural model of a FatA-type thioesterase from sunflower seeds (*Helianthus annuus* L.). *Planta* 221, 868–880. doi: 10.1007/s00425-005-1502-z

Shinozaki, K., and Yamaguchi-Shinozaki, K. (2000). Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. *Curr. Opin. Plant Biol.* 3, 217–223. doi: 10.1016/S1369-5266(00)00067-4

Singh, K., Foley, R. C., and O-ate-Sánchez, L. (2002). Transcription factors in plant defense and stress responses. *Curr. Opin. Plant Biol.* 5, 430–436. doi: 10.1016/S1369-5266(02)00289-3

Slabas, A. R., and Fawcett, T. (1992). The biochemistry and molecular biology of plant lipid biosynthesis. *Plant Mol. Biol.* 19, 169–191. doi: 10.1007/BF00015613

Smit, P., Raedts, J., Portyanko, V., Debellé, F., Gough, C., Bisseling, T., et al. (2005). NSP1 of the GRAS protein family is essential for rhizobial Nod factor-induced transcription. *Science* 308, 1789–1791. doi: 10.1126/science.1111025

Smith, A. F. A., and Hubley, R. (2014). Repeatmodeler open-1.0. [Internet]. Available online at: http://www.repeatmasker.org

Smith, B. W. (1950). Arachis hypogaea. Aerial flower and subterranean fruit. *Am. J. Bot.* 37, 802–815. doi: 10.1002/j.1537-2197.1950.tb11073.x

Song, L., Fan, C., Chen, Y., Zhang, X., and Hu, Z. (2016). The molecular regulation mechanism of the plant lipid biosynthesis. *Mol. Plant Breed.* 14, 2178–2187. doi: 10.13271/j.mpb.014.002178

Stracke, S., Kistner, C., Yoshida, S., Mulder, L., Sato, S., Kaneko, T., et al. (2002). A plant receptor-like kinase required for both bacterial and fungal symbiosis. *Nature* 417, 959–962. doi: 10.1038/nature00841

Stracke, S., Sato, S., Sandal, N., Koyama, M., Kaneko, T., Tabata, S., et al. (2004). Exploitation of colinear relationships between the genomes of *Lotus japonicus*, *Pisum sativum* and *Arabidopsis thaliana*, for positional cloning of a legume symbiosis gene. *Theor Appl Genet.* 108, 442–449. doi: 10.1007/s00122-003-1438-2

Sun, W., Van Montagu, M., and Verbruggen, N. (2002). Small heat shock proteins and stress tolerance in plants. *Biochim. Biophys. Acta* 1577, 1–9. doi: 10.1016/S0167-4781(02)00417-7

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tan, D. Y., Zhang, Y., and Wang, A. B. (2010). A review of geocarpy and amphicarpy in angiosperms, with special reference to their ecological adaptive significance. *Chin. J. Plant Ecol.* 34, 72–88. doi: 10.3773/j.issn.1005-264x.2010.01.011

Temsch, E. M., and Greilhuber, J. (2000). Genome size variation in *Arachis hypogaea* and *A. monticola* re-evaluated. *Genome* 43, 449–451. doi: 10.1139/g99-130

Thiel, T., Michalek, W., Varshney, R. K., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet*. 106, 411–422. doi: 10.1007/s00122-002-1031-0

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res*. 40:e115. doi: 10.1093/nar/gks596

Varshney, R. K., Chen, W., Li, Y., Bharti, A. K., Saxena, R. K., Schlueter, J. A., et al. (2011). Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol*. 30, 83–89. doi: 10.1038/nbt.2022

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., et al. (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol*. 31, 240–246. doi: 10.1038/nbt.2491

Wang, W., Vinocur, B., Shoseyov, O., and Altman, A. (2004). Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci*. 9, 244–252. doi: 10.1016/j.tplants.2004.03.006

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res*. 40:e49. doi: 10.1093/nar/gkr1293

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet*. 8, 973–982. doi: 10.1038/nrg2165

Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., et al. (2014). PredictProtein-an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res*. 42, W337–W343. doi: 10.1093/nar/gku366

Yen, C. L., Stone, S. J., Koliwad, S., and Harris, C. Jr. (2008). DGAT enzymes and triacylglycerol biosynthesis. *J. Lipid Res*. 49, 2283–2301. doi: 10.1194/jlr.R800018-JLR200

Young, N. D., Debellé, F., Oldroyd, G. E., Geurts, R., Cannon, S. B., Udvardi, M. K., et al. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*. 480, 520–524. doi: 10.1038/nature10625

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan-an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848. doi: 10.1093/bioinformatics/17.9.847

Zhao, L., Katavic, V., Li, F., Haughn, G. W., and Kunst, L. (2010). Insertional mutant analysis reveals that long-chain acyl-CoA synthetase 1 (LACS1), but not LACS8, functionally overlaps with LACS9 in Arabidopsis seed oil biosynthesis. *Plant J*. 64, 1048–1058. doi: 10.1111/j.1365-313X.2010.04396.x

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.

frontiers in Plant Science

METHODS
published: 26 April 2018
doi: 10.3389/fpls.2018.00564

Haplotype-Based Genotyping in Polyploids

Josh P. Clevenger[1], Walid Korani[2], Peggy Ozias-Akins[2] and Scott Jackson[3]*

[1] Mars-Wrigley Confectionery, Center for Applied Genetic Technologies, Athens, GA, United States, [2] Institute of Plant Breeding, Genetics, and Genomics, College of Agricultural and Environmental Sciences, University of Georgia, Tifton, GA, United States, [3] Institute of Plant Breeding, Genetics, and Genomics, University of Georgia, Athens, GA, United States

Accurate identification of polymorphisms from sequence data is crucial to unlocking the potential of high throughput sequencing for genomics. Single nucleotide polymorphisms (SNPs) are difficult to accurately identify in polyploid crops due to the duplicative nature of polyploid genomes leading to low confidence in the true alignment of short reads. Implementing a haplotype-based method in contrasting subgenome-specific sequences leads to higher accuracy of SNP identification in polyploids. To test this method, a large-scale 48K SNP array (Axiom Arachis2) was developed for *Arachis hypogaea* (peanut), an allotetraploid, in which 1,674 haplotype-based SNPs were included. Results of the array show that 74% of the haplotype-based SNP markers could be validated, which is considerably higher than previous methods used for peanut. The haplotype method has been implemented in a standalone program, HAPLOSWEEP, which takes as input bam files and a vcf file and identifies haplotype-based markers. Haplotype discovery can be made within single reads or span paired reads, and can leverage long read technology by targeting any length of haplotype. Haplotype-based genotyping is applicable in all allopolyploid genomes and provides confidence in marker identification and in silico-based genotyping for polyploid genomics.

Keywords: polyploid SNPs, SNP array, *Arachis*, haplotype markers, sequence-based genotyping

OPEN ACCESS

Edited by:
Shuizhang Fei,
Iowa State University, United States

Reviewed by:
Nahla Victor Bassil,
National Clonal Germplasm Repository (USDA-ARS), United States
Jacob A. Tennessen,
Oregon State University, United States

****Correspondence:***
Josh P. Clevenger
jclev@uga.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *30 January 2018*
Accepted: *10 April 2018*
Published: *26 April 2018*

Citation:
Clevenger JP, Korani W, Ozias-Akins P and Jackson S (2018) Haplotype-Based Genotyping in Polyploids. Front. Plant Sci. 9:564. doi: 10.3389/fpls.2018.00564

INTRODUCTION

The identification of functional variation controlling traits of interest relies on the ability to discern all true variation between accessions with discrete genotypes. The power of next-generation (short reads) and third-generation (long reads) sequencing is the ability to identify all variants. The size and complexity of polyploid genomes have led to the reliance on single nucleotide polymorphism (SNP) arrays and complexity reduction sequencing strategies such as genotyping-by-sequencing (GBS) and restriction site-associated DNA sequencing (RADSeq; Elshire et al., 2011; Willing et al., 2011). These methodologies have allowed access to unprecedented number of markers for genomics. There are drawbacks in using these technologies, however. One is the inherent ascertainment bias in using SNP arrays for genotyping (Lachance and Tishkoff, 2013). Ascertainment bias occurs from the bias associated with sampling smaller populations. Since the SNP probes on arrays are static, rare variants or subpopulation-specific variants will not be assayed. This will cause bias in population genetics studies, and will not allow the identification of rare functional variants controlling traits of interest. A method of identifying markers straight from sequence data alleviates ascertainment bias on an experiment-wise level by providing access to all potential polymorphisms in the population of interest and does not constrain analysis to discrete markers on an array.

With the cost of sequencing continuing to plummet and long read technologies increasing in efficiency and accuracy, the ability to generate sequence on whole populations is increasing. Having access to all potential polymorphisms can increase the resolution of genetic mapping and genome-wide association studies. In polyploids, confidence in sequence-based prediction of genotypes is confounded by the uncertain alignment of short reads in the genome. Mapping of homeologous reads causes confusion by the appearance *in silico* of a polymorphism between accessions that is only between subgenomes. This problem is confounded by variance in sequence coverage across genomic loci.

A method of untangling subgenomes in mapping experiments has been proposed and implemented in cultivated peanut [sliding window extraction of explicit polymorphisms (SWEEP)] and octoploid strawberry (Bassil et al., 2015; Clevenger and Ozias-Akins, 2015). Bassil et al. (2015) did not report the accuracy of marker identification, but SWEEP performed well in simulations. SWEEP was used to design a large scale SNP array and showed that accuracy was useful although lower than estimated (Clevenger et al., 2017b). A more precise method is needed to assign subgenome specificity to mapped reads to more definitively identify polymorphisms between accessions.

We propose a method of sequence-based genotyping in polyploids that instead of applying a filter to individual sites collects observed haplotypes from sequence reads and contrasts those haplotypes between accessions to identify polymorphic markers. To demonstrate the accuracy of the method, haplotype-based markers were validated on a new 48k SNP array for *Arachis*, Axiom Arachis2. Finally, a pipeline was developed to utilize the haplotype-based genotyping method as an easy-to-use one command program. Haplotype-based genotyping should be broadly applicable across allopolyploid species.

MATERIALS AND METHODS

Axiom Arachis2 Design

For SNP identification, a set of 21 *Arachis hypogaea* accessions were re-sequenced to 10X coverage (Clevenger et al., 2017b) and sequences from three accessions that are parents of two RIL populations ["T" (Qin et al., 2012) and "S" (Khera et al., 2016)] were also used. Analysis of sequence and SNP calling was carried out as in Clevenger et al. (2017b). Filtering for high-quality SNPs was then done using two methods. The first method, described here, was haplotype-based markers converted to SNPs. There were a total of 1,746 haplotype-based SNPs submitted to Affymetrix of which 1,674 were selected for the array. An alternative filtering method was used that took SWEEP-filtered SNPs (Clevenger and Ozias-Akins, 2015) and filtered them further using a machine learning approach (SNP-ML[1]). The models used for machine learning were trained using the true and false SNP sets from a previous array, Axiom Arachis v1 (Clevenger et al., 2017b; Pandey et al., 2017). A total of 133,162 putative SNPs were submitted to Affymetrix, of which 28,218 were selected for the array.

In addition, 6,407 markers between Tifrunner and GT-C20 were included that were identified using an early assembly of the cultivated Tifrunner genome[2]. Potential SNPs were filtered by only taking those SNP sites where all Tifrunner reads contained the reference base and all GT-C20 reads contained the alternate base. The remaining 22 markers were added based on their utility in marker-assisted selection. Seven markers select for an alien introgressed region that controls nematode resistance on chromosome A09, including a marker that is within the current candidate gene for resistance (Clevenger et al., 2017a). Seven markers were selected for late and early leaf spot resistance identified using QTL-seq (Clevenger et al., 2018). Eight markers were selected for two alien introgressed regions from *Arachis cardenasii* that control late leaf spot and rust resistance (Pandey et al., 2016; Clevenger et al., 2017c). The final 11,516 markers were included from the Axiom Arachis v1 that were identified as useful in interspecific populations. Of these, 4,489 were high-quality polymorphic markers in *A. hypogaea* populations. Supplementary Table S1 provides the final design of the Axiom Arachis2 SNP array.

Haplotyping Workflow

To identify haplotype-based markers, all possible polymorphic sites were called using Samtools mpileup. These potential polymorphisms were then used as a guide for the haplotyping procedure. The program is written in C++ as a standalone program. To access bam files and retrieve reads aligned to specific locations, the bamtools API is used (Barnett et al., 2011). First, all haplotypes for each accession are collected at every two-position haplotype where there is a potential polymorphism at both sites within a specified base window. Haplotypes are only collected if they occur within a single contiguous read. The haplotypes are stored in a data frame that is organized by genotype, haplotype position, and counts for each observed haplotype. Then the stored haplotypes are filtered based on the following criteria: (1) for a given pair of SNPs, there must be at least two accessions with more than one haplotype; (2) within each of these two accessions, both of its haplotypes must be observed at least twice; (3) within each of these two accessions, reads supporting the least observed haplotype must be at least 25% as frequent as the reads supporting the most observed haplotype (to exclude rare haplotypes that could be due to sequence error in one of the accessions); and (4) in at least one, but not all, accessions, the two haplotypes must be the same for one site and different from each other at the other site.

There are three different run modes to identify haplotypes. The default mode is described as above as haplotypes are collected only within a single contiguous read. Additionally, paired-end information can be leveraged as haplotypes present within two pairs will be considered. This mode is most useful when using large insert size libraries as haplotypes can be considered that span longer distances. The third mode is a simple "diploid" mode where an additional parameter to adjust the number of

[1]https://github.com/w-korani/SNP-ML/wiki

[2]peanutbase.org

bases considered for each haplotype is included. For this mode, the user can specify haplotypes from two up to *N* bases in length. This mode is useful when using long read technology as high-resolution haplotypes can be assayed.

The haplotyping procedure can be called individually or as a part of a pipeline. The pipeline can additionally take called haplotypes and genotype a population of individuals at those sites, giving as output an m × n matrix of genotypes, where m is individuals and n is haplotype markers. Usage and help file information is provided as a README in Supplementary File S1. HaploSWEEP is available under the MIT license at https://github.com/jclev-uga/HAPLOSWEEP.

Array-Based Marker Validation

Array genotype calls were manually curated within the Axiom Analysis Suite 3.1[3] based on methods described in Clevenger et al. (2017b). Each genotype that was used to design the array markers using HAPLOSWEEP was assayed with the array in duplicate. For validation of haplotype-derived markers, there were two considerations: (1) the marker shows true polymorphism between genotypes and (2) the genotype calls on the array match those called from the sequence data.

RESULTS AND DISCUSSION

Haplotype-Based Genotyping Identifies High-Quality Polymorphic SNPs in Polyploids

In an allopolyploid genome, there are generally at least two copies of any chromosomal region. As the divergence of the subgenomes decreases, the co-linearity and sequence similarity increase. When utilizing sequence-based genotyping, short reads originating from each subgenome can both map to the same duplicate location. Because of variance in coverage between sequenced samples, the reads do not always map in the ratio expected, i.e., 50% for each in an allotetraploid, 33% in an allohexaploid, etc.

The question becomes how to differentiate *in silico* between reads originating from each subgenome? One proposal is that using expected sequence coverage and eliminating from consideration any region with higher than expected coverage will properly filter out the regions where both subgenomes map simultaneously. Unfortunately, this strategy does not protect against false positive SNP calls. As an example in peanut, consider three samples sequenced using whole genome shotgun (WGS) sequencing with an expected sequence coverage of 10X. A coverage-based strategy would suggest that any sites be ignored with coverage above 15 reads and below four reads for each accession. Using the diploid progenitor genome sequences, an estimate can be made of the number and location of polymorphisms between subgenomes by fragmenting one genome into overlapping short sequences and mapping them to the alternate genome. Doing this with *Arachis duranensis* (A genome) fragments mapped to *Arachis ipaensis* identifies potentially 8,605,615 polymorphic sites that represent potential false positive SNP calls. After calling SNPs between the three accessions and filtering for expected coverage, there are 2,898,744 of those sites that fall within the expected coverage. In an experiment using coverage-based filtering, all these SNPs are potential false positives. Further, there are 25,459 sites where at least one of the three genotypes is scored as "homozygous" which would be considered a true SNP. Looking at each accession separately and filtering for expected coverage, there are 213,791, 113,340, and 81,120 false SNP sites where only one allele is represented and would be called SNPs, but are false positives. Given that the potential true polymorphisms between accessions are low, these potential false positives would drown out the true signal in a sequence-based genotyping experiment. Even when filtering for higher than expected coverage, the potential is high that many of the SNPs identified will be false positives.

Recently, a pipeline was developed called SWEEP (Clevenger and Ozias-Akins, 2015). This approach differentiates between subgenomes using an "anchor" within a window upstream and/or downstream of a potential SNP site. The anchor is simply an identified polymorphism between subgenomes. Using the genotypic likelihoods calculated from samtools mpileup (Li et al., 2009), putative true positive SNPs are selected when an anchor is within the base pair window in all of the accessions considered and at the SNP site at least one accession has a homozygous call. In simulations, SWEEP selected true SNPs with a rate above 95%; however, using real data, the validation rate was much lower, as seen after design of a 58K SNP array (Clevenger et al., 2017b; Pandey et al., 2017). The limiting factor is still being able to accurately differentiate the two subgenomes correctly using short reads.

Accurate identification of polymorphisms from sequence data in polyploids can be done after contrasting collected observed haplotypes within contiguous sequence reads that align to a common locus. A polymorphic haplotype is identified by containing at one position an "anchor" base that is the same in all observed haplotypes and a polymorphic position that is observed in at least one accession (**Figure 1**). To demonstrate this method, we developed HAPLOSWEEP, which collects all observed overlapping haplotypes within a user-specified base pair window and contrasts them between accessions to identify those haplotypes that are polymorphic between accessions (**Figure 1**).

Validation of the Haplotyping Method Using a 48K SNP Array

A 48K SNP array was developed for *Arachis*, Axiom Arachis2. A total of 1,674 haplotype-based markers were included on this array. To validate these haplotype-based markers, the parents used to identify them *in silico* were assayed on the array in duplicate. Analysis revealed that 1,243 (74%) of the haplotype-based markers could be validated (Supplementary Table S2). Of the 431 that were not validated, further analysis of their probe sequences revealed that the sequences were highly repetitive. The average positions aligned to completely with greater than

[3]www.thermofisher.com

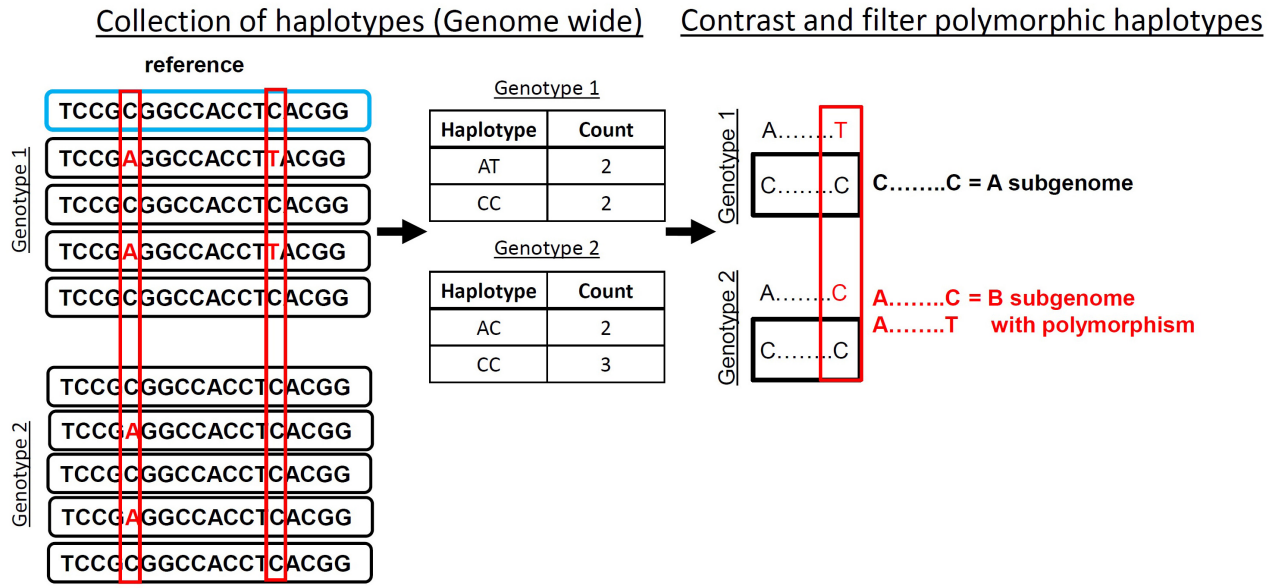

FIGURE 1 | HaploSWEEP pipeline. Top sequence (blue box) represents the reference being aligned too with reads (black boxes) representing two genotypes aligned to that sequence. All haplotypes (in polyploid mode all two base haplotypes; in diploid mode, the user can specify the length of the haplotype) are collected genome-wide that are within the user-specified read length and contain at least one potential polymorphism. Then the haplotypes are analyzed for potential haplotype polymorphisms that can be distinguished from differences between subgenomes. The filtering process, outlined in Section "Materials and Methods," is implemented and high-quality polymorphic haplotypes are output in two different files. One file lists every observed haplotype with number of observations in each accession. The other file lists the polymorphic haplotypes for each accession.

94% identity in validated marker probes was 1.9 where for invalidated marker probes it was 5.7. The enrichment of repetitive probes that are false positives will not allow a clear segregation of the polymorphic locus and so it is unclear if the *in silico* identified haplotype marker is a false positive or not.

Even with a high-quality polyploid reference genome, SNP calling is not trivial. Of the 6,407 markers that were identified by mapping to an early tetraploid assembly, only 2,888 (45%) were validated on the array. This result confirms that the problem of the multiple mapping of short reads is not fully solved by using a polyploid reference assembly.

The remaining markers were identified using a machine learning approach outlined in Korani et al. (unpublished). These markers were validated with a true positive rate (TPR) of 75%.

The true and false SNP sites obtained from both SNP arrays provide a resource for estimating validation rates. The sets provide 44,087 validated true SNP sites and 42,767 validated false SNP sites. These sets can be used to estimate validation for called haplotypes. Using whole genome resequencing from accessions used to design the array, haplotype markers were called in sets of two, three, four, and five accessions (Supplementary Table S3). The called haplotype markers were then compared to the true and false sets from the arrays for overlapping sites. Given the ratio of true to false positive known sites, the expected overlap of called markers is significantly greater with the true set and significantly lower with the false set for all experiments. The average estimated TPR across all experiments is greater than 89%. These validation results combined with the array-based validation show that the haplotype-based genotyping method can produce reliable genotyping results straight from sequence data.

Probability of Identifying Markers Genome-Wide

A caveat to the haplotype-based genotyping method described here is that a polymorphism must be in proximity to an anchor (subgenome polymorphism) within a distance that can be observed on a contiguous short read. In peanut, there are potentially 8,605,615 polymorphisms between the A and B subgenomes. These polymorphisms were identified by mapping fragmented, overlapping sequences from *A. duranensis* to the *A. ipaensis* genome. Given an experiment using 150 bp reads, there are 2,581,684,500 base pairs that are potentially within the range of those polymorphisms. With an estimated genome size of 2.7 billion base pairs, there is the potential to identify the majority of polymorphisms between any set of peanut accessions given a sequencing depth of at least 10-fold genome coverage.

Increased sensitivity and utility can be accomplished by incorporating paired read information. With that aim, a separate module was designed called HAPLOSWEEP_LONGRANGE. This module uses the same methodology to contrast observed haplotypes between accessions, but collects them differently. The data structure stores information on every read's status as a pair and the aligned base at every queried position. After haplotypes are captured that occur within single reads, an additional step is done to collect all haplotypes that occur between paired reads. Then the paired haplotypes are contrasted in the same

way as haplotypes occurring within single reads. Utility for this module is that large insert size libraries can be used to identify polymorphic markers across long distances.

Long Read Utility and Utility in Diploid Genomes

The haplotype-based genotyping framework can be applicable to diploid crops as well. Using long read technologies or large insert size libraries, long haplotypes can increase the resolution of genotypes and precision of mapping. An additional module, HAPLOSWEEP_DIPLOID, was designed to contrast haplotypes of any length. As haplotypes are observed, the program collects all haplotypes within the user-specified window up to the maximum haplotype length specified. When the window is not large enough to observe a maximum length haplotype, shorter haplotypes are collected.

Full Functionality

The modules described above are a part of a one-command pipeline that takes as input a vcf file of called SNPs and bam files of alignments. Maximum utility is achieved if no filtering of the vcf file is done beforehand as the haplotyping method needs to have access to all possible polymorphisms. The user can call any of the three modules based on needs of the experiment and set window size (length of reads) and max haplotype length. Additionally, given a set of called haplotypes, a population of individuals can be genotyped with the output a matrix of called genotypes for each individual.

Flexibility

Cultivated peanut is an allotetraploid. It is the simplest case to illustrate the method described because there are only two subgenomes. Further, the progenitor genomes have been sequenced and can be utilized with this method. Peanut is a special case because the progenitor genomes are very similar to the cultivated genome (Bertioli et al., 2016). In fact, the *A. ipaensis* genome is estimated to show 99.96% similarity to the cultivated B subgenome (Bertioli et al., 2016). In this case, the diploid genomes of peanut act as a proxy for the cultivated reference genome. The functionality of the pipeline is suitable for a case such as this as well as a polyploid reference genome.

As far as higher order allopolyploids such as wheat (hexaploid) or strawberry (octoploid), the pipeline should be applicable. It is designed to find at least one contrasting haplotype. As long as the haplotypes are the same on the other subgenomes, the markers will be identified. In a situation where there are more than two alleles at a specific site, these sites may be filtered out in a higher order polyploid. The pipeline is not designed to work for autopolyploids.

CONCLUSION

Large-scale validation using the Axiom Arachis2 48k SNP array has shown that haplotype-based genotyping in polyploids is more accurate than other methods. By leveraging the true and false SNP sites derived from the version 1 and version 2 *Arachis* SNP arrays, estimates of TPRs are above 89% for novel marker discovery. In order to allow all users access to this method, we developed a one-command pipeline that will call haplotype-based markers and use them to genotype populations. Shown to be accurate in allotetraploid peanut, this method should provide accurate marker identification in all allopolyploid species.

AUTHOR CONTRIBUTIONS

JC, PO-A, and SJ conceptualized the research. JC and WK performed the experiments, conducted data analysis, and curated data. JC wrote the original draft and was responsible for data visualization. SJ, PO-A, and JC revised the manuscript.

ACKNOWLEDGMENTS

Funding for this research was provided by the Peanut Foundation and the Agriculture and Food Research Initiative competitive grant 2017-67012-26118 of the USDA National Institute of Food and Agriculture (JC).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00564/full#supplementary-material

REFERENCES

Barnett, D., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., and Marth, G. T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* 27, 1691–1692. doi: 10.1093/bioinformatics/btr174

Bassil, N., Davis, T. M., Zhang, H., Ficklin, S., Mittmann, M., and Webster, T. (2015). Development and preliminary evaluation of a 90 K Axiom® SNP array for the allo-octoploid cultivated strawberry *Fragaria* × *ananassa*. *BMC Genomics* 16:155. doi: 10.1186/s12864-12015-11310-12861

Bertioli, D., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K., et al. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* 48, 438–446. doi: 10.1038/ng.3517

Clevenger, J., Bertioli, D. J., Leal-Bertioli, S. C. M., Chu, Y., Stalker, H. T., and Ozias-Akins, P. (2017c). IntroMap: a pipeline and set of diagnostic diploid *Arachis* SNPs as a tool for mapping alien introgressions in *Arachis hypogaea*. *Peanut Sci.* 44, 66–73. doi: 10.3146/PS17-5.1

Clevenger, J., Chu, Y., Arrais Guimaraes, L., Maia, T., Bertioli, D., Leal-Bertioli, S., et al. (2017a). Gene expression profiling describes the genetic regulation of *Meloidogyne arenaria* resistance in *Arachis hypogaea* and reveals a candidate gene for resistance. *Sci. Rep.* 7:1317. doi: 10.1038/s41598-017-00971-6

Clevenger, J., Chu, Y., Chavarro, C., Agarwal, G., Bertioli, D. J., Leal-Bertioli, S. C. M., et al. (2017b). Genome-wide SNP genotyping resolves signatures of selection and tetrasomic recombination in peanut. *Mol. Plant* 10, 309–322. doi: 10.1016/j.molp.2016.11.015

Clevenger, J., Chu, Y., Chavarro, C., Botton, S., Culbreath, A. K., Isleib, T. G., et al. (2018). Mapping late leaf spot resistance in peanut (*Arachis hypogaea*) using

QTL-seq reveals markers for marker assisted selection. *Front. Plant Sci.* 9:83. doi: 10.3389/fpls.2018.00083

Clevenger, J., and Ozias-Akins, P. (2015). SWEEP: a tool for filtering high-quality SNPs in polyploid crops. *G3* 5, 1797–1803. doi: 10.1534/g3.115.019703

Elshire, R., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. doi: 10.1371/journal.pone.0019379

Khera, P. P. M., Wang, H., Feng, S., Qiao, L., Culbreath, A. K., Kale, S., et al. (2016). Mapping quantitative trait loci of resistance to tomato spotted wilt virus and leaf spots in a recombinant inbred line population of peanut (*Arachis hypogaea* L.) from SunOleic 97R and NC94022. *PLoS One* 11:e0158452. doi: 10.1371/journal.pone.0158452

Lachance, J., and Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* 35, 780–786. doi: 10.1002/bies.201300014

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

Pandey, M., Agarwal, G., Kale, S. M., Clevenger, J., Nayak, S. N., Sriswathi, M., et al. (2017). Development and evaluation of a high density genotyping 'Axiom_*Arachis*' array with 58?K SNPs for accelerating genetics and breeding in peanut. *Sci. Rep.* 7:40577. doi: 10.1038/srep40577

Pandey, M., Khan, A. W., Singh, V. K., Vishwakarma, M. K., Shasidhar, Y., Kumar, V., et al. (2016). QTL-seq approach identified genomic regions and diagnostic markers for rust and late leaf spot resistance in peanut (*Arachis hypogaea* L.). *Plant Biotechnol. J.* 15, 927–941. doi: 10.1111/pbi.12686

Qin, H., Feng, S., Chen, C., Guo, Y., Knapp, S., Culbreath, A., et al. (2012). An integrated genetic linkage map of cultivated peanut (*Arachis hypogaea* L.) constructed from two RIL populations. *Theor. Appl. Genet.* 124, 653–664. doi: 10.1007/s00122-011-1737-y

Willing, E.-M., Hoffmann, M., Klein, J. D., Weigel, D., and Dreyer, C. (2011). Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics* 27, 2187–2193. doi: 10.1093/bioinformatics/btr346

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: 26 February 2018
doi: 10.3389/fpls.2018.00218

Development and Applications of Chromosome-Specific Cytogenetic BAC-FISH Probes in *S. spontaneum*

*Guangrui Dong[1,2], Jiao Shen[3], Qing Zhang[1,2], Jianping Wang[1,4], Qingyi Yu[1,5], Ray Ming[1,6], Kai Wang[1,2] and Jisen Zhang[1,2,3]**

[1] *Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Center for Genomics and Biotechnology, Haixia Institute of Science and Technology, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China,* [2] *Key Laboratory of Sugarcane Biology and Genetic Breeding, Ministry of Agriculture, Fujian Agriculture and Forestry University, Fuzhou, China,* [3] *College of Life Sciences, Fujian Normal University, Fuzhou, China,* [4] *Agronomy Department, University of Florida, Gainesville, FL, United States,* [5] *Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL, United States,* [6] *Texas A&M AgriLife Research Center, Department of Plant Pathology and Microbiology, Texas A&M University System, Dallas, TX, United States*

Saccharum spontaneum is a major *Saccharum* species that contributed to the origin of modern sugarcane cultivars, and due to a high degree of polyploidy is considered to be a plant species with one of the most complex genetics. Fluorescence *in situ* hybridization (FISH) is a powerful and widely used tool in genome studies. Here, we demonstrated that FISH based on bacterial artificial chromosome (BAC) clones can be used as a specific cytological marker to identify *S. spontaneum* individual chromosomes and study the relationship between *S. spontaneum* and other related species. We screened low-copy BACs as probes from the sequences of a high coverage of *S. spontaneum* BAC library based on BLAST search of the sorghum genome. In total, we isolated 49 positive BAC clones, and identified 27 BAC clones that can give specific signals on the *S. spontaneum* chromosomes. Of the 27 BAC probes, 18 were confirmed to be able to discriminate the eight basic chromosomes of *S. spontaneum*. Moreover, BAC-24, BAC-66, BAC-78, BAC-69, BAC-71, BAC-73, and BAC-77 probes were used to construct physical maps of chromosome 1 and chromosome 2 of *S. spontaneum*, which indicated synteny in Sb01 between *S. spontaneum* and sorghum. Furthermore, we found that BAC-14 and BAC-19 probes, corresponding to the sorghum chromosomes 2 and 8, respectively, localized to different arms of the same *S. spontaneum* chromosome, suggesting that there was an inter-chromosomal rearrangement event between *S. spontaneum* and sorghum. Our study provides the first set of chromosome-specific cytogenetic markers in *Saccharum* and is critical for future advances in cytogenetics and genome sequencing studies in *Saccharum*.

Keywords: *Saccharum spontaneum*, sorghum, polyploidy, bacterial artificial chromosome (BAC), fluorescence *in situ* hybridization (FISH)

OPEN ACCESS

Edited by:
Jun Yang,
Shanghai Chenshan Plant Science Research Center (CAS), China

Reviewed by:
Gabriel Rodrigues Alves Margarido, University of São Paulo, Brazil
Jianying Sun, Jiangsu Normal University, China

***Correspondence:**
Jisen Zhang
zjisen@126.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *29 November 2017*
Accepted: *05 February 2018*
Published: *26 February 2018*

Citation:
Dong G, Shen J, Zhang Q, Wang J, Yu Q, Ming R, Wang K and Zhang J (2018) Development and Applications of Chromosome-Specific Cytogenetic BAC-FISH Probes in S. spontaneum. Front. Plant Sci. 9:218. doi: 10.3389/fpls.2018.00218

INTRODUCTION

Sugarcane (*Saccharum* spp., *Poaceae*) is an annual or perennial crop grown in tropical and subtropical regions worldwide. Sugarcane is an important crop for sucrose production, and accounts for 70% of the world sugar production. In addition, as a C4 plant, sugarcane can efficiently convert solar energy into chemical energy, and is therefore an ideal biofuel crop for

ethanol production (Lam et al., 2009; Santchurn et al., 2014). The genus *Saccharum* has six species: *S. spontaneum*, *S. robustum*, *S. officinarum*, *S. barberi*, *S. sinense*, and *S. edule*, among which, *S. spontaneum* is considered the wild species with a basic chromosome set of $x = 8$. *S. spontaneum* is also one of the two main *Saccharum* species contributing to the modern sugarcane cultivars development (Dhont et al., 1998; Ha et al., 1999). *S. spontaneum* has a strong environmental adaptability and contains important genetic traits for disease and drought resistance, thus contributing to the stress tolerance of modern cultivar hybrids (Grivet et al., 2004). *S. spontaneum* has the widest geographic distribution, and its ploidy levels range from $2n = 5x = 40$ to $2n = 16x = 128$ (Panje and Babu, 1960).

Sugarcane chromosomes have similar morphologies and have a small size of 1–6 μm at the condensed metaphase stage (Ha et al., 1999; Dhont, 2005); hence, it is very challenging to discern between the different chromosomes based on traditional cytogenetic methods. Moreover, as an autopolyploid, its chromosomal structural alterations caused by polyploidization, duplication, deletion, and recombination are very common (Piperidis et al., 2010). In *S. spontaneum*, genetic linkage maps have been developed (Silva et al., 1995), but development of a cytological genetic map lagged behind other grass species. Reliable cytological tools to identify individual chromosomes can be used for effective genome research and germplasm resource utilization, specifically with the complex genome background of modern sugarcane cultivars. Chromosome-specific bacterial artificial chromosome (BAC) clones are invaluable resource for sugarcane genome researches, and will have many applications in physical mapping, chromosome identification, and marker-assisted breeding of the *Saccharum* spp., and in assisting sugarcane genome sequencing and assembly projects.

Fluorescence *in situ* hybridization (FISH) is a powerful tool for molecular cytology (Jiang and Gill, 2006). In *Saccharum* spp., 45S rDNA and 5S rDNA were used as probes to detect the basic chromosomes of *S. spontaneum*, *S. robustum*, and *S. officinarum*, to identify the different basic chromosomes, $x = 8$ in *S. spontaneum*, and $x = 10$ in *S. robustum* and *S. officinarum* (Dhont et al., 1998; Ha et al., 1999). Based on FISH, modern sugarcane cultivars were found to contain 70–80% of the chromosomes derived from *S. officinarum*, and 10–23% of chromosomes from *S. spontaneum*, while 5–17% appears to be the product of the recombination between *S. spontaneum* and *S. officinarum* (Dhont et al., 1996; Piperidis et al., 2001; Cuadrado et al., 2004). Moreover, chromosome elimination, recombination, and translocation events were detected in some progenitors derived from the hybridization between *Saccharum* and *Erianthus arundinaceus* using FISH (Dhont et al., 1995; Huang et al., 2015).

Sorghum has a small diploid genome (730 Mbp) with a low frequency of chromosome recombination events. Sorghum shares high synteny with the sugarcane genome, making it an ideal reference plant for comparative analyses with the sugarcane genome (Ming et al., 1998; Nazeema et al., 2007; Wang et al., 2010; Aitken et al., 2014). In this study, based on the available sequences for high coverage BAC resources, reliable BAC probes were developed for FISH chromosome identification and cytogenetic map construction. The objectives of this study were to: (1) develop a set of chromosome-specific BAC-FISH probes on *S. spontaneum* chromosome and (2) explore the chromosome rearrangement between *S. spontaneum* and sorghum.

MATERIALS AND METHODS

Materials

Saccharum spontaneum SES208 ($2n = 8x = 64$) was used for cytological analyses in this study. The SES208 plants were grown in the field on the campus of Fujian Agricultural and Forestry University (Fuzhou, China) in February of 2015 and maintained under regular sugarcane growth conditions.

A BAC library was constructed from the haploid genome of *S. spontaneum* AP85-441 ($4x = 32$), which was derived from *S. spontaneum* SES208 ($2n = 8x = 64$) via anther *in vitro* culture. The genome size of AP85-441 is about 3.2 Gbp. The library consisted of 38,400 clones, with an average insert size of 100 kb, covering 6× of the whole genome. This BAC library has been used for identifying the sucrose transporters and fructokinase gene families in *S. spontaneum* (Zhang et al., 2016; Chen et al., 2017).

Screening the BAC Library

35,156 BAC clones from the AP85-441 libraries were pooled into 701 libraries. Each library contains an average of 50 BAC clones. The DNA libraries were prepared with the PhasePrep^{TM} BAC DNA Kit (Sigma, United States) following the manufacturer's protocols. BAC DNA libraries were sequenced using Illumina HiSeq 2500 platform with PE250 model. A total of 686 libraries (18 libraries failed for sequencing) were sequenced, and 267.5 Gb of cleaned data were generated after trimming using Trimmomatic version 0.36 (Bankevich et al., 2012). Data from each BAC pool were assembled using SPAdes version 3.09 with default parameter. A total of 2,611,145 contigs were assembled with contig N50 of 7.38 kbp (Zhang and Ming, unpublished data).

The AP85-441 BAC library sequences were BLAST searched against the sorghum genome with an *E*-value of $1e^{-4}$. The sequences that had single BLAST hits in the sorghum genome were selected as candidate FISH probes. To identify low copy number BAC clones from the BAC library, sequence-specific primers were designed using Primer 5.0 software to enable BAC pools to be screened by PCR specifically. Primer lengths of 18–25 bp, T_m values of 55–65°C, and nucleotide compositions of 40–60% cytosine and guanine were selected. A two-step PCR method was used to screen the 3D dimension pools of the BAC library (Asakawa et al., 1997; Crooijmans et al., 2000) with some modifications. Each clone from a 384-well plate was cultured in 80 μl Lysogeny broth + 34 mg/ml chloramphenicol overnight at 37°C in a 384-well culture plate. In total, eight 384-well plates with a total of 3072 BAC clones were cultured. For each 384-well plate, we first constructed 16 row pools (row A to P) and 24 column pools (column 1–24) for each 384-well plate by mixing equal volumes of culture of each individual clone, and then mixed

the 16 row pools and 24 column pool together with equal volumes from each row and column pool as one plate pool or superpool. In this study, we focused on screening eight 384-well plates. Thus, eight superpools were prepared for the eight plates separately. The positive 384-well plate from the eight plates was identified and subsequently the positive row among the 16 mixtures was screened. Finally, the PCR positive columns were selected among 24 column mixtures. Twenty-four PCR reactions were performed to identify positive clones from the eight 384-well plates.

PCR amplifications were performed as previously described (Bouzidi et al., 2006). Each reaction included rTaq premix 7.5 μl [10× PCR buffer, $MgCl_2$ (25 mM), dNTPs (2.5 mM), rTaq polymerase (3 U/1.5 μl)], 0.6 μl of each primer (10 μM), 1.0 μl bacteria liquid template, to a final volume of 20 μl. Following initial denaturation at 95°C for 5 min, 35 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 1 min were performed. PCR products were analyzed on a 1.5% agarose gel.

Metaphase Chromosome Preparation

Chromosome preparation was performed as previously described (Lou et al., 2010) with minor modifications. In brief, *S. spontaneum* plants were grown in the greenhouse conditions for root tips harvesting. Excised root tips about 1–2 cm were treated with 0.002 mol/l 8-hydroxyquinoline at room temperature for 2–4 h, rinsed in water for 15 min, and fixed in ethanol:acetic acid (3:1) for at least 24 h at room temperature. The root tips were then digested in an enzyme solution (4% cellulose R-10 and 2% pectolyase in 0.1 M citrate buffer) at 37°C for 1 h, washed with ice deionized water for 30 min, and finally incubated in ethanol:acetic acid (3:1) for about 30 min. Slides were prepared according to using the "flame-dried" method (Iovene et al., 2008).

BAC DNA Purification and Probe Labeling

The BAC DNA was extracted with PhasePrep^{TM} BAC DNA Kit according to the manufacturer's manual. Purified BAC DNA was labeled by standard nick translation reaction, including diluted DNase I, 10× nick translation buffer, DNA Polymerase I, dNTPs, biotin-/digoxigenin-labeled dUTPs, and BAC DNA. The mixture was incubated at 15°C for 1.5 h. The cut products were then examined on a 1.5% agarose gel for the presence of a smear between 300 and 500 bp. The obtained probes were stored at −20°C until used.

In Situ Hybridization and Detection

Probe mix and hybridization: First, the probe mixture (50% deionized formamide, 2× SSC, 80 ng digoxygenin-/biotin-labeled DNA, 10% dextran sulfate, >1 μg C_ot-100) was placed in a 90°C hot block for 5 min, then immediately transferred on ice until ready to probe for hybridization. The previously prepared flamed-dried slides were treated with 70% deionized formamide, denatured on a heat block at 80°C for 1.5 min, immediately immersed sequentially in ice cold 70% ethanol, 90% ethanol, and then 100% ethanol each for 5 min, and then air dried on the bench. Finally, denatured probes were added to each slide, and covered with 24 × 32 mm coverslips. Slides were placed in a moist chamber at 37°C overnight.

Probe detection: the coverslips were removed and the slides were washed at room temperature in 2× SSC for 5 min, at 42°C in 2× SSC for 10 min and at room temperature in 1× PBS for 5 min in this sequence. Biotin-labeled probe signals were detected with 2 mg/ml Alexa Fluor 488 streptavidin and digoxygenin-labeled probe signals were detected with 2% anti-digoxigenin-rhodamine from sheep. The antibody cocktail (100 μl TNB buffer, 1 μl Alexa Fluor 488 streptavidin, 1 μl rhodamine anti-dig-sheep) was added to the slides, which were covered with 24 × 32 mm coverslips, incubated for 1 h at 37°C in a moist chamber, and washed at room temperature in 1× PBS three times for 5 min each. Excess liquid was removed and 4′,6-diamidino-2-phenylindole (DAPI) (Sigma, St. Louis, MO, United States) in the antifade solution Vectashield (Vector, Burlingame, CA, United States) was added to counterstain the chromosomes. Images were captured with an Olympus BX63 epifluorescence microscope. FISH signal images were analyzed using the CellSens Dimension software.

RESULTS

Screening Low-Copy BAC Clones

To screen the BAC clones that potentially have low copy number sequences in the *S. spontaneum* genome, we BLAST searched the sequenced AP85-441 BAC libraries against the sorghum genome. A total of 2000 BAC sequences corresponding to 10 sorghum chromosomes with good collinearity and showing sequence specificity were screened. The 2000 AP85-441 BAC sequences were masked by REPEATMASKER against the high-repeat sorghum DNA database and TIGR gramineae database to filter repeat sequences. Finally, 114 BACs distributed on 10 sorghum chromosomes were selected for FISH analysis (Supplementary Table 1) with 7–16 BACs on each of the 10 sorghum chromosome (**Figure 1**). To screen positive BAC clones, 114 specific primers were designed based on the BAC clone sequences (Supplementary Table 2) and 49 positive BAC clones were screened from the BAC libraries (Supplementary Table 3).

BAC-FISH Signal Strength and Distribution

DNA samples isolated from the 49 positive BAC clones were labeled with biotin or digoxigenin. 5S rDNA and 45S rDNA probes were used as control and were hybridized to *S. spontaneum* somatic metaphase chromosomes following the FISH procedure. Due to the correlation between signal strength and the variation of repeated sequence content of the BAC sequences, we used C_ot-100 as a competitor (**Table 1**). BAC clones displaying strong and steady hybridization signals were selected for further FISH analysis.

A total of 64 distinct chromosomes can be observed in the metaphase of the *S. spontaneum*. Each BAC-FISH was performed in four independent experiments (or slides). At least 10 spreads of somatic metaphase chromosomes were analyzed in each slide. The results showed that the 27 specific probes

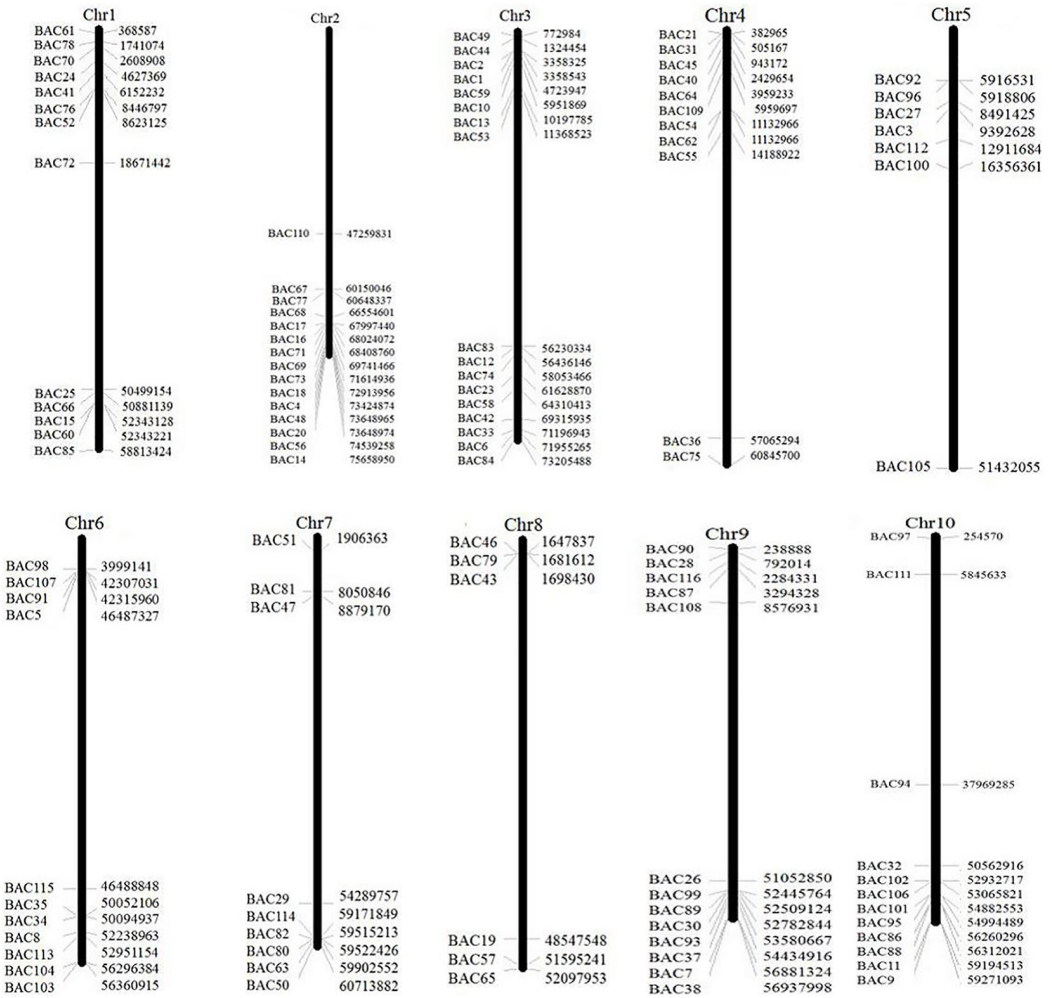

FIGURE 1 | The distributions of 114 BAC clones from AP85-441 in the chromosomes of sorghum. BAC1-114 represents the BAC number, the data on the right are the physical location of BAC in corresponding, region of the sorghum chromosomes. Unit = bp.

TABLE 1 | The classification of 27 positive clones with blocking DNA.

Groups	Blocking DNA	Signal numbers	BAC IDs
I	Not required	7, 8	2, 18, 20, 29, 84
II	25×	4, 6,7	11, 14, 19, 31, 33, 43, 71, 73,74
III	75×	8	26, 66, 76, 78
IV	100×	>8	24
V	150×	1∼6	4, 5, 32, 34, 35, 69, 77, 81

(Supplementary Table 4) can be classified into five groups based on the signal of hybridization (**Table 1**). In Group I (including BAC-18, BAC-20, BAC-29, and BAC-84), probes displayed eight distinctive sites without competitive in *situ* suppression (CISS) using C_ot-100 (**Figure 2A**), suggesting few repeated sequences existed in these BAC sequences. In Group II (including BAC-2, BAC-11, BAC-14, BAC-19, BAC-31, BAC-33, BAC-43, BAC-71, BAC-73, and BAC-74), C_ot-100 was used as competitor for BAC probe hybridization, and six to seven quite distinct sites were displayed in karyotypes of this group (**Figure 2B**). The hybridization result was similar to 45S rDNA with seven signal sites, which may be caused by chromosome structure variation in the homologous chromosome. In Group III (including BAC-26, BAC-66, BAC-76, and BAC-78), C_ot-100 DNA was used as a competitor and eight distinct sites could be observed in the karyotype, suggesting this group of BAC clones have repetitive DNA sequences (**Figure 2C**). In Group IV (including BAC-24), more than eight signal sites were observed in the karyotype and appeared in the pericentromeric and telomeric regions (**Figure 2D**), suggesting these BAC probes had repetitive sequences in *S. spontaneum*. In group V (including BAC-4, BAC-5, BAC-32, BAC-34, BAC-35, BAC-69, BAC-77, and BAC-81), the probes showed dispersed signals in all chromosomes with limited regions of individual chromosomes displaying specific signals (**Figure 2E**). The hybridization result indicated that this group of BAC clones contained large amounts of repetitive DNA in *S. spontaneum*, which may be caused by a rapid evolutionary divergence for the repetitive regions

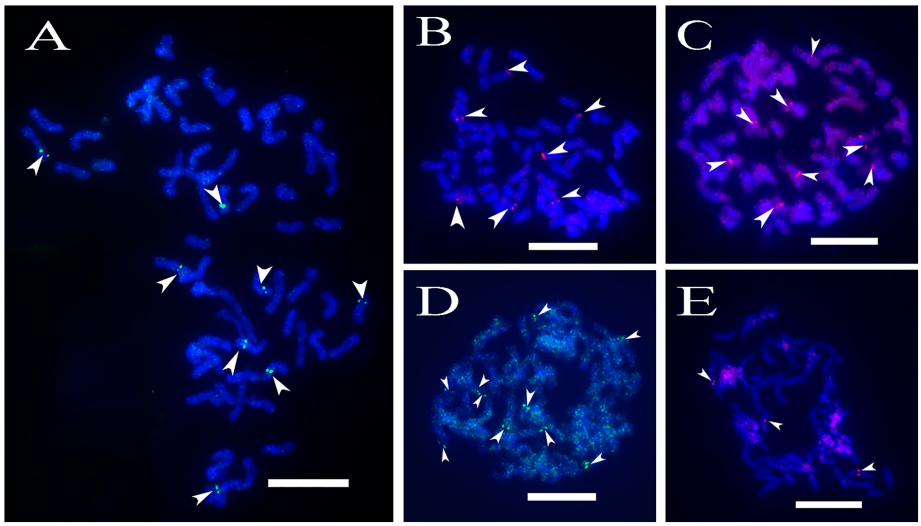

FIGURE 2 | (A–E) The spread of the *S. spontaneum* chromosome hybridized with chromosome-specific BAC clones using BAC-FISH. **(A–E)** show hybridization of BAC probes (**Table 2**) to *S. spontaneum* chromosomes. **(A)** BAC-20, **(B)** BAC-43, **(C)** BAC-66, **(D)** BAC-24, and **(E)** BAC-81 probes labeled with biotin (green) and digoxigenin (red). The arrowheads indicate the signals. Scale bars = 10 μm.

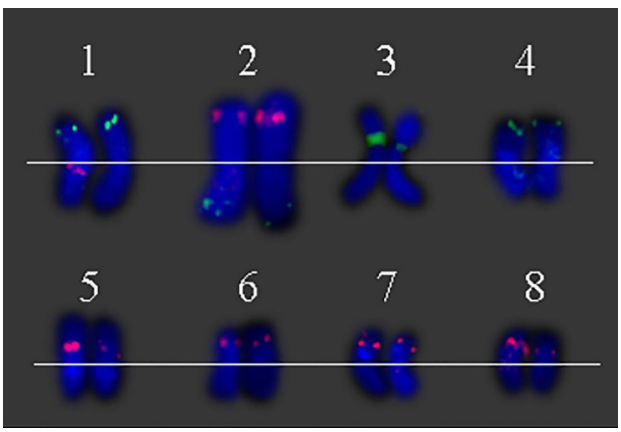

FIGURE 3 | Eight individual *S. spontaneum* chromosomes with chromosome-specific BAC clones. (1) BAC-24 (green) and BAC-66 (red), (2) BAC-14 (green) and BAC-19 (red), (3) BAC-2 (green), (4) BAC-31 (green), (5) BAC-35 (red), (6) 29 (red), (7) BAC-26 (red), and (8) BAC-32 (red) probes labeled with biotin and digoxigenin.

after the split of sorghum and *S. spontaneum*. Therefore, of the 49 BAC probes, a set of 19 BAC probes (Supplementary Table 3) from Groups I, II, and III showed specific signals on *S. spontaneum* chromosomes corresponding to nine of the sorghum chromosomes beside Sb05. The signals of the other 22 probes were dispersed in all chromosomes. We selected a set of 19 BAC probes as *S. spontaneum* chromosome-specific signals (**Figure 3**). Of the 19 BAC probes, one probe was located on Sb02, Sb04, and Sb09; two probes on Sb07, Sb08, and Sb10; three probes on Sb01 and Sb06; and four probes on Sb03. This probe set provided a useful tool for comparative analysis of *S. spontaneum* and sorghum (**Table 2**).

TABLE 2 | The distribution of BAC-FISH probes in chromosomes of *S. bicolor* and *S. spontaneum*.

S. bicolor		*S. spontaneum*	
Chromosome ID	**BAC IDs**	**Chromosome ID**	**BAC IDs**
1	24, 66, 78	1	24*, 66, 78
2	14	2	14*, 19(Sb8)
3	2, 33, 74, 84	3	2*, 33, 74, 84
4	31	4	31*
5	N/A	N/A	N/A
6	5, 34, 35	5	35*, 5, 35
7	29, 81	6	29*, 81
8	19, 43	N/A	N/A
9	26	7	26*
10	11, 32	8	32*, 11

The BAC-FISH results were presented in **Figure 2.*

Construction of a BAC Library Specific for *S. spontaneum* Chromosomes

Sb01 and Sb02 are the two largest sorghum chromosomes. To detect the genome synteny between sorghum and *S. spontaneum*, 10 chromosome-specific BAC clones were selected from Groups I, II, and III. Of these 10 BAC probes, three probes (BAC clone ID: 24, 66, 78) were homologous to Sb01 and seven probes (BAC clone ID: 14, 18, 20, 69, 71, 73, 77) corresponded to Sb02. We utilized dual-color detection of FISH for complementarily labeled BAC pairs and the probes for mapping the chain cytogenetic relationship for the BAC clones. By doing so, we constructed the cytogenetic map of *S. spontaneum* based on the 10 BAC probes.

For three probes corresponding to Sb01, FISH analysis of *S. spontaneum* (SES208) showed that BAC-24 and BAC-78 probes

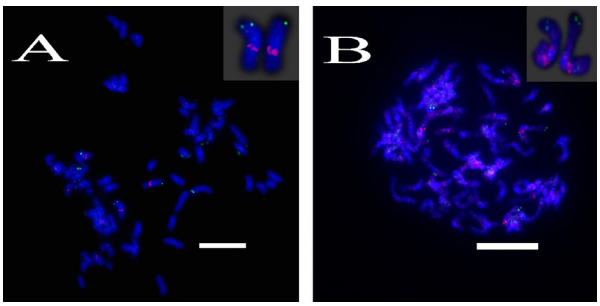

FIGURE 4 | FISH mapping of BAC-24, BAC-66, BAC-78 on SsChr1 in *S. spontaneum*. **(A)** BAC-24 and BAC-66 probes labeled with biotin (green) and digoxigenin (red); **(B)** BAC-78 and BAC-66 probes labeled with biotin (green) and digoxigenin (red). Scale bars = 10 μm.

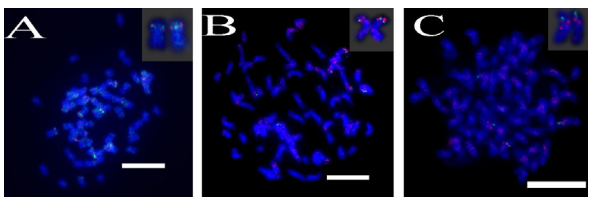

FIGURE 5 | FISH mapping of BAC-69, BAC-71, BAC-73, BAC-77 on SsChr2 in *S. spontaneum*. **(A)** BAC-71 and BAC-69 probes; **(B)** BAC-73 and BAC-69 probes; **(C)** BAC-71 and BAC-77 probes labeled with biotin (green) and digoxigenin (red). Scale bars = 10 μm.

were mainly located on the distal region of *S. spontaneum* chromosome 1, and BAC-66 was mapped in close proximity to the centromeric region (**Figure 4**). BAC-24 and BAC-66 located to the same chromosome (*S. spontaneum* chromosome 1); simultaneously whereas probe 66 and probe 78 also located on the same chromosome, demonstrating that the three BAC probes distributed to the same chromosomes in *S. spontaneum* thus suggesting synteny between *S. spontaneum* and sorghum for chromosome 1. FISH results revealed that the seven BAC probes aligned to Sb02; only probes 69, 71, 73, and 77 of the seven BACs generated intense signals on one *S. spontaneum* chromosome (SsChr2) (**Figure 5**); while probes 14, 18, and 20 were undetectable on *S. spontaneum* chromosome 2 but were observed on other *S. spontaneum* chromosomes. These results demonstrated the chromosome rearrangements occurred in the corresponding Sb02 between *S. spontaneum* and sorghum.

Moreover, the sequence of BAC-2 shared high similarity (97%) to a fragment on chromosome 3 of sorghum. Double color FISH analysis showed that both BAC-2 and 45s rDNA were mapped to the same location of the *S. spontaneum* chromosome (**Figure 6**).

Chromosome Rearrangement on Sb02 between Sorghum and *S. spontaneum*

To further investigate the chromosome rearrangement on *Sb02* between sorghum and *S. spontaneum*, chromosome-specific

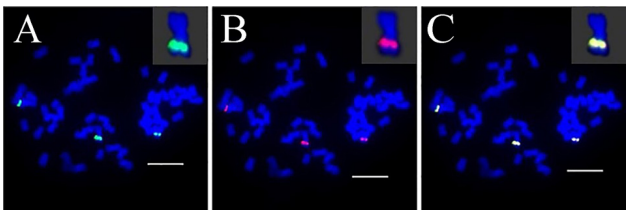

FIGURE 6 | FISH mapping of BAC-2, 45srDNA on SsChr3 in *S. spontaneum*. **(A)** BAC-2 labeled with biotin (green); **(B)** 45s rDNA probes labeled with digoxigenin (red). **(C)** Merged images of **(A)** and **(B)**. Scale bars = 10 μm.

probes corresponding to Sb02 (BAC-14, BAC-18, and BAC-20) and corresponding to Sb08 (BAC-19 and BAC-43) were used for FISH analysis. The results showed that BAC-14 and BAC-19 were mapped to different arms of the same *S. spontaneum* chromosomes, indicating that chromosome rearrangements took place between chromosomes 2 and 8 in *S. spontaneum* and sorghum (**Figure 7**).

DISCUSSION

Saccharum is a complex genus characterized by high polyploidy levels, with small chromosomes (Ha et al., 1999; Dhont, 2005). Distinguishing individual *Saccharum* chromosomes based on their morphology is very challenging. Taking advantage of sequences that are highly covered by the BAC library of *S. spontaneum*, we were able to screen for the potential low copy number BAC clones in *S. spontaneum* using the sorghum genome as a reference. As a high polyploid species, *S. spontaneum* may contain more genome rearrangements than diploids. The differences of the genomes between *S. spontaneum* and sorghum may be the cause of inexactitude in the predictions of the low copy number BAC sequences in *S. spontaneum* genome. In this study, we have initially obtained 114 potential BAC sequences with single-copy based on the sorghum genome. However, only 49 could be identified in the partial BAC libraries, and 27 were found to be chromosome-specific BAC-FISH probes. There are at least two reasons to explain these results. Firstly, the 114 BAC sequences were partial fragments of the BAC insertion due to the limitations of NGS sequence assembly, and may contain repeat sequences in the other regions that have no sequence information. These potentially uncovered repeat sequences could mislead the detection of low-copy sequence through sequence comparison. Secondly, the recent polyploidization in *S. spontaneum* may cause the variation of repeat sequences between *S. spontaneum* and sorghum, and thus result in the nonspecific FISH signal in *S. spontaneum*. In a previous study, sequencing 20 BACs in sugarcane hybrids generated 1.45 Mb contig sequences, the sequences aligning with sorghum genome spanned 0.99 Mb, which accounted for about half of the sugarcane BAC sequences (Wang et al., 2010). Obviously, deletions/insertions existed between sorghum and *Saccharum*, and the utilization of the sorghum genome as reference cannot replace the unique features

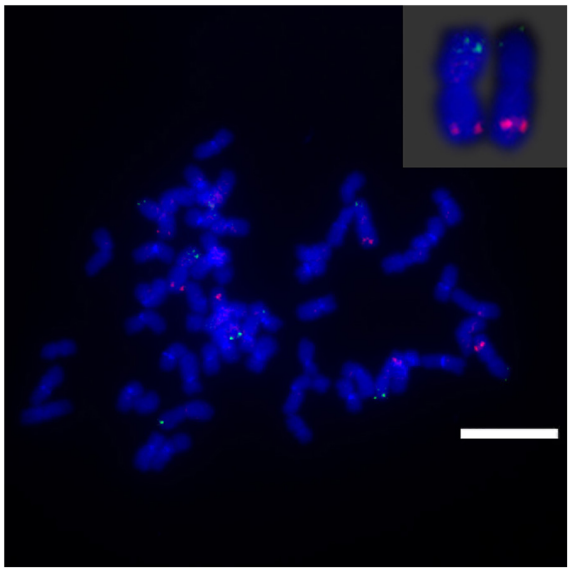

FIGURE 7 | The rearrangement of chromosome in sugarcane and sorghum as revealed by FISH. BAC-14 and BAC-19 probes labeled with biotin (green) and digoxigenin (red). Scale bars = 10 μm.

of the *Saccharum* genome, and thus may cause unpredicted FISH results.

Some FISH slides had strong background noise, which may be due to small regions of repeat sequences (such as SSR sequences) in the BAC clone, which consequently produce interference during hybridization. The sorghum genome has a repeat content of approximately 61% (Paterson et al., 2009), whereas about half of the genome is composed of repeat sequences in *Saccarrum* hybrids (De et al., 2014). In these repeat-rich genomes, it is difficult to develop signal-specific FISH probes, which distinguish chromosomes cytogenetically with similar morphologies. CISS with blocking DNA C_ot-100 can efficiently preclude repeat sequences. In this study, blocking DNA C_ot-100 was used for FISH analysis of the BAC probes in Groups II and III (**Figures 2D–F**), which were verified to be the chromosome-specific BAC probes, while the FISH of BAC probes in Groups IV and V produced strong background interference (**Figures 2D–F**). It is obvious that there are variations of repeat sequences among the examined BAC probes. Therefore, optimization of experiments would be necessary for BAC-FISH analysis in *Saccharum*. Six to seven sites were displayed in Group II, which could be caused by the homologous chromosome structure variations, such as the fragment deletion in one or two homologous chromosomes. In hexaploid wheat, BAC 676D4 hybridized more strongly to the A-genome chromosomes than to the B- and D-genome chromosomes (Zhang et al., 2004). These undetectable of one to two signals in the homologous chromosomes also could be caused by the technical issues of BAC-FISH for such many chromosomes with small size.

In this study, we identified 27 chromosome-specific BAC-FISH probes that correspond to 9 of the 10 sorghum chromosomes (all except Sb05). In a study of the genetic map derived from a cross between *S. officinarum* and sugarcane cultivar, Sb05 was merged with Sb06 in sugarcane (Aitken et al., 2014). In our study, a genetic map based on the F1 population of *S. spontaneum* revealed that Sb05 was divided into two segments, and the two segments were merged with Sb06 and Sb07, respectively (Zhang and Ming, unpublished data). Recently, the genetic map of a member of the *Andropogoneae*, *Miscanthus sinensis*, demonstrated that Sb05 has poor collinearity with the corresponding linkage group in *M. sinensis* (Ma et al., 2012). Similarly, the two ancestral maize chromosomes orthologous to Sb05 retain the smallest number of syntenic orthologs to sorghum genes (Schubert and Lysak, 2011). Therefore, the absence of the BAC-FISH corresponding to Sb05 may indicate chromosome fusion in *S. spontaneum*.

Sorghum and *S. spontaneum* diverged 12 million years ago (MYA), the basic chromosome number was reduced from $x = 10$ to $x = 8$. In Aitken et al. (2014), the HG2 (homologous group 2) of sugarcane aligned to Sb05 and Sb06, and HG8 (homologous group 8) to Sb02 and Sb08, providing evidence for the basic chromosome reduction event (Aitken et al., 2014). In this study, we observed that the sorghum chromosomes Sb02 and Sb08 had interchromosomal rearrangement in *S. spontaneum* as demonstrated by the evidence that BAC-14 aligned to Sb02, and BAC-19 aligned to Sb08 (**Figure 7**). Our study provided the first physical and cytogenetic evidence for the sorghum inter-chromosome rearrangement in *S. spontaneum*. Unpublished data from our group also revealed that Sb08 is divided into two segments, and were merged with of Sb02 and Sb09 in *S. spontaneum*, respectively (Zhang and Ming, unpublished data). The probes corresponding to Sb09, Sb08, Sb05, and Sb06 could be further used for investigating the inter-chromosomal events between sorghum and *S. spontaneum*. *S. spontaneum* has a wide range of ploidy levels ($2n = 40–128$) (Irvine, 1999). These BAC probes could be used to confirm the inter-chromosomal rearrangement of *S. spontaneum* with the different polyploidy levels.

Due to different basic chromosome number between sorghum ($x = 10$) and *S. spontaneum* ($x = 8$), the sorghum chromosomes were not a one-to-one correspondence with *S. spontaneum*. As our genetic mapping study mentioned herein before, Sb08 was divided into two segments which merged with segments of Sb02 and segments of Sb09 in *S. spontaneum*; Sb05 was divided into two segments which merged with segments of Sb06 and segments of Sb07 *S. spontaneum*. Therefore, the probes corresponding to Sb08 and Sb05 were not specific to single *S. spontaneum* chromosomes, whereas the other probes corresponding to the other eight sorghum chromosomes can be used as chromosome-specific cytogenetic BAC-FISH probes for *S. spontaneum*. Thus, the BAC probes tested were based on the eight sorghum chromosomes (**Table 2**) were sufficient for chromosome identification using BAC-FISH. A satellite chromosome was found in one *S. spontaneum* chromosome (chromosome 3) (Ha et al., 1999). Previously, simultaneous FISH revealed that the signals of 45S rDNA were located on the

secondary constrictions of the satellite chromosomes within the chromosome 3 from the anther culture-derived *S. spontaneum* clone (AP85-361) (Ha et al., 1999). In this study, BAC-2 which corresponds to Sb03 and 45s rDNA were mapped to the same location of the *S. spontaneum* chromosome, thus further supporting previous findings (Ha et al., 1999). Our results also provided direct evidence that chromosome 3 of *S. spontaneum* named in the previous study is homologous to Sb03.

CONCLUSION

In this study, we developed chromosome-specific BACs of *S. spontaneum* as a step toward the development of a simple and reproducible method for chromosome identification using BAC-FISH cytogenetic markers, confirming the feasibility of isolating chromosome-specific BACs based on the sorghum genome to construct a physical map of sugarcane. We also provide the first cytogenetic evidence of inter-chromosomal rearrangement between sorghum and *S. spontaneum*. The establishment of the sugarcane BAC-FISH technology system offers new opportunities and the means for sugarcane molecular cytogenetics research, including karyotype analysis, gene localization, and physical map construction. These results are essential for assembly of *S. spontaneum* genome required for whole-genome sequencing.

AUTHOR CONTRIBUTIONS

GD, JS, and JZ conceived the study and designed the experiments. GD, JS, QZ, JW, QY, RM, KW, and JZ carried out the experiments and analyzed the data. GD and JZ wrote the manuscript. All authors read and approved the final paper.

FUNDING

This project was supported by grants from the 863 program (2013AA102604), NSFC (31201260), Program for New Century Excellent Talents in Fujian Province University, and the funding from the Fujian Agriculture and Forestry University.

ACKNOWLEDGMENTS

We thank Dr. Xintan Zhang for assembling the BAC sequences and Dr. Irene Lavagi for editing the English.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00218/full#supplementary-material

REFERENCES

Aitken, K. S., McNeil, M. D., Berkman, P. J., Hermann, S., Kilian, A., Bundock, P. C., et al. (2014). Comparative mapping in the Poaceae family reveals translocations in the complex polyploid genome of sugarcane. *BMC Plant Biol.* 14:190. doi: 10.1186/s12870-014-0190-x

Asakawa, S., Abe, I., Kudoh, Y., Kishi, N., Wang, Y., Kubota, R., et al. (1997). Human BAC library: construction and rapid screening. *Gene* 191, 69–79. doi: 10.1016/S0378-1119(97)00044-9

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Bouzidi, M. F., Franchel, J., Tao, Q., Stormo, K., Mraz, A., Nicolas, P., et al. (2006). A sunflower BAC library suitable for PCR screening and physical mapping of targeted genomic regions. *Theor. Appl. Genet.* 113, 81–89. doi: 10.1007/s00122-006-0274-6

Chen, Y., Zhang, Q., Hu, W., Zhang, X., Wang, L., Hua, X., et al. (2017). Evolution and expression of the fructokinase gene family in *Saccharum*. *BMC Genomics* 18:197. doi: 10.1186/s12864-017-3535-7

Crooijmans, R. P., Vrebalov, J., Dijkhof, R. J., van der Poel, J. J., and Groenen, M. A. (2000). Two-dimensional screening of the Wageningen chicken BAC library. *Mamm. Genome* 11, 360–363. doi: 10.1007/s003350010068

Cuadrado, A., Acevedo, R., de la Espina, M. D. S., Jouve, N., and Torre, C. D. L. (2004). Genome remodelling in three modern *S. officinarum* × *S. spontaneum* sugarcane cultivars. *J. Exp. Bot.* 55, 847–854. doi: 10.1093/jxb/erh093

De, S. N., Monteiro-Vitorello, C. B., Metcalfe, C. J., Cruz, G. M., Del Bem, L. E., Vicentini, R., et al. (2014). Building the sugarcane genome for biotechnology and identifying evolutionary trends. *BMC Genomics* 15:540. doi: 10.1186/1471-2164-15-540

Dhont, A. (2005). Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenet. Genome Res.* 109, 27–33. doi: 10.1159/000082378

Dhont, A., Grivet, L., Feldmann, P., Rao, S., Berding, N., and Glaszmann, J. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Mol. Genet. Genomics* 250, 405–413. doi: 10.1007/BF02174028

Dhont, A., Ison, D., Alix, K., Roux, C., and Glaszmann, J. (1998). Determination of basic chromosome numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes. *Genome* 41, 221–225. doi: 10.1139/g98-023

Dhont, A., Rao, P. S., Feldmann, P., Grivet, L., Islamfaridi, N., Taylor, P. W. J., et al. (1995). Identification and characterisation of sugarcane intergeneric hybrids, *Saccharum officinarum* x *Erianthus arundinaceus*, with molecular markers and DNA in situ hybridisation. *Theor. Appl. Genet.* 91, 320–326. doi: 10.1007/BF00220894

Grivet, L., Daniels, C., Glaszmann, J. C., and D'Hont, A. (2004). A review of recent molecular genetics evidence for sugarcane evolution and domestication. *Ethnobot. Res. Appl.* 2, 9–17. doi: 10.17348/era.2.0.9-17

Ha, S., Moore, P. H., Heinz, D. J., Kato, S., Ohmido, N., and Fukui, K. (1999). Quantitative chromosome map of the polyploid Saccharum spontaneum by multicolor fluorescence in situ hybridization and imaging methods. *Plant Mol. Biol.* 39, 1165–1173. doi: 10.1023/A:1006133804170

Huang, Y., Wu, J., Wang, P., Lin, Y., Fu, C., Deng, Z., et al. (2015). Characterization of chromosome inheritance of the intergeneric BC2 and BC3 progeny between *Saccharum* spp. and *Erianthus arundinaceus*. *PLoS One* 10:e0133722. doi: 10.1371/journal.pone.0133722

Iovene, M., Wielgus, S. M., Simon, P. W., Buell, C. R., and Jiang, J. (2008). Chromatin structure and physical mapping of chromosome 6 of potato and comparative analyses with tomato. *Genetics* 180, 1307–1317. doi: 10.1534/genetics.108.093179

Irvine, J. E. (1999). Saccharum species as horticultural classes. *Theor. Appl. Genet.* 98, 186–194. doi: 10.1007/s001220051057

Jiang, J., and Gill, B. S. (2006). Current status and the future of fluorescence in situ hybridization (FISH) in plant genome research. *Genome* 49, 1057–1068. doi: 10.1139/g06-076

Lam, E., Shine, J., Silva, J. D., Lawton, M., Bonos, S. A., Calvino, M., et al. (2009). Improving sugarcane for biofuel: engineering for an even better feedstock. *Glob. Change Biol. Bioenergy* 1, 251–255. doi: 10.1111/j.1757-1707.2009.01016.x

Lou, Q., Iovene, M., Spooner, D. M., Buell, C. R., and Jiang, J. (2010). Evolution of chromosome 6 of *Solanum species* revealed by comparative fluorescence in situ hybridization mapping. *Chromosoma* 119, 435–442. doi: 10.1007/s00412-010-0269-6

Ma, X. F., Jensen, E., Alexandrov, N., Troukhan, M., Zhang, L., Thomas-Jones, S., et al. (2012). High resolution genetic mapping by genome sequencing reveals genome duplication and tetraploid genetic structure of the diploid *Miscanthus sinensis*. *PLoS One* 7:e33821. doi: 10.1371/journal.pone.0033821

Ming, R., Liu, S., Lin, Y., Silva, J. D., Wilson, W. A., Braga, D. G., et al. (1998). Detailed alignment of Saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. *Genetics* 150, 1663–1682.

Nazeema, J., Laurent, G., Nathalie, C., Olivier, G., Christophe, G. J., Paulo, A., et al. (2007). Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. *Plant J.* 50, 574–585. doi: 10.1111/j.1365-313X.2007.03082.x

Panje, R. R., and Babu, C. N. (1960). Studies in *Saccharum spontaneum* distribution and geographical association of chromosome numbers. *Cytologia* 25, 152–172. doi: 10.1508/cytologia.25.152

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

Piperidis, G., D'Hont, A., and Hogarth, D. M. (2001). "Chromosome composition analysis of various *Saccharum* interspecific hybrids by genomic in situ (hybridisation) (GISH)," in *Proceedings of the International Society of Sugar Cane Technologists, XXIV Congress* (Brisbane, QLD: ASSCT), 565–566.

Piperidis, G., Piperidis, N., and Dhont, A. (2010). Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. *Mol. Genet. Genomics* 284, 65–73. doi: 10.1007/s00438-010-0546-3

Santchurn, D., Ramdoyal, K., Badaloo, M. G. H., and Labuschagne, M. T. (2014). From sugar industry to cane industry: evaluation and simultaneous selection of different types of high biomass canes. *Biomass and Bioenergy* 61, 82–92. doi: 10.1016/j.biombioe.2013.11.023

Schubert, I., and Lysak, M. A. (2011). Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* 27, 207–216. doi: 10.1016/j.tig.2011.03.004

Silva, J. D., Honeycutt, R. J., Burnquist, W. L., Aljanabi, S. M., Sorrells, M. E., Tanksley, S. D., et al. (1995). *Saccharum spontaneum* L. *'SES* 208' genetic linkage map combining RFLP- and PCR-based markers. *Mol. Breed.* 1, 165–179. doi: 10.1007/BF01249701

Wang, J., Roe, B. A., Macmil, S. L., Yu, Q., Murray, J. E., Tang, H., et al. (2010). Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics* 11:261. doi: 10.1186/1471-2164-11-261

Zhang, P., Li, W., Friebe, B., and Gill, B. S. (2004). Simultaneous painting of three genomes in hexaploid wheat by BAC-FISH. *Genome* 47, 979–987. doi: 10.1139/G04-042

Zhang, Q., Hu, W., Zhu, F., Wang, L., Yu, Q., Ming, R., et al. (2016). Structure, phylogeny, allelic haplotypes and expression of sucrose transporter gene families in *Saccharum*. *BMC Genomics* 17:88. doi: 10.1186/s12864-016-2419-6

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

REVIEW
published: 16 February 2018
doi: 10.3389/fpls.2018.00151

Potentials, Challenges, and Genetic and Genomic Resources for Sugarcane Biomass Improvement

*Ramkrishna Kandel[1,2], Xiping Yang[1], Jian Song[1,3] and Jianping Wang[1,4]**

[1] Agronomy Department, University of Florida, Gainesville, FL, United States, [2] Horticultural Sciences Department, University of Florida, Gainesville, FL, United States, [3] College of Life Sciences, Dezhou University, Dezhou, China, [4] FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems, Fujian Agriculture and Forestry University, Fuzhou, China

Lignocellulosic biomass has become an emerging feedstock for second-generation bioethanol production. Sugarcane (*Saccharum* spp. hybrids), a very efficient perennial C4 plant with a high polyploid level and complex genome, is considered a top-notch candidate for biomass production due to its salient features viz. fast growth rate and abilities for high tillering, ratooning, and photosynthesis. Energy cane, an ideal type of sugarcane, has been bred specifically as a biomass crop. In this review, we described (1) biomass potentials of sugarcane and its underlying genetics, (2) challenges associated with biomass improvement such as large and complex genome, narrow gene pool in existing commercial cultivars, long breeding cycle, and non-synchronous flowering, (3) available genetic resources such as germplasm resources, and genomic and cell wall-related databases that facilitate biomass improvement, and (4) mining candidate genes controlling biomass in genomic databases. We extensively reviewed databases for biomass-related genes and their usefulness in biofuel generation. This review provides valuable resources for sugarcane breeders, geneticists, and broad scientific communities involved in bioenergy production.

Keywords: biomass, sugarcane, energy cane, cell wall databases, biomass candidate genes, second-generation biofuel

OPEN ACCESS

Edited by:
Shuizhang Fei,
Iowa State University, United States

Reviewed by:
Yi-Hong Wang,
University of Louisiana at Lafayette, United States
Michael Butterfield,
Monsanto UK Ltd., United Kingdom

****Correspondence:***
Jianping Wang
wangjp@ufl.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: *15 November 2017*
Accepted: *29 January 2018*
Published: *16 February 2018*

Citation:
Kandel R, Yang X, Song J and Wang J (2018) Potentials, Challenges, and Genetic and Genomic Resources for Sugarcane Biomass Improvement. Front. Plant Sci. 9:151. doi: 10.3389/fpls.2018.00151

SUGARCANE, A POTENTIAL BIOFUEL CROP

Sugarcane (*Saccharum* spp.) is a perennial, tropical or subtropical non-cereal grass mainly grown for sugar, contributing to approximately 75% to total global sugar production (Commodity Research Bureau, 2015). It has a close genetic relationship with sorghum (*Sorghum bicolor*) and other members of *Poaceae* family, namely, *Miscanthus* (*Miscanthus x giganteus*) and *Erianthus* (*Erianthus arundinaceus*) (Amalraj and Balasundaram, 2006). As a C4 plant, sugarcane has one of the highest solar energy conversion efficiency and highest biomass yield among the known crops (Henry, 2010; Byrt et al., 2011). Biomass accumulation in sugarcane could reach up to 550 kg/ha/day (Muchow et al., 1994). More recently, sugarcane has been increasingly exploited as a second-generation (i.e., lignocellulosic-based) biofuel feedstock (Lam et al., 2009). Sugarcane had the highest dry biomass yield (39 t/ha/yr), followed by *Miscanthus* (29.6 t/ha/yr), maize (17.6 t/ha/yr), and switchgrass (10.4 t/ha/yr) (Heaton et al., 2008), which could vary depending on the growing season and conditions. The average dry lignocellulosic biomass yield of sugarcane was approximately 22.9 ton/ha/yr (van Der Weijde et al., 2013) with some exceptional genotypes

reaching up to 80–85 ton/ha/yr (Moore et al., 1998) and theoretical potential yield can exceed 100 ton/ha/yr (Jakob et al., 2009; Moore, 2009). Sugarcane has been grown in more than 100 countries in the world with Brazil, India, and China as the top sugarcane producers (FAOSTAT, 2016).

Sugarcane Biomass Potentials

Biomass, an alternative source to fossil resources, offers a promising opportunity for renewable energy (Lynd et al., 2008). Plant biomass, specifically lignocellulose, composed of plant cell walls from grass family, is considered sustainable and renewable feedstock for biofuels (Ragauskas et al., 2006). This concept prompted the establishment of biomass industries across the globe (**Table 1**). Sugarcane is a standout among the bioenergy crops for bioethanol production because of its fast growing and high biomass yielding capacity (Waclawovsky et al., 2010). Sugarcane biomass mainly comes from stalks and straw which respectively constituted 80–85% and 10–15% of total biomass (Carvalho-Netto et al., 2014). Tops, the plant parts between the upper end and the last stalk node with attached green leaves, constituted up to 26% of the total stem weight at harvest (Miocque, 1999).

The oil crisis in the US in the late 1970s spurred the use of sugarcane as an energy plant, (Alexander, 1985; Bischoff et al., 2008). Energy cane, an ideal type of sugarcane showing high biomass yield, was specifically selected for biofuel production (Knoll et al., 2013). Two very distinct traits of energy canes included high number of tillers or stalks per stool and vigorous ratooning ability (Matsuoka and Stolf, 2012). Compared to conventional sugarcane, energy cane hybrids produced 138 and 235% more total biomass (green matter) and fiber, respectively (Matsuoka et al., 2012). With the availability of technologies that convert lignocellulosic biomass into ethanol, the cultivation of energy cane is recently widely increasing (Carvalho-Netto et al., 2014). This emerging biofuel crop is currently being expanded commercially to achieve an annual yield target of one million tons of cane in Florida State alone.

Energy cane has been divided into Type I and Type II physiological types based on its sugar and fiber content (Tew and Cobill, 2008). Type I energy cane contains comparable level of sugar (>13%) but higher fiber content (>17%) than conventional sugarcane. In contrast, Type II energy cane has marginal sugar content (<5%) but very high fiber content (>30%) and is exclusively bred for biomass production. Lignin content in Type I and Type II energy canes was slightly more than that of conventional sugarcane (Knoll et al., 2013). Energy cane fulfills all the requirements for a renewable bioenergy source (Matsuoka et al., 2014). In marginal land of low-fertility where sugarcane cultivation is not profitable, growers may consider growing energy cane for lignocellulosic ethanol production (Sandhu and Gilbert, 2014). Recently, energy cane hybrids of both Type I and Type II varieties are being developed by various private breeding companies in Brazil (Matsuoka et al., 2014). These energy cane varieties can be expanded in geographical range beyond tropical and subtropical regions owing to its wider adaptation and cold tolerance characteristics (Knoll et al., 2013; van Antwerpen et al., 2013).

Sugarcane Biomass Quality

The second-generation bioethanol production not only depends on cellulose content of biomass, the major component for biofuel production, but also on the quality of plant cell wall. Cellulose accounted for about 43–49% of above-ground dry matter in sugarcane and energy cane cultivars (Sanjuan et al., 2001; Kim and Day, 2011), which is comparable to wood (∼45%; Smook, 1992) and more than typical forage grasses (∼30%; Theander and Westerlund, 1993). Plant cell wall is composed of 'complex and dynamic extracellular matrices' that regulate cell growth, provide mechanical support, and protect against pathogens. There are two types of plant cell wall on the bases of the architecture, chemical composition, and biosynthetic processes involved (Carpita, 1996). Primary cell wall is formed by deposition of polysaccharides, predominantly cellulose, hemicellulose, and pectin (Cosgrove, 2005). Secondary cell wall (SCW) is deposited inside primary wall to provide mechanical strength after cells cease to grow and accounts for most of the biomass for biofuel production.

The SCW in sugarcane is composed of mostly cellulose (∼50%), lignin (∼25%), and hemicellulose (∼25%) (Loureiro et al., 2011). Cellulose and hemicellulose serve as the skeletons of plants and are further 'strengthened by lignin and phenolic cross-linkages' (Carpita, 1996). Cellulose and hemicellulose are composed of different carbohydrate polymers and can be converted into fermentable sugars for bioethanol. However, this requires chemical processes such as pretreatment and enzymatic hydrolysis of cellulose, depolymerization, and distillation. Lignocellulosic biomass is recalcitrant to bioethanol conversion, mainly due to lignin and monolignol in cell wall (Weng et al., 2008; Vanholme et al., 2010). Lignification process in sugarcane when studied using histological, biochemical, and transcriptional data obtained from two sugarcane genotypes with contrasting lignin contents, revealed a total of 35 compounds that were related to lignin biosynthesis in sugarcane stems (Bottcher et al., 2013).

Besides composition and content of lignin, composition, structure, and interactions of other polysaccharides in the cell wall play a vital role in the efficient conversion of lignocellulosic biomass to ethanol. Various studies have reported that 'degree of cell wall porosity,' 'cellulose crystallinity,' 'polysaccharide accessible surface area,' and 'protective sheathing of cellulose by hemicellulose' contributed to recalcitrance of cell wall to 'enzymatic degradation' (Himmel et al., 2007; Gross and Chu, 2010; Zhang et al., 2012; Zhao et al., 2012). Understanding the biochemistry of cell wall, genes involved in its biosynthesis, and development of sugarcane genotypes to fulfill the requirements for efficient conversion of biomass to ethanol should be the focus of sugarcane bioethanol production in the future.

Genetic Studies of Sugarcane Biomass

Biomass yield in sugarcane could be improved by an enhanced understanding of underlying genetics of biomass yield components: stalk number (SN), stalk diameter (SD), stalk height (SH), and stalk weight (SW). These components were controlled by genes with additive and non-additive effects and their interactions (Zhou et al., 2009; Carvalho-Netto et al., 2014).

TABLE 1 | Biomass-related databases.

Database	Link	Country	Purpose
Biomass Energy Centre	http://www.biomassenergycentre.org.uk/portal/page?_pageid=73,1&_dad=portal&_schema=PORTAL	United Kingdom	Bioenergy
European Biomass Industry Association	http://www.eubia.org/	European Union	Bioenergy
Louisiana Biomass Resources Database	http://www2.lsuagcenter.com/biomass/about.aspx	United States	Bioenergy
Biomass Power Association	http://www.biomasspowerassociation.com/	United States	Electricity
Sugarcane	http://sugarcane.org/	Brazil	Bioenergy
SAHYOG Project	http://www.sahyog-europa-india.eu/	European Union and India	Bioenergy
BioEnergy Science Center	http://www.bioenergycenter.org/besc/	United States	Cellulosic biofuels
Russian Biofuel Association	http://www.biofuels.ru/	Russia	Bioethanol, biodiesel

Relative contribution of additive variance was the highest for SN among these components. Similarly, high genetic variability and heritability existed in sugarcane for SD, SN, and SW (Sanghera et al., 2014), implying that selection of sugarcane clones for biomass trait is feasible. Further, an attempt was made to identify the quantitative trait loci (QTLs) controlling biomass yield components such as SH, SN, SD, and brix with a population consisting of 295 progeny developed by selfing 'R570' (Hoarau et al., 2002). A total of 40 putative quantitative trait alleles (QTAs) were identified, with each QTA contributing only 3–7% toward total phenotypic variation. Another effort made by Aitken et al. (2004) reported 32 putative QTLs associated with SN, SD, and SW in an F_1 segregating population. Similarly, the phenotypic variations explained by each QTL were very low, ranging from 3 to 9%. Interestingly, 11 of the 32 QTLs identified were associated with more than one trait. Molecular markers linked to biomass yield components have been identified and thus could be used in introgression breeding programs. Recently, an association mapping conducted on 28 genotypes of sugarcane identified a few simple sequence repeat (SSR) markers associated to SW and SN (Bilal et al., 2015). So, numerous biomass yield components of sugarcane can be targeted to enhance biomass production. Recently, gene expression analysis showed that 1,649 and 555 differently expressed (DE) transcripts were revealed between young and mature tissues and between 10 sugarcane genotypes with different level of fiber content, respectively. Of these DE transcripts, 151 and 23, respectively were directly related to fiber and sugar accumulation. In addition, the analysis also found full-length candidate transcripts and pathways that could determine the contrasting fiber accumulation in genotypes with varying content and tissue types (Hoang et al., 2017). The results from gene expression analysis is more reliable than that of molecular marker analysis as it offers the ability to discriminate closely related gene transcripts (Hoang et al., 2017). Thus, biomass yield improvement in sugarcane could be feasible if we could couple the information on molecular markers linked to QTLs controlling biomass yield components with gene expression analysis.

The high tillering ability usually corresponded to an increased number of harvestable stalks and consequent production of high number of favorable ratoons in the following seasons (Matsuoka and Stolf, 2012). Thus, tillering has been considered as a critical biomass trait. Dissecting the genetics of tillering ability based on the information available in other species can also aid the effort in utilization of various genetics approaches for biomass improvement of sugarcane. Four QTLs that control tillering in sorghum were identified (Hart et al., 2001). Importantly, markers associated with SN in sugarcane have been identified, which were 'co-located within or near QTLs that control tillering and rhizomatousness in sorghum' (Jordan et al., 2004). Tillering characteristics in maize was reported to have an incomplete dominance (Rogers, 1950). Two genes g*rassy tiller1 (gt1)* and *teosinte branched1 (tb1)* acted in a common pathway that control tillering in maize (Whipple et al., 2011). A homolog of *tb1* gene in sorghum (Kebrom et al., 2006), *BRANCHED1* (*BRC1*) in *Arabidopsis* controlled formation of axillary buds. Similarly, *MONOCULM 1 (MOC1)*, likely a 'master regulator' of tillering has been isolated and characterized in rice (Li et al., 2003). Over-expression of *tb1* gene in rice reduced SN, though formation of axillary buds was not affected (Takeda et al., 2003). Targeted mutagenesis to *tb1* gene using CRISPR/Cas9 in switchgrass resulted in mutant plants with increased tiller production compared to wild types (Liu et al., 2017). With the homology and gene function conservation across grass species, most likely *tb1* gene would control tillering in sugarcane. Pribil et al. (2007) reported that plants with over-expressed sugarcane *tb1* were significantly taller than untransformed lines. However, SN was not significantly different between transformed and non-transformed lines. An effect of manipulating gibberellins (GA) metabolic pathway in the shoot architecture of sugarcane was also studied by Pribil et al. (2007). The genetically transformed sugarcane lines with over-expression of *GA 2-oxidase*, coding an enzyme that converts GA into non-functional GA, in the cultivar Q117 exhibited variations in height reduction (47 ± 4 cm) and tiller production (5 ± 0.6) relative to control plants (174 ± 21 cm; 1.8 ± 0.9). In contrary, over-expression of *GA 20-oxidase* gene

increased stem elongation and stem weight, while substantially reducing SN. In another effort by Pribil et al. (2007), the data obtained from a total of 31 Q117 transgenic sugarcane lines produced with reduced expression of another branching gene, *MAX3,* involved in strigolactone biosynthesis, indicated that regulation of axillary branching affect plant height in sugarcane. These studies suggested that tillering characteristics in sugarcane could be manipulated by introgressing the genes that control tillering. However, with the complex genomes in sugarcane and species-specific genetic composition, the gene interaction network, dosage effects, and various genetic backgrounds of the recipient clones could remarkably complicate the gene effects after introgression and manipulating process in sugarcane.

CHALLENGES OF SUGARCANE BIOMASS IMPROVEMENT

Biomass yield is a complex concept. Broadly speaking, goals of sugarcane breeders should be to enhance overall biomass yield, biomass quality, and adaptation to wider environment etc. The biomass yield trait could be explained at three levels, and is usually intertwined with biomass quality. At the field level, the biomass trait is dry biomass yield per acre, which is determined by planting density if plant genotype is fixed. At the individual plant level, biomass can be further dissected into SH, SD, SN, and leaf biomass. Thus, selecting genotypes with enhanced SH, SD, SN, and leaf biomass is crucial for higher biomass yield. At the cellular level, cellulose, hemicellulose, and lignin in the cell wall constitute the plant biomass. Increasing the relative cellulose and hemicellulose content as well as balancing the lignin content was vital for increasing biomass yield and enhancing biofuel conversion efficiency (Li et al., 2014). Jung et al. (2013) reported a compromised biomass yield in sugarcane when *caffeic acid O-methyltransferase (COMT)*, a key enzyme in lignin biosynthesis, was suppressed by 91% and lignin content was reduced by 12%. However, 80% suppression of *COMT* and 6% reduction in lignin content made no impact on biomass yield significantly. So, integrating all these different levels of traits in one systematic crop is very challenging because it depends on identifying the genetic basis or components of each specific trait, and balancing those components.

Sugarcane biomass improvement faces additional and specific inherent challenges viz. narrow gene pool in modern sugarcane cultivars, poor synchronization and fertility of flowers in parental clones, long breeding/selection cycle, and genomic complexity (Manickavasagam et al., 2004; Lakshmanan et al., 2005). These issues have been hindering the ability of breeders to efficiently improve biomass traits and thus must be dealt with selection of parents with wide genetic variability, synchronous flowering and cross-fertility, coupled with molecular markers to improve the efficiency of genotype selection.

Narrow Genetic Bases of Current Cultivars

Modern sugarcane cultivars were derived from only a handful of sugarcane clones (Arceneaux, 1967; Roach, 1989) including eight *Saccharum officinarum*, two *S. spontaneum*, one natural hybrid of *S. spontaneum* and *S. officinarum*, and two *S. sinense*. In addition, commercial cultivars were further developed from intercrossing of hybrids and their subsequent backcrosses to *S. officinarum*, called nobilization. These hybrids were repeatedly used in developing modern sugarcane cultivars, which contributed to narrow genetic bases of current sugarcane cultivars. Consequently, sugarcane cultivars became vulnerable to various diseases and insect pests in addition to a diminished genetic gain for both sugar content and biomass yield.

Poor Synchronization and Fertility of Flowers

The synchronization in flowering between clones selected for crossing is very critical in sugarcane breeding programs. Sugarcane clones tend to flower up to 8 weeks apart (Nuss, 1982), and it is especially pronounced between *S. officinarum* and *S. spontaneum* (Moore and Nuss, 1987), thus requiring breeders to artificially induce flowering in an attempt to facilitate cross pollination. This lack of overlap in flowering periods between desired clones could debilitate the breeding programs. Thus, studies have been conducted to synchronize flowering in desired parents through manipulation of the photoperiod (Bull and Glasziou, 1979). In addition, sugarcane flowers have 'low male fertility' and reduced pollen viability at high latitudes (Moore and Nuss, 1987), and in some cases, were self-sterile (Skinner, 1959). Moreover, progeny derived from crosses involving high degree of self-pollination showed decreased viability and vigor (Skinner, 1959; Tew and Pan, 2010). Thus, selection of desired parents that are cross-fertile, yet with wide genetic distance to ensure fertile progeny with broader genetic base is critical in sugarcane improvement.

Long Breeding/Selection Cycle

Hybridization in sugarcane is tedious, time consuming, and requires special skills to perform. Conventional sugarcane breeding takes 10–15 year to create new cultivars because sugarcane has a long growing season of 10–12 months (one generation/year). Basically, sugarcane breeding program involves three basic steps: (i) parental clone selection, (ii) hybridization, and (iii) selection of superior progeny in several vegetatively propagated generations based on their phenotypic performance (11 year). At early generations, selection of superior genotypes is performed for the traits with high heritability, albeit, using low selection intensity. Broad-sense heritability for biomass yield components such as SD, SN, and SH (Sanghera et al., 2014) and overall cane yield was high (0.51–0.84) (Racedo et al., 2016) in sugarcane and thus can be selected for in early generations. At later generation of selection, significantly reduced number of clones will be planted in replications at different environments for performance and thus helps increase the experimental accuracy to screen the traits with low heritability (Gazaffi et al., 2014). Final characterization involves further evaluation of selected genotypes for stability, uniformity, yield, and uniqueness by assessing over several cuts. Superior genotypes are then released as cultivars for commercial production.

Genomic Complexity and Genome Size

Most commercial sugarcane cultivars are interspecific hybrids that have chromosome from 100 to130, with approximately 80% of chromosomes inherited from *S. officinarum*, 10–20% from *S. spontaneum*, and less than 5–17% from recombination between the two species (D'Hont et al., 1996; Piperidis and D'Hont, 2001; Cuadrado et al., 2004). Thus, each locus of sugarcane cultivars has up to 12 alleles (Le Cunff et al., 2008). The somatic cell genome (2C) size of the modern cultivar 'R570' (2n = 115) was approximately 10,000 Mbp (10 Gbp) (D'Hont, 2005) with an average size of 87 Mbp per chromosome, which is larger than the 73 Mbp per chromosome in sorghum (Wang et al., 2010). The monoploid sugarcane genome (750–930 Mbp) is twice the size of rice (389 Mbp), similar to sorghum (760 Mbp), and much smaller than maize (2500 Mbp) (D'Hont and Glaszmann, 2001). Thus, high polyploidy and large genome size pose considerable challenges in sugarcane improvement through QTL identification and marker assisted selection (MAS) (Sreenivasan et al., 1987; Lu et al., 1994; Jannoo et al., 1999).

Allele segregation and inheritance in sugarcane are much more complicated than diploid species. Multiple homologous chromosomes with multi-dose alleles commonly occur in *Saccharum* spp., which complicate the segregation ratio in the crosses and thus required evaluation of thousands of progeny to sort out the segregation of alleles (Matsuoka et al., 2009). In addition, large and complex genome required a large number of molecular markers to sufficiently cover the genome (Gouy et al., 2013). Consequently, development of markers linked with desirable traits is challenging tasks in sugarcane. Furthermore, selection of superior F_1 hybrids with favorable alleles became difficult due to a substantial random sorting of homologous and homoelogous chromosomes and the formation of recombinants (Grivet and Arruda, 2002). Thus, genomic complexity hinders the dissection of biomass traits at the molecular level, complicating sugarcane improvement program through MAS. The current selection of sugarcane genotypes with improved biomass yield mainly relied on visual and labor-intensive field traits measurements.

GENETIC RESOURCES FOR SUGARCANE BIOMASS IMPROVEMENT

Sugarcane Germplasm and Their Utilization

Sugarcane germplasm collection is the potential source of genetic variation for many traits including biomass. For example, *S. spontaneum* possessed wide genetic variability for morphology, ratooning, and tolerance for biotic and abiotic stresses (Aitken and McNeil, 2010; Govindaraj et al., 2014). Modern sugarcane cultivars inherited the tillering and ratooning ability from *S. spontaneum* through hybridization (Matsuoka and Stolf, 2012). In addition, *S. spontaneum* is genetically more diverse than *S. officinarum*, thus contributing to ecological adaptation of sugarcane (Jackson, 1994; Tew and Cobill, 2008), which allowed sugarcane to grow even in marginal land. Sugarcane or energy cane breeding should tap into all the relevant information on genetic variances in the germplasm associated with biomass yield traits to not only improve but also to broaden the genetic base of biomass traits (Todd et al., 2014).

Organized attempts were made to collect genotypes that were highly productive, resistant to diseases, and had high sugar content (Berding and Roach, 1987). International Board of Plant Genetic Resources (IBPGR) and International Society of Sugar Cane Technologists (ISSCT) undertook efforts to collect sugarcane accessions (Anonymous, 1982) and consequently, two duplicated world sugarcane (*Saccharum* spp.) collections are maintained in India and USDA, known collectively as the 'World Collection of Sugarcane and Related Grasses' (WCSRG). The National Plant Repository in Miami, FL, United States maintains over 1000 accessions of *Saccharum* germplasm collected from 45 different countries all over the world (Berding and Roach, 1987; Comstock et al., 1995). The WCSRG contains enormous genetic variability for various morphological traits, biomass yield components, adaption to stresses, and other agronomic or quality traits. The WCSRG provides a repository for many valuable alleles of lignocellulosic biomass traits, which could be targeted to enhance biomass production through energy cane breeding directly or can be utilized for identifying alleles associated with biomass traits for marker development and MAS.

Characterization of germplasm serves as an important bridge linking the collection and utilization phases of germplasm conservation (Heinz, 1987). In efforts to use the WCSRG in breeding program and to broaden the genetic base of sugarcane cultivars, the genetic diversity analysis on partial genotypes in WCSRG was conducted (Tai and Miller, 2002; Brown et al., 2007). The CP 96–1252 was released with a widened germplasm base through introgression program among WCSRG (Miller et al., 2005). In addition, 1002 accessions from WCSRG, presumed to possess valuable alleles for biomass and other agronomic traits (Nayak et al., 2014), were genotyped with SSR markers. A core collection of 300 accessions that represented the genetic diversity of WCSRG was developed according to genotypic data (Nayak et al., 2014). On the other hand, the WCSRG was phenotypically characterized by eight traits and a similar core collection was developed based on morphological traits (Todd et al., 2014). A diversity panel representing the WCSRG were selected by weighing in different parameters from combination of both phenotypic and genotypic data, which, not only serves as an association population to discover the desirable alleles in the future, but also can be utilized in the breeding program for crop improvement as they have been thoroughly evaluated for various traits (Todd et al., 2017).

Sugarcane Genomic Databases

Though sugarcane has a complex genome to decipher, sugarcane geneticists have invested significant efforts to explore and dissect its complex genome using different genomic tools. Genomic databases are critical reservoirs and important foundations for molecular breeders to mine the candidate genes and to facilitate molecular crop

improvement through MAS. Below, we summarized the publicly available genomic databases (**Table 2**), which can be mined and utilized for sugarcane molecular improvement.

SUCEST-FUN Database[1] is a large sugarcane functional genomics database including approximately 237,954 expressed sequence tags (ESTs) from 26 diverse cDNA libraries constructed from different sugarcane varieties with different developmental stages and different tissues and organs (Vettore et al., 2003). The ESTs were further assembled into 42,982 distinct contigs, which had 71 and 82% of contigs significantly matching the *Arabidopsis* and rice genome, respectively. The database webserver integrates transcripts, molecular markers, gene categories, gene expression studies, and data mining tools to provide comprehensive access to sugarcane genomic resources (Nishiyama et al., 2012). This is the most comprehensive web portal for sugarcane genomic resources as it houses not only the sugarcane transcript sequences but also other related databases such as Sugarcane Gene Index (SGI), and Sugarcane Signal Transduction (SUCAST), and Sugarcane Metabolism (SUCAMET) as well.

Sugarcane transcription factor database[2] has a collection of 1,177 predicted sugarcanes (*S. officinarum*) transcription factors (TFs). It is a part of plant transcription factor database (plantTFDB)[3], which in turn catalogs all the genes involved in plant transcriptional activities and provides a repository for 320,370 putative TFs from 165 species (Jin et al., 2017) including sorghum, a close diploid relative of sugarcane, detailing ontology, domain feature, expression pattern, and orthologous groups of genes (Zhang et al., 2011). This database sheds light on interactions between TFs and target genes in order to explore functional mechanisms of TFs.

GRASSIUS[4] is a publicly available web resource that integrated different databases as well as computational and experimental resources related to the control of gene expression in the grasses and associated agronomic traits and also links four databases: GrassTFDB (Grass transcription factor database), GrassCoRegDB (co-regulator database), GrassPROMDB (promoter database), and TFome collection (TF open reading frame) as well. GRASSIUS provides information on TFs from maize, sugarcane, rice, sorghum, and *Brachypodium distachyon* and contains the collection of grass TFome, which provides information on full-length ORFs. GrassPROMDB furnishes the data on promoters and cis-regulatory elements for the aforementioned grass species (Yilmaz et al., 2009). Overall, it contains 9,044 TFs, 579 co-regulators, 149,075 promoter sequences, 2,114 TF ORF clones and 180 TFomes from five grass species. Recently, TFome for maize has been updated with 2,017 unique maize TFs including 24 families of co-regulators (Burdo et al., 2014). So, GRASSIUS especially focuses on regulatory elements and their interactions in grass species and can be utilized as backup sources and cross species comparative genome studies in sugarcane.

TropGENE[5] database is a genetic information system for tropical crops. The most commonly stored information on this database included the genetic resources (agro-morphological traits, parentages, reactions to diseases and drought, and allelic diversity), molecular markers, genetic maps, sequences, genes, QTLs information, physical maps, and corresponding references (Ruiz et al., 2004; Hamelin et al., 2013). It contained about 19,800 molecular markers and 9,500 germplasm entries for 10 tropical crops with their accession number, country of origin, taxonomy, ploidy level, and phenotypic information on agronomic and morphological traits (Hamelin et al., 2013). TropGENE differs from other sugarcane-related databases in that it provides both genetic and phenotypic resources for tropical crops including sugarcane. Thus, a typical agronomic trait can be explored at both molecular and phenotypical levels in this database.

[1] http://sucest-fun.org

[2] http://planttfdb.cbi.pku.edu.cn/index.php?sp=Sof

[3] http://planttfdb.cbi.pku.edu.cn

[4] http://grassius.org/

[5] http://tropgenedb.cirad.fr/tropgene/JSP/index.jsp

TABLE 2 | Publicly available genomic resources and tools for sugarcane and its allied species.

Database	Link	Species	Type
Sugarcane transcription factor database	http://planttfdb.cbi.pku.edu.cn/index.php?sp=Sof	*S. officinarum*	Transcription factor
SUCEST-FUN	http://sucest-fun.org	Sugarcane	EST
TropGENE	http://tropgenedb.cirad.fr/tropgene/JSP/index.jsp	Tropical crops (banana, cocoa, breadfruit, coconut, coffee, cotton, oil palm, rice, rubber tree, sugarcane)	QTLs, genetic and physical maps, Phenotypes, Parentage, allelic diversity
Grassius	http://grassius.org/	*Brachypodium* maize, sugarcane, sorghum, and rice,	Transcription factor
Phytozome	https://phytozome.jgi.doe.gov/pz/portal.html	Eighty-six green plants	Whole genome sequences and annotation
Sorghum transcription factor database	http://planttfdb_v1.cbi.pku.edu.cn:9010/web/index.php?sp=sb	Sorghum	Transcription factor
MOROKOSHI	http://sorghum.riken.jp/morokoshi/Home.html	Sorghum	Transcriptome, FL-cDNA

Phytozome[6] serves as a comparative portal for green plant genomics. It is a centralized platform that provides evolutionary history of plant gene at the sequence level in addition to offering information on gene structure, gene family, genome organization, and functional annotations of complete plant genomes (Goodstein et al., 2012). Sorghum belongs to the same subtribe Saccharine as sugarcane which makes a reliable model because of its small genome (730 Mbp) for functional genomics of sugarcane and other C4 grasses. Besides, its high level of inbreeding and the partitioning of carbon into sugar make it a model for biomass crops like sugarcane (Paterson et al., 2009). About 85% of sorghum genes are orthologous to sugarcane genes thus sorghum genome provides an excellent resource to study sugarcane genome (Wang et al., 2010). Currently, of the 86 sequenced and annotated plant genomes, 52 have been clustered into gene families at 15 evolutionarily significant nodes[6]. In addition to comparative genomics, phytozome also provides information on expression data and proteome of different organisms. It is the most comprehensive database for retrieving green plant genomes.

Cell Wall Composition Databases of Related Species

Because lignocellulose is very recalcitrant to enzymatic degradation, bioenergy researchers should have the knowledge of the genes particularly involved in its biosynthetic pathways so that those genes could be selected or modified to achieve readily degradable biomass (Ekstrom et al., 2014). In quest for efficient conversion of lignocellulose into ethanol, many cell wall-related databases have been developed and updated regularly with new findings on cell wall genomics. These databases will be excellent resources for comparative genomics study in identifying target genes (Saballos, 2013) for biological and genetic studies and for biofuel crop improvement (Yin, 2014). The plant cell wall-related databases[7] were divided into general, species-specific, and family specific databases (reviewed by Cao et al., 2010). We provide brief discussions on these databases as in-depth review for most of the databases is provided previously (Cao et al., 2010).

General Databases Provide Information about Cell Wall-Related Genes and Their Biosynthetic Pathway for Different Species

Cell wall genomics (CWG)[8] was created and maintained by collaborative efforts of scientists at different universities and research institutions. CWG is supported by the NSF Plant Genome Research Program and provides huge resources for plant biologists studying mutants of 'cell wall-related genes' in *Arabidopsis*, rice, maize, and sorghum. Specifically, this database provides the information on cell wall biogenesis pathway, T-DNA insertional mutants, and forward and reverse genetics for insertional mutants. CWG characterizes the cell wall phenotypes of homozygous cell wall mutants of *Arabidopsis* (dicot) and maize (monocot), providing large scale insertional DNA lines for both plant species as well as characterizing the genes associated with architectural assembly of the cell wall. Despite the lack of functional annotation, an estimated 1,000 genes were reported to be involved in biosynthesis of cell wall-related proteins (Yong et al., 2005). CWG provides information on gene families involved in cell wall biogenesis for both monocot (maize) and dicot (*Arabidopsis*) plant species. Six stages of cell wall formation have been outlined including substrate generation, synthases and glycosyl transferases, secretory pathway, wall assembly, wall dynamics, and wall disassembly. Basically, CWG is a complete repository for gene families and their pathways involved in cell wall formation.

Cell wall navigator (CWN) integrates cell wall-related protein families from many plant and non-plant species, allowing comparison of sequences derived from different species. It has four unique features; (1) an adaptive design for organizing complex protein families from many organisms to cover all the known sequences, (2) a flexible architecture to integrate new families rapidly, (3) an automated update and analysis pipeline for maintaining current information, and (4) many visualization and interactive mining tools. It has information for more than 30 gene families comprising more than 5,000 coding genes involved in primary cell wall metabolism. It incorporates sequences from three different resources: *Arabidopsis* and *Oryza sativa* sp. *japonica* from The Institute for Genomic Research (TIGR), the UniProt database, and the EST division of the National Center for Biotechnology Information (NCBI). The organism-unspecific EST search tool allows the comparative genomic study of novel genes in organisms with distinctive cell wall compositions (Girke et al., 2004). Thus, CWN provides information on detailed functional genomic data involved cell wall biosynthesis as opposed to CWG.

Plant cell walls[9] was created and maintained by complex carbohydrate research center (CCRC) at the University of Georgia (UGA). The CCRC in turn was founded in 1985 at UGA to better understand the chemical structure and biological functions of complex carbohydrates. The research was carried out by six independently funded groups that studied various areas including primary cell wall structures, three-dimensional conformations of cell wall components, the interactions and biosynthesis of cell wall components, and functional role of cell wall as a barrier to plant pathogens and source of biofuels. Plant cell walls focuses on cell wall formation with regard to structural, mechanical, and defensive roles mostly at the biochemical level.

Plant database of annotated cell wall genomes contains genome information on annotated genes, gene structures, and protein functions for seven plant genomes (e.g., rice, *Arabidopsis*, sorghum etc.), 12 algal genomes, as well as individual proteins encoded in these genomes. The information on cell wall-related gene families such as carbohydrate active enzyme (CAZy) family, protein family (Pfam) domain information, 3-D protein structures, homology-based functional prediction, phylogenetic trees of CAZy family proteins (133 CAZy), and their subcellular localizations and interactions allows users to conduct comparative genomic analyses of cell wall-related genes

[6]http://www.phytozome.org

[7]http://plantcellwalls.ucdavis.edu/

[8]http://cellwall.genomics.purdue.edu

[9]http://cell.ccrc.uga.edu/~mao/cellwall/main.htm

(Mao et al., 2009). This database analyzes only annotated cell wall-related genes for comparative genomics.

CAZy database[10] is the most comprehensive repository of Carbohydrate-Active enZYmes (CAZymes) (Park et al., 2010), an important class of proteins that synthesizes, modifies, and degrades structural and storage biomass polysaccharides (Cantarel et al., 2009). Thus, knowledge of CAZymes is crucial to biofuel industry (Yin et al., 2012). The database comprised five classes of protein families: glycosyltransferases (GTs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), carbohydrate-binding modules (CBMs), and glycoside hydrolases (GHs). CAZy provides genomic, biochemical, taxonomical, and structural information on many cell wall-related proteins while providing sequence annotation information from other publicly available resources. It contains the regularly updated information on CAZy protein family, incorporation of new family members and their biochemical information obtained from literature. It reports sequence information for about 340,000 CAZymes, which includes 12,700 biochemically characterized CAZymes and 1400 CAZymes with 3D structures (Lombard et al., 2014). Further, CAZymes Annotation Tools (CAT) was developed for systematic annotation of CAZy proteins. CAT utilizes information collected in the CAZy database, analyzes it, and supplements it with information from other databases (Park et al., 2010). As of November 2017, the database contains CAZymes information for 8,436 Bacteria, 283 Archaea, 212 Eukaryota, and 332 Viruses. Basically, CAZymes studies storage biomass polysaccharides that are directly involved in plant biomass formation.

Database for automated carbohydrate active enzyme annotation (dbCAN[11]) is an improvement on CAZy database in a way that it provides an automated and comprehensive annotation of CAZymes in addition to an easy access to sequences, signature domains, alignments, and phylogeny data of CAZyme-related enzyme families (Yin et al., 2012). PlantCAZyme[12] is a web resource built upon dbCAN and is especially dedicated to providing pre-computed sequence and annotation data on CAZymes. It has information on 43,7900 CAZymes of 159 protein families from 35 plants and chlorophyte algae of fully sequenced genomes (Ekstrom et al., 2014).

Species-Specific Databases

Species-specific databases provide information on cell wall-related genes for particular species. Thus, they complement the general database for deeper understanding of cell wall genes for the species (Cao et al., 2010).

MAIZEWALL[13] provides a public repository on 'a bioinformatic analysis and gene expression data' related to 'cell wall biosynthesis and assembly in maize.' It has 735 contigs that have been classified into 174 gene families and which in turn are classified into 19 functional cell wall-related categories based on known gene annotations. Of the 735 contigs, 651 have full set of developmental gene expression data. Gene expression data are easily accessible and are ranked based on their expression level for each organ and internode stage. Maize homologs were obtained based on 100 cell-wall related keywords and BLAST search against the available cell wall-related genes and homology search against ESTs obtained from cell wall-forming TEs in Zinnia (Guillaumie et al., 2007). MAIZEWALL 'allowed alignments of multiple sequence, identification of predicted protein domains, and sub-cellular localizations of target sequences using user-friendly bioinformatics software' (Cao et al., 2010). In addition, it provided the complete bioinformatic information of each gene as well as gene-specific tags and organ specific fingerprint of each cell wall-related gene (Guillaumie et al., 2007).

Wheat GlycosylTransferase Inventory database (GTIdb[14]) has been used for searching exhaustive candidate genes in wheat that play roles in particular biological process. It provides comprehensive analysis of glycosyltransferases (GT), a multi-gene superfamily involved in biosynthesis of cell wall and storage polysaccharides plus glycosylation of various metabolites. Wheat GT sequences were identified based on sequence homology with *Arabidopsis* and rice GT's found in CAZy database. The database is comprised of two sections: the wheat section and the core database. A total of 912,573 wheat ESTs extracted from 220 EST libraries were used to 'characterize 833 contigs and 2,296 singletons into 41 GT families.' The database provides the sequences of GT for wheat, *Arabidopsis*, and rice in a downloadable format. In addition, phylogenetic trees that provide information on each family of GT from all three species are available in PDF format (Sado et al., 2009).

Rice GT database[15] integrates and hosts functional genomic data for putative rice GTs. It displays user-selected functional genomic data on phylogenetic tree that included sequence and mutant lines information, and expression data. In addition, interactive chromosomal map delineating positions of GTs are included. There are 617 putative GT genes that corresponded to 793 transcripts (gene models) in rice. Links are provided to BLAST, CAZy database, Rice Annotation Project Database (RAP-DB), MSU/TIGR rice database, GRAMENE database, and other rice related databases. Of the 33 rice-diverged GT genes that expressed strongly in above-ground, vegetative tissues, 21 were strong candidates for understanding and manipulating cell walls for biofuel production (Cao et al., 2008).

Cell Wall-Related Gene Family Databases

Expansin Central[16] provides information solely on expansin proteins. Expansins, involved in cell growth, cell wall disassembly, cell separation, and cell wall loosening, are plant cell wall proteins (Cho and Kende, 1997a,b; Li et al., 2003). Expansin central details on protein structure, mechanism of action, nomenclature, genes involved in biosynthetic pathway, protocols, phylogenetic tree, and references for expansin genes. Currently, the database contains a total of 226 expansin gene

[10]www.cazy.org

[11]http://csbl.bmb.uga.edu/dbCAN/annotate.php

[12]http://cys.bios.niu.edu/plantcazyme/

[13]http://www.polebio.scsv.ups-tlse.fr/MAIZEWALL

[14]http://wwwappli.nantes.inra.fr:8180/GTIDB/

[15]http://ricephylogenomics.ucdavis.edu/cellwalls/gt/

[16]http://www.personal.psu.edu/fsl/ExpCentral/index.htm

sequences for *Arabidopsis*, rice, maize, tomato, papaya, poplar and many other species.

Xyloglucan endotransglycosylases/hydrolases (XTH World) provides wealth of information related to composition and organization of primary cell wall and its spatial and temporal variability. In addition, it gives the information on how different cell wall microfibrils interact to form the primary cell wall in dicotyledonous plants as well as different genes involved in cell wall biosynthesis in rice, *Arabidopsis*, tomato, and other crops. To avoid the confusion due to contradictory series of nomenclature for essentially the same class of genes or proteins, the unifying nomenclature was proposed to classify a class of genes that encoded a spectrum of biochemical activities under xyloglucan endotransglucosylase/hydrolase (Rose et al., 2002). Xyloglucan binds cellulose non-covalently and also cross-links cellulose microfibrils (McCann et al., 1992). The database focused on standardized nomenclature and systematic identification of genes/proteins that fell under *xyloglucan endotransglucosylase-hydrolases* (*XTH*) gene family (Cosgrove, 2005). In addition, it provides the links to different databases for rice, tomato, and *Arabidopsis*.

Glycoside Hydrolases Database (GHDB) provides information on CAZy family GH16 glycoside hydrolases, including sequences of 260 amino acids that belong to the family. It provides 3D protein structures, functional annotation, phylogenetic trees, multiple sequence alignments of subfamilies: GH16a and GH16b, and homologous subgroups (Strohmeier et al., 2004). In addition, automatic BLAST search was also incorporated into the database in order to provide comprehensive analysis of the stored data (Strohmeier et al., 2004).

In summary, CWG and CWN databases have exclusive information on cell wall biogenesis pathways in general and are easily accessible. CWN provides the comparative study on sequences of protein families from different plant species that are involved in plant cell wall metabolism. 'Plant cell walls' is good resource for the scientific community interested in biofuel potential of cell wall whereas 'plant database of annotated cell wall genomes' is a huge resource for comparative genomic study of cell wall-related genes across plant and non-plant genomes. CAZy database is a resource dedicated to CAZy protein family involved in cell wall synthesis across all kingdoms, such as Bacteria, Archea, Eukayota, and Viruses. It is the most useful cell wall database for bioenergy research as CAZymes are the integral parts of cell wall biosynthesis. The dbCAN along with PlantCAZyme and CAZy database are dedicated to providing information on CAZymes to enhance bioenergy related studies. 'MAIZEWALL' solely delves into the biosynthesis of maize cell wall through transcriptome analysis of different developmental stages of maize. Wheat GTIdb focuses on candidate genes in wheat that play a key role in cell wall formation and storage polysaccharides. Rice GTdb is dedicated in integrating and hosting functional genomic data on GT genes, candidate genes for biofuel traits in rice. Expansin Central mainly focuses on expansin protein. The XTH database provides information primarily on XTH compound and its role in architectural assembly of the primary cell wall. The GHDB is a database that provides functional annotation and multiple sequence alignments of glycoside hydrolase enzymes of CAZy family.

Application of Genomic Databases for Sugarcane Biomass Improvement

In the past decade, sugarcane became an attractive feedstock for second-generation biofuel production. Due to its complex genome structure and genetic inheritance, the genome sequencing progress is slow. In this vein, public genomic databases of related species and database searching tools provide powerful queries to get insight into biomass related genes from *Saccharum* genome before its whole genome sequence information is released.

Search for Biomass-Related Candidate Genes

Genetics and genomics of model species have uncovered many genes underlying the architecture of biomass yield components at individual plant level such as tillering pattern, SH, SN, leaf number and area, and structure and size of reproductive organs (Long et al., 2006; Jahn et al., 2011). Though we have summarized the sugarcane genomic databases and cell wall related databases, the plant architecture related database is currently not available yet. To retrieve plant architecture genes in sugarcane genome, the first step is to identify the candidate genes to form a candidate gene pool. Keywords defined based on relevant literature description of genes involved in plant architecture, such as tillering, vegetative growth, flowering time, leaf morphology, and secondary xylem and tracheary element differentiation can be used to search the published literature related to characterization of genes associated with biomass production. After evaluating the evidence presented in the paper, the gene sequences can be downloaded from the sources provided to form a plant architecture gene pool. Then the summarized sugarcane genomic databases can be searched through sequence blasting. The first databases to be searched can be the updated genomic sequences (CDS or protein sequences of the annotated gene models) of *Sorghum bicolor*, the closest species to sugarcane with a complete genome sequence from Phytozome database. The top hits are basically the corresponding nucleotide and protein sequences of the candidate genes in sorghum genome, which can then be BLASTed against the available sugarcane EST databases (**Table 2**) to retrieve the sugarcane nucleotide sequences.

Besides the genes involved in the plant architecture, genes related to cell wall biogenesis are important factors controlling biomass. Although many genes putatively involved in different aspects of cell wall biogenesis have been identified in a variety of model plants, relatively few genes contributing to biomass have been explicitly identified in *Sorghum bicolor*. Plant cell wall related gene databases can be searched. For example, the CWN, CWG, and MAIZEWALL databases classify cell wall-related genes into different categories: substrate generation, polymer synthesis, secretion, assembly, rearrangement during development, and disassembly (Girke et al., 2004; Penning et al., 2009). These databases give us an inventory of the genes that could become possible targets in the production of biomass. In order to obtain *Sorghum bicolor* homologs, *Arabidopsis* (6093), Rice (2002)

and Maize (734) cell wall genes can be combined and used to BLAST search for their corresponding coding sequences in sorghum genome, then the transcript sequence in sugarcane through blasting the sugarcane related genomic databases (**Table 2**).

In sugarcane, a huge number of ESTs contain characterized candidate genes involved in important agronomic traits such as sucrose accumulation, biomass yield, and plant architecture etc. (Souza et al., 2001; Kido et al., 2012). Gene expression profiling database allows mining of large number of genes associated with biomass traits. For example, the sugar metabolism related genes have been assessed by transcriptome analysis to reveal the regulation of metabolic enzymes and sugar transporters in sugarcane stem (Casu et al., 2003, 2004, 2007; Watt et al., 2005). Cellulose synthase (*Ces*A) and cellulose synthase-like (*Cs*l) families were identified from 119 differentially expressed genes and further characterized in sugarcane (Casu et al., 2007). In two genotypes IACSP04-065 and IACSP04-627 with different lignin content, more than 2,000 transcripts along with genes that control lignin biosynthetic pathway showed differential expression, which can help us identify genes from the lignin biosynthesis and its interactions (Vicentini et al., 2015). The expression profile was analyzed between two genotypes contrasting for lignin content which showed that transcription factor ShMYB58/63 was correlated with ratio of Syringyl (S) and Guaiacyl (G) lignin substructures and interaction between ShMYB58/63 and ShF5H (Santos Brito et al., 2015). In addition, the EST database has proven to be a useful resource to discover sequence polymorphism in three genes of alcohol dehydrogenases (*Adh*) family (Grivet et al., 2003).

With the candidate gene pool, after exploring all the related databases, candidate gene association analysis can be conducted to identify alleles contributing to sugarcane biomass in a large sugarcane germplasm diversity panel with biomass traits and candidate gene sequence variations. Markers associated with biomass traits can be developed from the association analysis. MAS comes in handy especially to improve crops such as sugarcane that is propagated vegetatively and takes many years of selection for varietal development. QTLs for biomass traits can also be interrelated by the candidate genes in the QTL intervals. A substantial progress has been made to identify molecular markers linked to key biomass-related traits. Molecular markers linked to QTLs for biomass traits such as SD, SW, SN, and SH have been identified in prior studies (Hoarau et al., 2002; Aitken et al., 2004; Bilal et al., 2015). These molecular markers if validated could be utilized to select seedlings that possess QTLs controlling biomass yield traits. Selection of genotypes in seedling stage speeds up the breeding cycle and genetic gain. Besides, selection for these traits could be carried out in early generations because of their high heritability. An incorporation of desirable alleles from diverse germplasm into elite cultivars through MAS leads to improved genetic gain in the breeding programs. So, future studies should focus on molecular markers utilization, targeted mutagenesis, and gene expression analysis for introgression of genes that control biomass yield.

Modification of Biomass-Related Candidate Genes

Breeding endeavors in the future should focus not only on the high biomass yield of sugarcane, but also for high quality of biomass. Sugarcane biomass composition has been genetically modified to increase cellulose and hemicellulose content while balancing the lignin content for enhanced biofuel conversion efficiency (Li et al., 2014). *Cinnamyl alcohol dehydrogenase (CAD)* and *COMT* are two key enzymes involved in lignin synthesis. Plant growth and development were not affected when these enzymes were manipulated. However, doing so would change the quality and composition in cell wall (Saathoff et al., 2011). Additionally, transgenic sugarcanes produced increased sucrose and fiber contents in immature internodes, when activities of *pyrophosphate: fructose 6-phosphate 1-phosphotransferase (PFP)* were down-regulated (Groenewald and Botha, 2008; Van der Merwe et al., 2010). Transgenic sugarcanes with bacterial isomerase gene had a doubled sugar content, as well as 'increased photosynthesis, sucrose transport and sink strength' (Wu and Birch, 2007). Recently, engineering of lignin biosynthesis pathway genes by modulating lignin content has been a strategy to reduce the costs of enzymatic digestion of cellulosic biomass and improve cell wall digestibility. In fact, down-regulation of the *COMT* gene in sugarcane using RNA interference has shown decreased lignin content by 3.9–13.7% and thus required less enzyme and hydrolysis time to generate more fermentable sugar than control (Jung et al., 2012, 2013). Further, reduced cell wall lignin content improved enzymatic digestibility in sugarcane (Jung et al., 2012), maize (Park et al., 2012), and switchgrass (Fu et al., 2011; Saathoff et al., 2011; Yee et al., 2012). Though so much of the focus has been in down-regulating *COMT* or *CAD* genes in sugarcane, sorghum, maize, and switchgrass in order to reduce the lignin content, the improvement of sugarcane genotypes with improved lignocellulosic biomass quality is still at its infancy. As sugarcane has highly complex and polyploid genome, targeted mutagenesis using CRISPR/Cas9 could be a valuable tool to characterize target genes and sort out desirable genotypes. In addition, gene expression analysis could enhance the reliability of genes controlling biomass yield components.

CONCLUSION

Sugarcane has a significant potential as a biomass crop due to its highly efficient photosynthetic rate, high tillering, and ratooning abilities. More recently, newly developed energy cane cultivars have higher fiber content and biomass yield than conventional sugarcane cultivars, specifically at marginal land, thus produce more second-generation biofuel.. However, most of the sugarcane and energy cane cultivars are hybrids developed from interspecific crosses of *S. spontaneum*, and *S. officinarum* with large and complex genome, which obscures molecular and genetic studies for crop improvement. In addition, narrow gene pool, non-synchronous and poor fertility of flowers

among desired parents, and long breeding cycle bottleneck the efficient crop improvement for various economically important traits. Despite the challenges in sugarcane breeding, many genetic resources and genomic databases are available for the sugarcane biomass improvement at molecular level. Specifically, cell wall-related databases offer comprehensive information on biomass-related genes. Dissecting genes involved in biosynthesis of biomass polysaccharides help us better understand the biosynthetic pathways underlying primary and secondary cell wall synthesis, which will be helpful to improve the quality and yield of sugarcane biomass. The available genomic databases are valuable sources to aid studies for genetic improvement of sugarcane biomass quality and yield as the genetic analysis tools for polyploid become available. This review should be helpful for the scientists working on sugarcane biomass improvement through biological, genetic, and genomic approaches.

AUTHOR CONTRIBUTIONS

JW conceived the topic and outline and critically revised the manuscript. RK prepared the manuscript draft. XY and JS contributed critical components to the draft. All authors reviewed the manuscript.

ACKNOWLEDGMENTS

This research was financially supported by the Office of Science (BER), United States Department of Energy, Plant Feedstock Genomics project, Florida Sugar Cane League, and the United States of Department of Agriculture, National Institute of Food and Agriculture, Hatch Project 1011664. Publication of this article was funded in part by the University of Florida Open Access Publishing Fund.

REFERENCES

Aitken, K., and McNeil, M. (2010). "Diversity analysis," in *Genetics, Genomics and Breeding of Sugarcane*, eds R. Henry and C. Kole (Enfield, NH: Science Publishers), 19–42.

Aitken, K. S., Jackson, P. A., Piperidis, G., and McIntyre, C. L. (2004). "QTL identified for yield components in a cross between a sugarcane (*Saccharum* spp.) cultivar Q165^{A} and a S. officinarum clone IJ76-514," in *Proceedings of the Australian Agronomy Conference, 12*th *AAC, 4*th *ICSC*, Wagga Wagga, NSW.

Alexander, A. G. (1985). *The Energy Cane Alternative*. Amsterdam: Elsevier Science Publishers BV.

Amalraj, A. V., and Balasundaram, N. (2006). On the taxonomy of the members of '*Saccharum* Complex'. *Genet. Resour. Crop Evol.* 53, 35–41. doi: 10.1007/s10722-004-0581-1

Anonymous (1982). *Genetic Resources of Sugarcane. International Board for Plant Genetic Resources Working Group on the Genetic Recources of Sugar Cane*. Rome: IBPGR Secretariat, 1–19.

Arceneaux, G. (1967). Cultivated sugarcanes of the world and their botanical derivation. *Proc. Int. Soc. Sugarcane Technol.* 12, 844–845.

Berding, N., and Roach, B.T. (1987). "Germplasm collection, maintenance, and use," in *Developments in Crop Science II. Sugarcane Improvement through Breeding*, ed. D. J. Heinz (Amsterdam: Elsevier), 143–210. doi: 10.1016/B978-0-444-42769-4.50009-6

Bilal, M., Saeed, M., Nasir, I. A., Tabassum, B., Zameer, M., Khan, A., et al. (2015). Association mapping of cane weight and tillers per plant in sugarcane. *Biotechnol. Equip.* 29, 617–623. doi: 10.1080/13102818.2015.1008203

Bischoff, K. P., Gravois, K. A., Eagan, T. E., Hoy, J. W., Kimbeng, C. A., LaBorde, C. M., et al. (2008). Registration of "L79-1002" sugarcane. *J. Plant Regist.* 2, 211-217. doi: 10.3198/jpr2007.12.0673crc

Bottcher, A., Cesarino, I., Santos, A. B., Vicentini, R., Mayer, J. L., Vanholme, R., et al. (2013). Lignification in sugarcane: biochemical characterization, gene discovery, and expression analysis in two genotypes contrasting for lignin content. *Plant Physiol.* 163, 1539–1557. doi: 10.1104/pp.113.225250

Brown, J. S., Schnell, R., Power, E., Douglas, S. L., and Kuhn, D. N. (2007). Analysis of clonal germplasm from five *Saccharum* species: *S. barberi*, *S. robustum*, *S. officinarum*, *S. sinense* and *S. spontaneum*. A study of inter- and intra-species relationships using microsatellite markers. *Genet. Resour. Crop Evol.* 54, 627–648. doi: 10.1007/s10722-006-0035-z

Bull, T. A., and Glasziou, K. T. (1979). "Sugarcane," in *Australian Field Crops: Tropical Celreals, OIlseeds, Grain Legumes and Other Crops*, eds J. V. Lovett and A. Lazenby (Sydney, NSW: Angus and Robertson Publishers), 95–113.

Burdo, B., Gray, J., Goetting-Minesky, M. P., Wittler, B., Hunt, M., Li, T., et al. (2014). The Maize TFome – development of a transcription factor open reading frame collection for functional genomics. *Plant J.* 80, 356–366. doi: 10.1111/tpj.12623

Byrt, C. S., Grof, C. P., and Furbank, R. T. (2011). C4 plants as biofuel feedstocks: Optimising biomass production and feedstock quality from a lignocellulosic perspective. *J. Integr. Plant Biol.* 53, 120–135. doi: 10.1111/j.1744-7909.2010.01023.x

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* 37, D233–D238. doi: 10.1093/nar/gkn663

Cao, P., Jung, K.-H., and Ronald, P. C. (2010). A survey of databases for analysis of plant cell wall-related enzymes. *BioEnergy Res.* 3, 108–114. doi: 10.1007/s12155-010-9082-6

Cao, P. J., Bartley, L. E., Jung, K. H., and Ronald, P. C. (2008). Construction of a rice glycosyltransferase phylogenomic database and identification of rice-diverged glycosyltransferases. *Mol. Plant* 1, 858–877. doi: 10.1093/mp/ssn052

Carpita, N. C. (1996). Structure and biogenesis of the cell walls of grasses. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 47, 445–476. doi: 10.1146/annurev.arplant.47.1.445

Carvalho-Netto, O. V., Bressiani, J. A., Soriano, H. L., Fiori, C. S., Santos, J. M., Barbosa, G. V. S., et al. (2014). The potential of the energy cane as the main biomass crop for the cellulosic industry. *Chem. Biol. Technol. Agric.* 1, 1–8. doi: 10.1186/s40538-014-0020-2

Casu, R. E., Dimmock, C. M., Chapman, S. C., Grof, C. P., McIntyre, C. L., Bonnett, G. D., et al. (2004). Identification of differentially expressed transcripts from maturing stem of sugarcane by in silico analysis of stem expressed sequence tags and gene expression profiling. *Plant Mol. Biol.* 54, 503–517. doi: 10.1023/B:PLAN.0000038255.96128.41

Casu, R. E., Grof, C. P., Rae, A. L., McIntyre, C. L., Dimmock, C. M., and Manners, J. M. (2003). Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis. *Plant Mol. Biol.* 52, 371–386. doi: 10.1023/A:1023957214644

Casu, R. E., Jarmey, J. M., Bonnett, G. D., and Manners, J. M. (2007). Identification of transcripts associated with cell wall metabolism and development in the stem of sugarcane by Affymetrix GeneChip Sugarcane Genome Array expression profiling. *Funct. Integr Genomics* 7, 153–167. doi: 10.1007/s10142-006-0038-z

Cho, H. T., and Kende, H. (1997a). Expansins and internodal growth of deepwater Rice. *Plant Physiol.* 113, 1145–1151.

Cho, H. T., and Kende, H. (1997b). Expression of expansin genes is correlated with growth in deepwater rice. *Plant Cell* 9, 1661–1671.

Commodity Research Bureau (2015). *The 2015 CRB Commodity Yearbook*. Chicago, IL: Commodity Research Bureau.

Comstock J., Schnellt, R., and Miller, J. (1995). "Current status of the world sugarcane germplasm collection in Florida," in *Sugarcane Germplasm Conservation and Exchange*, eds B. J. Croft, C. M. Piggin, E. S. Wallis, and D. M. Hogarth (Canberra, ACT: ACIAR Proceedings), 17–18.

Cosgrove, D. J. (2005). Growth of the plant cell wall. *Nat. Rev. Mol. Cell Biol.* 6, 850–861. doi: 10.1038/nrm1746

Cuadrado, A., Acevedo, R., Moreno Diaz de la Espina, S., Jouve, N., and de la Torre, C. (2004). Genome remodelling in three modern S. *officinarum* x *S. spontaneum* sugarcane cultivars. *J. Exp. Bot.* 55, 847–854. doi: 10.1093/jxb/erh093

D'Hont, A. (2005). Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. *Cytogenet. Genome Res.* 109, 27–33. doi: 10.1159/000082378

D'Hont, A., and Glaszmann, J. C. (2001). Sugarcane genome analysis with molecular markers, a first decade of research. *Proc. Int. Soc. Sugarcane Technol.* 24, 556–559.

D'Hont, A., Grivet, L., Feldmann, P., Rao, P. S., Berding, N., and Glauszmann, J. C. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Mol. Gen. Genet.* 250, 405–413. doi: 10.1007/BF02174028

Ekstrom, A., Taujale, R., McGinn, N., and Yin, Y. (2014). PlantCAZyme: A database for plant carbohydrate-active enzymes. *Database (Oxford)* 2014:bau079. doi: 10.1093/database/bau079

FAOSTAT (2016). *FAOSTAT. Food and Agriculture Organization of the United Nations*. Rome: FAOSTAT.

Fu, C., Mielenz, J. R., Xiao, X., Ge, Y., Hamilton, C. Y., Rodriguez, M. Jr., et al. (2011). Genetic manipulation of lignin reduces recalcitrance and improves ethanol production from switchgrass. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3803–3808. doi: 10.1073/pnas.1100310108

Gazaffi, R., Oliveira, K. M., Souza, A. P., and Garcia, A. A. F. (2014). "Sugarcane: breeding methods and genetic mapping," in *Sugarcane Bioethanol R &D for Productivity and Sustainability*, ed. L. A. B. Cortez (Sao Paulo: Blucher).

Girke, T., Lauricha, J., Tran, H., Keegstra, K., and Raikhel, N. (2004). The cell wall navigator database. A systems-based approach to organism-unrestricted mining of protein families involved in cell wall metabolism. *Plant Physiol.* 136, 3003–3008. doi: 10.1104/pp.104.049965

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Gouy, M., Rousselle, Y., Bastianelli, D., Lecomte, P., Bonnal, L., Roques, D., et al. (2013). Experimental assessment of the accuracy of genomic selection in sugarcane. *Theor. Appl. Genet.* 126, 2575–2586. doi: 10.1007/s00122-013-2156-z

Govindaraj, P., Amalraj, V. A., Mohanraj, K., and Nair, N. V. (2014). Collection, characterization and phenotypic diversity of *Saccharum spontaneum* L. from arid and semi arid zones of Northwestern India. *Sugar Tech* 16, 36–43. doi: 10.1007/s12355-013-0255-4

Grivet, L., and Arruda, P. (2002). Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr. Opin. Plant Biol.* 5, 122–127. doi: 10.1016/S1369-5266(02)00234-0

Grivet, L., Glaszmann, J., Vincentz, M., Da Silva, F., and Arruda, P. (2003). ESTs as a source for sequence polymorphism discovery in sugarcane: example of the *Adh* genes. *Theor. Appl. Genet.* 106, 190–197. doi: 10.1007/s00122-002-1075-1

Groenewald, J. H., and Botha, F. C. (2008). Down-regulation of pyrophosphate: fructose 6-phosphate 1-phosphotransferase (PFP) activity in sugarcane enhances sucrose accumulation in immature internodes. *Transgenic Res.*17, 85–92. doi: 10.1007/s11248-007-9079-x

Gross, A. S., and Chu, J. W. (2010). On the molecular origins of biomass recalcitrance: the interaction network and solvation structures of cellulose microfibrils. *J. Phys. Chem. B* 114, 13333–13341. doi: 10.1021/jp106452m

Guillaumie, S., San-Clemente, H., Deswarte, C., Martinez, Y., Lapierre, C., Murigneux, A., et al. (2007). MAIZEWALL. Database and developmental gene expression profiling of cell wall biosynthesis and assembly in maize. *Plant Physiol.* 143, 339–363. doi: 10.1104/pp.106.086405

Hamelin, C., Sempere, G., Jouffe, V., and Ruiz, M. (2013). TropGeneDB, the multi-tropical crop information system updated and extended. *Nucleic Acids Res.* 41, D1172–D1175. doi: 10.1093/nar/gks1105

Hart, G. E., Schertz, K. F., Peng, Y., and Syed, N. H. (2001). Genetic mapping of *Sorghum bicolor* (L.) Moench QTLs that control variation in tillering and other morphological characters. *Theor. Appl. Genet.* 103, 1232–1242. doi: 10.1007/s001220100582

Heaton, E. A., Dohleman, F. G., and Long, S. P. (2008). Meeting US biofuel goals with less land: the potential of Miscanthus. *Glob. Change Biol.* 14, 2000–2014. doi: 10.1111/j.1365-2486.2008.01662.x

Heinz, D. J. (1987). *Sugarcane Improvement Through Breeding*, Amsterdam: Elsevier.

Henry, R. J. (2010). Evaluation of plant biomass resources available for replacement of fossil oil. *Plant Biotechnol. J.* 8, 288–293. doi: 10.1111/j.1467-7652.2009.00482.x

Himmel, M. E., Ding, S. Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W., et al. (2007). Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* 315, 804–807. doi: 10.1126/science.1137016

Hoang, N. V., Furtado, A., O'keeffe, A. J., Botha, F. C., and Henry, R. J. (2017). Association of gene expression with biomass content and composition in sugarcane. *PLOS ONE* 12:e0183417. doi: 10.1371/journal.pone.0183417

Hoarau, J. Y., Grivet, L., Offmann, B., Raboin, L. M., Diorflar, J. P., Payet, J., et al. (2002). Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). II. Detection of QTLs for yield components. *Theor. Appl. Genet.* 105, 1027–1037. doi: 10.1007/s00122-002-1047-5

Jackson, P. (1994). Genetic relationships between attributes in sugarcane clones closely related to *Saccharum spontaneum*. *Euphytica* 79, 101–108. doi: 10.1007/BF00023581

Jahn, C. E., Mckay, J. K., Mauleon, R., Stephens, J., McNally, K. L., Bush, D. R., et al. (2011). Genetic variation in biomass traits among 20 diverse rice varieties. *Plant Physiol.* 155, 157–168. doi: 10.1104/pp.110.165654

Jakob, K., Zhou, F., and Paterson, A. H. (2009). Genetic improvement of C4 grasses as cellulosic biofuel feedstocks. *In Vitro Cell. Dev. Biol. Plant* 45, 291–305. doi: 10.1007/s11627-009-9214-x

Jannoo, N., Grivet, L., Seguin, M., Paulet, F., Domaingue, R., Rao, S. P., et al. (1999). Molecular investigation of the genetic base of sugarcane cultivars. *Theor. Appl. Genet.* 99, 171–184. doi: 10.1007/s001220051222

Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D1040–D1045. doi: 10.1093/nar/gkw982

Jordan, D. R., Casu, R. E., Besse, P., Carroll, B. C., Berding, N., Mclntyre, C., et al. (2004). Markers associated with stalk number and suckering in sugarcane colocate with tillering and rhizomatousness QTLs in Sorghum. *Genome* 5, 988–993. doi: 10.1139/g04-040

Jung, J. H., Fouad, W. M., Vermerris, W., Gallo, M., and Altpeter, F. (2012). RNAi suppression of lignin biosynthesis in sugarcane reduces recalcitrance for biofuel production from lignocellulosic biomass. *Plant Biotechnol. J.* 10, 1067–1076. doi: 10.1111/j.1467-7652.2012.00734.x

Jung, J. H., Vermerris, W., Gallo, M., Fedenko, J. R., Erickson, J. E., and Altpeter, F. (2013). RNA interference suppression of lignin biosynthesis increases fermentable sugar yields for biofuel production from field-grown sugarcane. *Plant Biotechnol. J.* 11, 709–716. doi: 10.1111/pbi.12061

Kebrom, T. H., Burson, B. L., and Finlayson, S. A. (2006). Phytochrome B represses Teosinte Branched1 expression and induces sorghum axillary bud outgrowth in response to light signals. *Plant Physiol.* 3, 1109–1117. doi: 10.1104/pp.105.074856

Kido, E. A., Neto, J. R. C. F., de Oliveira Silva, R. L., Pandolfi, V., Guimaraes, A. C. R., Veiga, D. T., et al. (2012). New insights in the sugarcane transcriptome responding to drought stress as revealed by supersage. *Sci. World J.* 2012:821062. doi: 10.1100/2012/821062

Kim, M., and Day, D. F. (2011). Composition of sugar cane, energy cane, and sweet sorghum suitable for ethanol production at Louisiana sugar mills. *J. Ind. Microbiol. Biotechnol.* 38, 803–807. doi: 10.1007/s10295-010-0812-8

Knoll, J. E., Anderson, W. F., Richard, E. P., Doran-Peterson, J., Baldwin, B., Hale, A. L., et al. (2013). Harvest date effects on biomass quality and ethanol yield of new energycane (*Saccharum* hyb.) genotypes in the Southeast USA. *Biomass Bioenerg.* 56, 147–156. doi: 10.1016/j.biombioe.2013.04.018

Lakshmanan, P., Geijskes, R. J., Aitken, K. S., Grof, C. L. P., Bonnett, G. D., and Smith, G. R. (2005). Sugarcane biotechnology: the challenges and opportunities. *In Vitro Cell. Dev. Biol. Plant* 4, 345–363. doi: 10.1016/j.bjm.2016.10.003

Lam, E., Shine, J., Da Silva, J., Lawton, M., Bonos, S., Calvino, M., et al. (2009). Improving sugarcane for biofuel: engineering for an even better feedstock. *GCB Bioenerg.* 1, 251–255. doi: 10.1111/j.1757-1707.2009.01016.x

Le Cunff, L., Garsmeur, O., Raboin, L. M., Pauquet, J., Telismart, H., Selvi, A., et al. (2008). Diploid/polyploid syntenic shuttle mapping and haplotype-specific

chromosome walking toward a rust resistance gene (*Bru*1) in highly polyploid sugarcane (2n approximately 12x approximately 115). *Genetics* 180, 649–660. doi: 10.1534/genetics.108.091355

Li, Q., Song, J., Peng, S., Wang, J. P., Qu, G. Z., Sederoff, R. R., et al. (2014). Plant biotechnology for lignocellulosic biofuel production. *Plant Biotechnol. J.* 12, 1174–1192. doi: 10.1111/pbi.12273

Li, X., Qian, Q., Zhiming, F., Yonghong, W., and Guosheng, X. (2003). Control of tillering in rice. *Nature* 42, 618–621. doi: 10.1038/nature01518

Liu, Y., Merrick, P., Zhang, Z., Ji, C., Yang, B., and Fei, S. Z. (2017). Targeted mutagenesis in tetraploid switchgrass (*Panicum virgatum* L.) using CRISPR/Cas9. *Plant Biotechnol. J.* 16, 381–393. doi: 10.1111/pbi.12778

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495. doi: 10.1093/nar/gkt1178

Long, S. L., Zhu, X., Naidu, S. L., and Ort, D. R. (2006). Can improvement in photosynthesis increase crop yields? *Plant Cell Environ.* 29, 315–330. doi: 10.1111/j.1364-3040.2005.01493.x

Loureiro, M. E., Barbosa, M. H. P., Lopes, F. J. F., and Silvério, F. O. (2011). "Sugarcane breeding and selection for more efficient biomass conversion in cellulosic ethanol," in *Routes to Cellulosic Ethanol*, eds S. M. Buckeridge and H. G. Goldman (New York, NY: Springer), 199–239.

Lu, Y. H., D'Hont, A., Walker, D. I. T., Rao, P. S., Feldmann, P., and Glaszmann, J. C. (1994). Relationships among ancestral species of sugarcane revealed with RFLP using single copy maize nuclear probes. *Euphytica* 78, 7–18.

Lynd, L. R., Laser, M. S., Bransby, D., Dale, B. E., Davison, B., Hamilton, R., et al. (2008). How biotech can transform biofuels. *Nat. Biotechnol.* 26, 169–172. doi: 10.1038/nbt0208-169

Manickavasagam, M., Ganapathi, A., Anbazhagan, V. R., Sudhakar, B., Selvaraj, N., Vasudevan, A., et al. (2004). *Agrobacterium*-mediated genetic transformation and development of herbicide-resistant sugarcane (*Saccharum* species hybrids) using axillary buds. *Plant Cell. Rep.* 23, 134–143. doi: 10.1007/s00299-004-0794-y

Mao, F. L., Yin, Y. B., Zhou, F. F., Chou, W.-C., Zhou, C., Chen, H., et al. (2009). pDAWG: An integrated database for plant cell wall genes. *Bioenerg. Res.* 2, 209–216. doi: 10.1007/s12155-009-9052-z

Matsuoka, S., Bressiani, J., Maccheroni, W., Fouto, I., Santos, F., Borém, A., et al. (2012). "Sugarcane bioenergy," in *Sugarcane: Bioenergy, Sugar and Ethanol-Technology and Prospects*, F. Santos, A. Borém, and C. Caldas (Viçosa: Suprema), 471–500.

Matsuoka, S., Ferro, J., and Arruda, P. (2009). The Brazilian experience of sugarcane ethanol industry. *In Vitro Cell. Dev. Biol. Plant* 45, 372–381. doi: 10.1007/s11627-009-9220-z

Matsuoka, S., Kennedy, A. J., dos Santos, E. G. D., Tomazela, A. L., and Rubio, L. C. S. (2014). Energy cane: its concept, Development, Characteristics, and Prospects. *Adv. Bot.* 2014:597275. doi: 10.1155/2014/597275

Matsuoka, S., and Stolf, R. (2012). "Sugarcane tillering and ratooning: key factors for a profitable cropping," in *Sugarcane: Production, Cultivation and Uses*, eds J. F. Goncalves and K. D. Correia (Hauppauge, NY: Nova Science Publishers, Inc), 137–157.

McCann, M. C., Wells, B., and Roberts, K. (1992). Complexity in the spatial localization and length distribution of plant cell wall matrix polysaccharides. *J. Microsc.* 166, 123–136. doi: 10.1111/j.1365-2818.1992.tb01511.x

Miller, J. D., Tai, P. Y., Edme, S. J., Comstock, J. C., Glaz, B. S., and Gilbert, R. A. (2005). Basic germplasm utilization in the sugarcane development program at Canal Point, FL, USA. *Int. Soc. Sugar Cane Technol.* 2, 532–536.

Miocque, J. (1999). Evaluation of growth and yield of green matter of sugarcane from Araraquara, SP region. *STAB Sugar Ethanol Byproducts* 17, 45–47.

Moore, P. H. (2009). "Sugarcane biology, yield, and potential for improvement," in *Proceedings of the Workshop BIOEN on Sugarcane Improvement* San Pablo, CA.

Moore, P. H., Botha, F. C., Furbank, R. T., and Grof, C. P. L. (1998). "Potential for overcoming physio-biochemical limits to sucrose accumulation," in *Intensive Sugarcane Production:Meeting the Challenges Beyond 2000*, eds B. A. Keating, and J. R. Wilson (New York, NY: CAB International), 141–155.

Moore, P. H., and Nuss, K. J. (1987). "Flowering and flower synchronization," in *Sugarcane Improvement Through Breeding*, ed D. J. Heinz (Amsterdam: Elsevier), 273–311. doi: 10.1016/B978-0-444-42769-4.50012-6

Muchow, R. C., Spillman, M. F., Wood, A. W., and Keating, B. A. (1994). Radiation interception and biomass accumulation in a sugarcane crop grown under irrigated tropical conditions. *Aust. J. Agric. Res.* 45, 37–49. doi: 10.1071/AR9940037

Nayak, S. N., Song, J., Villa, A., Pathak, B., Ayala-Silva, T., Yang, X., et al. (2014). Promoting utilization of *Saccharum* spp. genetic resources through genetic diversity analysis and core collection construction. *PLOS ONE* 9:110856. doi: 10.1371/journal.pone.0110856

Nishiyama, M. Y., Vicente, F., Lembke, C. G., Sato, P. M., Dal-Bianco, M. L., Fandino, R. A., et al. (2012). The SUCEST-FUN regulatory network database: designing and energy grass. *Int. Sugar J.* 114, 821–826.

Nuss, K. J. (1982). Flowering of sugarcane in a photoperiod house from 1971 to 1981. *Proc. S. Afr. Sugar Technol. Assoc.* 56, 140–142.

Park, B. H., Karpinets, T. V., Syed, M. H., Leuze, M. R., and Uberbacher, E. C. (2010). CAZymes Analysis Toolkit (CAT): Web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology* 20, 1574–1584. doi: 10.1093/glycob/cwq106

Park, S. H., Mei, C., Pauly, M., Ong, R. G., Dale, B. E., Sabzikar, R., et al. (2012). Downregulation of maize cinnamoyl-coenzyme A reductase via RNA interference technology causes brown midrib and improves ammonia fiber expansion-pretreated conversion into fermentable sugars for biofuels. *Crop Sci.* 52, 2687–2701. doi: 10.2135/cropsci2012.04.0253

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

Penning, B. W., Hunter, C. T., Tayengwa, R., Eveland, A. L., Dugard, C. K., Olek, A. T., et al. (2009). Genetic resources for maize cell wall biology. *Plant Physiol.* 151, 1703–1728. doi: 10.1104/pp.109.136804

Piperidis, A., and D'Hont, A. (2001). Chromosome composition analysis of various *Saccharum* interspecific hybrids by genomic in situ hybridization (GISH). *Proc. Int. Soc. Sugarcane Technol.* 24, 556–559.

Pribil, M., Hermann, S. R., Dun, G. D., Karno, X. X., Ngo, C., O'Neill, S. W. et al. (2007). "Altering sugarcane shoot architecture through genetic engineering: prospects for increasing cane and sugar yield," in *Proceedings of the 2007 Conference of the Australian Society of Sugar Cane Technologists held at Cairns*, Cairns, QLD, 251–257.

Racedo, J., Gutierrez, L., Perera, M. F., Ostengo, S., Pardo, E. M., Cuenya, M. I., et al. (2016). Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. *BMC Plant Biol.* 16:142. doi: 10.1186/s12870-016-0829-x

Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cairney, J., Eckert, C. A., et al. (2006). The path forward for biofuels and biomaterials. *Science* 311, 484–489. doi: 10.1126/science.1114736

Roach, B. T. (1989). Origin and improvement of the genetic base of sugarcane. *Proc. Aust. Soc. Sugar Cane Technol.* 11, 34–47.

Rogers, J. S. (1950). The inheritance of inflorescence characters in maize-teosinte hybrids. *Genetics* 35, 541–558.

Rose, J. K., Braam, J., Fry, S. C., and Nishitani, K. (2002). The XTH family of enzymes involved in xyloglucan endotransglucosylation and endohydrolysis: current perspectives and a new unifying nomenclature. *Plant Cell Physiol.* 43, 1421–1435. doi: 10.1093/pcp/pcf171

Ruiz, M., Rouard, M., Raboin, L. M., Lartaud, M., Lagoda, P., and Courtois, B. (2004). TropGENE-DB, a multi-tropical crop information system. *Nucleic Acids Res.* 32, 364–367. doi: 10.1093/nar/gkh105

Saathoff, A. J., Sarath, G., Chow, E. K., Dien, B. S., and Tobias, C. M. (2011). Downregulation of cinnamyl-alcohol dehydrogenase in switchgrass by RNA silencing results in enhanced glucose release after cellulase treatment. *PLOS ONE* 6:16416. doi: 10.1371/journal.pone.0016416

Saballos, A. (2013). "Development and utilization of sorghum as a bioenergy crop," in *Genetic Improvement of Bioenergy Crops*, ed. W. Vermerris (New York, NY: Springer Science and Business Media, LLC), 211–248.

Sado, P. E., Tessier, D., Vasseur, M., Elmorjani, K., Guillon, F., and Saulnier, L. (2009). Integrating genes and phenotype: a Wheat-*Arabidopsis*-rice glycosyltransferase database for candidate gene analyses. *Funct. Integr. Genomics* 9, 43–58. doi: 10.1007/s10142-008-0100-0

Sandhu, H. S., and Gilbert, R. (2014). *Production of Biofuel Crops in Florida: Sugarcane/Energy Cane*. Gainesville, FL: University of Florida.

Sanghera, G. S., Tyagi, V., Kumar, R., Thind, K. S., and Sharma, B. (2014). "Quality parameters and their association with cane yield in sugarcane under subtropical

conditions," in *Proceedings of the National Symposium Crop Improvement Inclusive Sustainable Development held at Punjab Agricultural University*, Ludhiana, 796–798.

Sanjuan, J., Anzaldo, V., Vargas, J., Turrado, J., and Patt, R. (2001). Morphological and chemical composition of pith and fibers from Mexican sugarcane bagasse. *Holz Roh Werkstoff* 59, 447–450. doi: 10.1007/s001070100236

Santos Brito, M., Nobile, P., Bottcher, A., Dos Santos, A., Creste, S., De Landell, M., et al. (2015). Expression profile of sugarcane transcription factor genes involved in lignin biosynthesis. *Trop. Plant Biol.* 8, 19–30. doi: 10.1007/s12042-015-9147-y

Skinner, J. C. (1959). Controlled pollination of sugarcane. *Bur. Sugar. Exp. Station* 1, 7–20. doi: 10.1007/BF00307721

Smook, G. A. (1992). *Handbook for Pulp and Paper Technologists*, Peachtree Corners, GA: TAPPI Press.

Souza, G. M., Simoes, A. C. Q., Oliveira, K. C., Garay, H. M., Fiorini, L. C., Gomes, F., et al. (2001). The sugarcane signal transduction (SUCAST) catalogue: prospecting signal transduction in sugarcane. *Genet. Mol. Biol.* 24, 1–4. doi: 10.1590/S1415-47572001000100005

Sreenivasan, T. V., Ahloowalia, B. S., and Heinz, D. J. (1987). "Cytogenetics," in *Sugarcane Improvement Through Breeding*, ed. D. J. Heinz (Amsterdam: Elsevier), 211–254. doi: 10.1016/B978-0-444-42769-4.50010-2

Strohmeier, M., Hrmova, M., Fischer, M., Harvey, A. J., Fincher, G. B., and Pleiss, J. (2004). Molecular modeling of family GH16 glycoside hydrolases: potential roles for xyloglucan transglucosylases/hydrolases in cell wall modification in the poaceae. *Protein Sci.* 13, 3200–3213. doi: 10.1110/ps.04828404

Tai, P., and Miller, J. (2002). Germplasm diversity among four sugarcane species for sugar composition. *Crop Sci.* 42, 958–964. doi: 10.2135/cropsci2002.0958

Takeda, T., Suwa, Y., Suzuki, M., Kitano, H., Ueguchi-Tanaka, M., Ashikari, M., et al. (2003). The OsTB1 gene negatively regulates lateral branching in rice. *Plant J.* 33, 513–520. doi: 10.1046/j.1365-313X.2003.01648.x

Tew, T. L., and Cobill, R. M. (2008). "Genetic improvement of sugarcane (*Saccharum* spp.) as an energy crop," in *Genetic Improvement of Bioenergy Crops*, ed. W. Vermerris (New York, NY: Springer), 273–294. doi: 10.1007/978-0-387-70805-8_9

Tew, T. L., and Pan, Y. B. (2010). Microsatellite (simple sequence repeat) marker-based paternity analysis of a seven-parent sugarcane polycross. *Crop Sci.* 4, 1401–1408. doi: 10.2135/cropsci2009.10.0579

Theander, O., and Westerlund, E. (1993). "Quantitative analysis of cell wall components," in *Forage Cell Wall Structure and Digestibility*, eds H. G. Jung, D. R. Buxton, R. D. Hatfield, and J. Ralph, (Madison, WI: ASA-CSSA-SSSA), 83–104.

Todd, J. R., Sandhu, H., Hale, A. L., Glaz, B., and Wang, J. (2017). Phenotypic evaluation of a diversity panel selected from the world collection of sugarcane (*Saccharum* spp.) and related grasses. *Maydica* 62:M19.

Todd, J. R., Wang, J., Glaz, B., Sood, S., Ayala-Silva, T., Nayak, S. N., et al. (2014). Phenotypic characterization of the Miami World Collection of sugarcane (*Saccharum* spp.) and related grasses for selecting a representative core. *Genet. Resour. Crop Evol.* 61, 1581–1596. doi: 10.1007/s10722-014-0132-3

van Antwerpen, R., Berry, S. D., van Antwerpen, T., Smithers, J., Joshi, S., and van der Laan, M. (2013). "Sugarcane as an energy crop: its role in biomass economy," in *Biofuel Crop Sustainability*, ed. B. Singh (Hoboken, NJ: John Wiley & Sons, Ltd), 53–108. doi: 10.1002/9781118635797.ch3

Van der Merwe, M. J., Groenewald, J. H., Mark, S., Kossmann, J., and Botha, F. C. (2010). Downregulation of pyrophosphate: *D-fructose-6-phoshate 1-phosphotransferase* activity in sugarcane culms enhances sucrose accumulation due to elevated hexose-phosphate levels. *Planta* 231, 595–608. doi: 10.1007/s00425-009-1069-1

van Der Weijde, T., Alvim Kamei, C. L., Torres, A. F., Vermerris, W., Dolstra, O., Visser, R. G. F., et al. (2013). The potential of C4 grasses for cellulosic biofuel production. *Front. Plant Sci.* 4:107. doi: 10.1186/s13068-016-0479-0

Vanholme, R., Van Acker, R., and Boerjan, W. (2010). Potential of *Arabidopsis* systems biology to advance the biofuel field. *Trends Biotechnol.* 28, 543–547. doi: 10.1016/j.tibtech.2010.07.008

Vettore, A. L., da Silva, F. R., Kemper, E. L., Souza, G. M., da Silva, A. M., Ferro, M. I., et al. (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* 13, 2725–2735. doi: 10.1101/gr.1532103

Vicentini, R., Bottcher, A., Brito, M. S., Santos, A. B., Creste, S., Landell, M. G., et al. (2015). Large-scale transcriptome analysis of two sugarcane genotypes contrasting for lignin content. *PLOS ONE* 10:e0134909. doi: 10.1371/journal.pone.0134909

Waclawovsky, A. J., Sato, P. M., Lembke, C. G., Moore, P. H., and Souza, G. M. (2010). Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content. *Plant Biotechnol. J.* 8, 263–276. doi: 10.1111/j.1467-7652.2009.00491.x

Wang, J., Roe, B., Macmil, S., Yu, Q., Murray, J. E., Tang, H., et al. (2010). Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. *BMC Genomics* 11:261. doi: 10.1186/1471-2164-11-261

Watt, D. A., McCormick, A. J., Govender, C., Carson, D. L., Cramer, M. D., Huckett, B. I., et al. (2005). Increasing the utility of genomics in unravelling sucrose accumulation. *Field Crops Res.* 92, 149–158. doi: 10.1016/j.fcr.2005.01.012

Weng, J.-K., Li, X., Bonawitz, N. D., and Chapple, C. (2008). Emerging strategies of lignin engineering and degradation for cellulosic biofuel production. *Curr. Opin. Biotechnol.* 19, 166–172. doi: 10.1016/j.copbio.2008.02.014

Whipple, C. J., Kebrom, T. H., Weber, A. L., Yang, F., Hall, D., Meeley, R., et al. (2011). Grassy tillers1 promotes apical dominance in maize and responds to shade signals in the grasses. *Proc. Natl. Acad. Sci. U.S.A.* 108, 506–512. doi: 10.1073/pnas.1102819108

Wu, L., and Birch, R. G. (2007). Double sugar content in sugarcane plants modified to produce a sucrose isomer. *Plant Biotechnol.* 5, 109–117. doi: 10.1111/j.1467-7652.2006.00224.x

Yee, K. L., Rodriguez, M. Jr., Tschaplinski, T. J., Engle, N. L., Martin, M. Z., Fu, C., et al. (2012). Evaluation of the bioconversion of genetically modified switchgrass using simultaneous saccharification and fermentation and a consolidated bioprocessing approach. *Biotechnol. Biofuels* 5:81. doi: 10.1186/1754-6834-5-81

Yilmaz, A., Nishiyama, M. Y., Fuentes, B. G., Souza, G. M., Janies, D., Gray, J., et al. (2009). GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol.* 149, 171–180. doi: 10.1104/pp.108.128579

Yin, Y. (2014). "Databases for bioenergy-related enzymes," in *Bioenergy Research: Advances and Applications*, eds V. Gupta, M. Tuohy, C. Kubicek, J. Saddler, and F. Xu (Amsterdam: Elsevier BV), 95–107. doi: 10.1016/B978-0-444-59561-4.00006-1

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 40, W445–W451. doi: 10.1093/nar/gks479

Yong, W., Link, B., O'Malley, R., Tewari, J., Hunter, C. T., Lu, C. A., et al. (2005). Genomics of plant cell wall biogenesis. *Planta* 221, 747–751. doi: 10.1007/s00425-005-1563-z

Zhang, H., Jin, J., Tang, L., Zhao, Y., Gu, X., Gao, G., et al. (2011). PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res.* 39, D1114–D1117. doi: 10.1093/nar/gkq1141

Zhang, T., Wyman, C. E., Jakob, K., and Yang, B. (2012). Rapid selection and identification of *Miscanthus* genotypes with enhanced glucan and xylan yields from hydrothermal pretreatment followed by enzymatic hydrolysis. *Biotechnol. Biofuels* 5:56. doi: 10.1186/1754-6834-5-56

Zhao, X., Zhang, L., and Liu, D. (2012). Biomass recalcitrance. Part I: the chemical compositions and physical structures affecting the enzymatic hydrolysis of lignocellulose. *Biofuels Bioprod. Bioref.* 6, 465–482. doi: 10.1002/bbb.1331

Zhou, H., Liu, G., Liu, J., and He, J. (2009). Genetic analysis of sugarcane biomass yield and its component traits using ADAA models. *J. Trop. Agric.* 47, 70–73.

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: 01 February 2018
doi: 10.3389/fpls.2018.00089

Quantitative Trait Transcripts Mapping Coupled with Expression Quantitative Trait Loci Mapping Reveal the Molecular Network Regulating the Apetalous Characteristic in *Brassica napus* L.

Kunjiang Yu[1,2,3†], *Xiaodong Wang*[1†], *Feng Chen*[1], *Qi Peng*[1], *Song Chen*[1], *Hongge Li*[1], *Wei Zhang*[1], *Sanxiong Fu*[1], *Maolong Hu*[1], *Weihua Long*[1], *Pu Chu*[2], *Rongzhan Guan*[2*] and *Jiefu Zhang*[1*]

[1] Key Laboratory of Cotton and Rapeseed, Ministry of Agriculture, Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing, China, [2] State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing, China, [3] College of Agriculture, Guizhou University, Guiyang, China

OPEN ACCESS

Edited by:
Shuizhang Fei,
Iowa State University, United States

Reviewed by:
Xusheng Wang,
St. Jude Children's Research Hospital, United States
Xin Li,
AgReliant Genetics, LLC, United States
Reka Howard,
University of Nebraska System, United States

***Correspondence:**
Rongzhan Guan
guanrzh@njau.edu.cn
Jiefu Zhang
jiefu_z@163.com

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

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Plant Science

Received: 03 November 2017
Accepted: 16 January 2018
Published: 01 February 2018

Citation:
Yu K, Wang X, Chen F, Peng Q, Chen S, Li H, Zhang W, Fu S, Hu M, Long W, Chu P, Guan R and Zhang J (2018) Quantitative Trait Transcripts Mapping Coupled with Expression Quantitative Trait Loci Mapping Reveal the Molecular Network Regulating the Apetalous Characteristic in Brassica napus L. Front. Plant Sci. 9:89. doi: 10.3389/fpls.2018.00089

The apetalous trait of rapeseed (*Brassica napus*, AACC, $2n = 38$) is important for breeding an ideal high-yield rapeseed with superior klendusity to *Sclerotinia sclerotiorum*. Currently, the molecular mechanism underlying the apetalous trait of rapeseed is unclear. In this study, 14 petal regulators genes were chosen as target genes (TGs), and the expression patterns of the 14 TGs in the AH population, containing 189 recombinant inbred lines derived from a cross between apetalous "APL01" and normal "Holly," were analyzed in two environments using qRT-PCR. Phenotypic data of petalous degree (PDgr) in the AH population were obtained from the two environments. Both quantitative trait transcript (QTT)-association mapping and expression QTL (eQTL) analyses of TGs expression levels were performed to reveal regulatory relationships among TGs and PDgr. QTT mapping for PDgr determined that *PLURIPETALA* (*PLP*) was the major negative QTT associated with PDgr in both environments, suggesting that *PLP* negatively regulates the petal development of line "APL01." The QTT mapping of *PLP* expression levels showed that *CHROMATIN-REMODELING PROTEIN 11* (*CHR11*) was positively associated with *PLP* expression, indicating that *CHR11* acts as a positive regulator of *PLP* expression. Similarly, QTT mapping for the remaining TGs identified 38 QTTs, associated with 13 TGs, and 31 QTTs, associated with 10 TGs, respectively, in the first and second environments. Additionally, eQTL analyses of TG expression levels showed that 12 and 11 unconditional eQTLs were detected in the first and second environment, respectively. Based on the QTTs and unconditional eQTLs detected, we presented a hypothetical molecular regulatory network in which 14 petal regulators potentially regulated the apetalous trait in "APL01" through the *CHR11-PLP* pathway. *PLP* acts directly as the terminal signal integrator negatively regulating petal development in the *CHR11-PLP* pathway. These findings will aid in the understanding the molecular mechanism underlying the apetalous trait of rapeseed.

Keywords: *Brassica napus* L., apetalous, quantitative trait transcript, expression QTL, regulatory network

INTRODUCTION

Flowers of angiosperms are typically composed of four organ types inclined to four floral whorls. From the outside of the flower to the center, these organs are orderly sepals, petals, stamens, and carpels (the subunits of the gynoecium). Over the last 20 years, the molecular mechanism of flower development have been adequately elucidated in several angiosperm species, such as *Arabidopsis thaliana*, *Antirrhinum majus*, *Petunia hybrid,* and *Oryza sativa* (Schwarz-Sommer et al., 1990; Bowman et al., 1991; van der Krol and Chua, 1993; Li et al., 2011; Hirano et al., 2014). Recently, the genetics of flower development in *Ranunculales* were also decoded successfully (Damerval and Becker, 2017). The "ABC model" as the basic model explaining both floral patterning and floral organ identity has been endlessly enriched by works in several eudicot species (Pelaz et al., 2000; Jack, 2001; Theissen and Saedler, 2001). Currently, the "ABCE model," as the most detailed floral model, is guiding investigations that will aid in understanding the origin and diversification of angiosperm flowers.

Petal initiation, a key unit of flower development, is crucial in revealing the evolutionary history of flowering plants. According to the "floral quarter model," A class (*APETALA 1*, *AP1*), B class (*APETALA3* and *PISTILLALA*, *AP3* and *PI*, respectively), and E class (*SEPALLALA 1/2/3*, *SEP1/2/3*) genes are simultaneously required for petal identity in *Arabidopsis* (Theissen and Saedler, 2001; Ditta et al., 2004). Molecular evolutionary studies indicated that B class genes underwent two vital duplication and divergence events, in which the first event generated the *PI* and paleo*AP3* lineages, while the second event generated eu*AP3* and *TM6* lineages (Kramer et al., 1998; Kim et al., 2004). Both paleo*AP3* and *TM6* have the same paleo*AP3* motif regulating stamen development, but they are not involved in petal development (Kramer et al., 1998; Kim et al., 2004; Rijpkema et al., 2006). Eu*AP3* contains the eu*AP3* motif required for development of both petals and stamens (Vandenbussche et al., 2004; de Martino et al., 2006; Rijpkema et al., 2006; Drea et al., 2007; Kramer et al., 2007; Hileman and Irish, 2009). Strangely, although there are both eu*AP3* and *TM6* in most eudicots, there is only eu*AP3* in *Arabidopsis* and snapdragon (Lamb and Irish, 2003; Vandenbussche et al., 2004). In addition to B class genes, there are a number of genes involved in petal development in *Arabidopsis*, many of which function upstream or downstream of ABE class genes (Kaufmann et al., 2009, 2010; Wuest et al., 2012). However, the locations of some genes in the regulatory network of petal development are unclear, such as *PLURIPETALA* (*PLP*) (Running et al., 2004) and *CHROMATIN-REMODELING PROTEIN 11*(*CHR11*) (Smaczniak et al., 2012).

Apetalous rapeseed, which is a novel floral mutant in which the whorl organs are perfectly developed separate from the petals, has advantages of low-energy consumption, high photosynthetic efficiency and superior klendusity to *Sclerotinia sclerotiorum* (Chapman et al., 1984; Yates and Steven, 1987; Morrall, 1996; Jamaux and Spire, 1999). Thus, apetalous rapeseed is considered the ideotype of high-yield rapeseed (Mendham and Rao, 1991; Rao et al., 1991), and it has attracted the attention of botanists and breeders since its appearance. Currently, the molecular mechanism underlying the apetalous characteristic of rapeseed is poorly known because of the lack of stable apetalous mutants and the complexity of polygenic inheritance (Kelly et al., 1995; Fray et al., 1997; Wang et al., 2015; Yu et al., 2016). The apetalous characteristic of rapeseed is mainly governed by recessive genes, usually by two to four loci (Kelly et al., 1995), and several quantitative trait loci (QTLs) regulating petal development on chromosomes A3, A4, A5, A6, A9, C4, and C8 have been identified (Fray et al., 1997; Wang et al., 2015). A deficiency in eu*AP3* expression may give rise to the apetalous characteristic, while the paleo*AP3* expression ensures stamen development in *Brassica napus* (Zhang et al., 2011). This theory, coupled with the "ABCE model," predicts that sepals of apetalous rapeseed should increase, but the number of sepals is actually normal (Zhang et al., 2011). This indicates that the molecular mechanism controlling the apetalous characteristic of rapeseed is more complex than initially believed.

In our previous study (Wang et al., 2015), nine QTLs associated with petalous degree (PDgr) have been detected on chromosomes A3, A5, A6, A9, and C8 in the AH population, containing 189 recombinant inbred lines derived from a cross between an apetalous line "APL01" and a normal petalled variety "Holly." Interestingly, three QTLs, *qPD.A9-2*, *qPD.C8-2,* and *qPD.C8-3*, are stably expressed in multiple environments (Wang et al., 2015). In another study (Yu et al., 2016), genome-wide transcriptomic analyses of the apetalous line "APL01" and another normally petalled line "PL01" both derived from the F_6 generation of crosses between apetalous "Apetalous No. 1" and normal petalous "Zhongshuang No. 4" rapeseed have been performed. Further analysis suggested that a large number of genes involved in protein biosynthesis were differentially expressed at the key stage of petal primordium initiation in "APL01" compared with in "PL01," and 36 petal regulators implicated in the apetalous trait of line APL01 were identified (Yu et al., 2016). Interestingly, the 36 petal regulators were outside of the confidence intervals (CIs) of nine QTLs regulating PDgr, implying that these genes maybe function at the downstream of the QTLs (Yu et al., 2016). However, it's worth noting that mutants of the 36 petal regulators result in defective floral phenotypes other than abnormal petals in *Arabidopsis*, such as (*PLP*) (Running et al., 2004) and (*CHR11*) (Smaczniak et al., 2012). For the aptelous characteristic of rapeseed, these genes collaboratively participate in the regulation of petal development, leading to the unique floral phenotype of "APL01." However, the specifics of this collaborative participation are unclear. Thus, it is necessary to analyze relationships among petal regulators and PDgr using multiple approaches.

A quantitative trait transcript (QTT) analysis is a mixed linear model approach of association mapping of a transcriptome (Zhang et al., 2015). So far, QTT has been applied to detect the transcripts associated with complex traits in mice (Zhang et al., 2015), rice (Zhou et al., 2016), and human (Chen et al., 2016) populations, and it has efficiently identified the genetic effects of individual loci, and epistatic interactions of pair-wise loci or gene-by-gene (G×G) (Zhang et al., 2015; Chen et al., 2016; Zhou et al., 2016). Expression QTL (eQTL) analysis based on linkage mapping is an approach to determining gene expression levels

(Jansen and Nap, 2001). This approach can identify the genetic determinants of gene expression levels and has been successfully used to investigate gene regulatory pathways in plants (DeCook et al., 2005; Jordan et al., 2007; Yin et al., 2010; Wang et al., 2014), animals (Sun et al., 2003; Ghazalpour et al., 2006; Li et al., 2006), and humans (Cheung et al., 2003; Göring et al., 2007; Battle and Montgomery, 2014). Conditional QTL mapping is a method that can exclude the contribution of a causal trait to the variation of the resultant trait (Zhu, 1995). Unconditional QTL mapping coupled with conditional QTL analysis could dissect the genetic relationships between two traits at the QTL level, and then it has been broadly applied to exploring the relationships between QTLs and the corresponding conditional traits (Zhao et al., 2006; Cui et al., 2011; Zhang et al., 2013).

In this study, we analyzed the expression levels of the 36 petal regulators genes and 1 candidate gene *CG1* (*BnaC08g10840D*), underlying the CI of the major QTL *qPD.C8-2* in "APL01," "PL01," and "Holly" by using qRT-PCR. The comparative analyses indicated that both 13 petal regulators genes and *CG1* showed the same dynamic expression levels between "APL01" and "PL01" as between "APL01" and "Holly." Thus, the 14 genes were chosen as target genes (TGs) for quantitative reverse transcription-PCR (qRT-PCR) analyses. The expression patterns of the 14 TGs in the AH population were analyzed in two environments using qRT-PCR. Phenotypic data of PDgr in the AH population were obtained from the two environments. Regulatory relationships among TGs and PDgr were discovered, genomic regions influencing TGs expression were identified, and molecular networks regulating the petal development of an apetalous line "APL01" were constructed as a result of QTT-association mapping coupled with eQTL analyses of TGs expression levels.

MATERIALS AND METHODS

Plant Materials

"APL01" and "PL01" was selected from the F_6 generation of crosses between apetalous ("Apetalous No. 1") and normal petalous ("Zhongshuang No. 4") rapeseed in 1998. "Apetalous No. 1" had been developed from the F_8 generation of crosses between a Chinese rapeseed cultivar with smaller petals (SP103) and *B. rapa* variety with a lower PDgr (LP153). "Zhongshuang No. 4" was bred at the Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences, Wuhan, China. The AH population, containing 189 recombinant inbred lines (RILs), was derived from a cross between an apetalous line "APL01" and a normally petalled variety "Holly." The genotype "Holly" is a completely petalled variety. The AH population was planted in two different districts, Lishui County (coded 2015a) and Xuanwu District (coded 2015b), in Nanjing of Jiangsu Province for one year (September-May of 2014-2015) with good field management measures. The subsequent works were independently performed in both environments.

Collection of Samples, and Evaluation of PDgr

According to our previous study (Yu et al., 2016) and with early flower development studies in *B. napus* (Polowick and Sawhney, 1986) and in *Arabidopsis* (Smyth et al., 1990), the petal primordia appear in the second whorl later in stage 5, but the petal primordia begin growing rapidly at the start of stage 9 in *B. napus*. The length of buds in stage 10 is at least double that of buds in stage 9. To minimize the sampling error, young inflorescences only containing buds at stages 1 to 9 were gathered for the subsequent works after removing stage 10 to 12 buds during flower bud development. At least five young inflorescences derived from five plants in each RIL of the AH population were collected in each environment. A total of two biological samples were collected in each RIL of the AH population. For lines "APL01," "PL01," and "Holly," three biological samples of each line were separately collected. The actual and theoretic numbers of flower petals were recorded in each RIL at early blooming stage. The evaluation of PDgr was carried out as described in our previous study (Wang et al., 2015).

Total RNA Exaction, cDNA Synthesis, and qRT-PCR Assay

Total RNA was isolated using MagaZorb® Total RNA Mini-Prep Kit (Promega, Madison, WI, USA). RNA degradation and contamination were checked on 1% agarose gels. The RNA concentration was measured using the Q3000® Micro-Ultraviolet Spectrophotometer (Quawell, Sunnyvale, CA, USA). First-strand cDNAs were synthesized in a final volume of 20 μL containing 4 μL of 5×PrimeScript RT Master Mix (Perfect Real Time), ≤1 μg of total RNA, and <16 μL of RNase Free dH_2O using PrimeScript™ RT Master Mix (Perfect Real Time) (TaKaRa, Da Lian, China). Sequences of TGs and paralogs were obtained from the *B. napus* genome database (http://www.genoscope.cns.fr/brassicanapus/) (Chalhoub et al., 2014). Primers for the qRT-PCR assay were designed using Primer 5 software and synthesized by Sangon Biotech (Shanghai, China) (**Table S1**). The rapeseed *ACTIN* (*BnaA05g21350D*) gene was chosen as the endogenous reference gene to examine the sample-to-sample variation in the amount of cDNA. Each reaction (20 μL) contained 10 μL of 2×SYBR Premix Ex Taq (Tli RNaseH Plus), 0.8 μL of 10 μM gene-specific primers, 0.4 μL of 50×ROX Reference Dye II, <100 ng of first-strand cDNAs, and <8.8 μL of RNase Free dH_2O according to SYBR® Premix Ex Taq™ (Tli RNaseH Plus) (TaKaRa). The three-step PCR (95°C for 30 s, followed by 40 cycles of 95°C for 5 s, 55°C for 30 s, and 72°C for 30 s) was performed with the ABI PRISM 7500 Real-Time PCR System (Applied Biosystems, Foster, CA, USA). For the qRT-PCR assay on "APL01" vs. "PL01," or "Holly," the later was chosen as the sample for reference. For the qRT-PCR assay in the AH population, RIL43 was chosen as the reference sample. Triplicate replicates for each qRT-PCR assay were performed independently.

Data Collection, Identification of TGs, and Drafting of Standard Curves

PCR cycles (C_t) for all genes were determined in each amplification reaction after removing the reactions with nonspecific and/or unrepeatable amplifications. The relative expression levels of the genes in different samples were calculated using $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001), defined as: $\Delta\Delta Ct = (C_{t,\ target} - C_{t,\ actin})_{genotype} - (C_{t,\ target} - C_{t,\ actin})_{calibrator}$,

in which "genotype" indicates the target sample and "calibrator" indicates the reference sample. In our previous study, 36 petal regulators and 1 candidate gene were identified as differentially expressed genes in line APL01 compared with line PL01 (Yu et al., 2016). In this study, whether the differences in these genes' expression levels between "APL01" and "PL01" or "Holly" are significant depends on the *P*-value estimated using SPSS Statistics 19.0 software (IBM, Armonk, NY, USA) (non-paired *t*-test, $P < 0.05$). Genes showing the same expression patterns between "APL01" and "PL01" as between "APL01" and "Holly" were regarded as TGs for the subsequent analyses. Standard cDNA was diluted 10, 15, 20, 25, 30, and 35 times before the qRT-PCR analysis. The cDNA's dilution ratio is the independent variable of the standard curve, while the C_t values of the TGs and *ACTIN* are the dependent variables. Standard curves of TGs were drawn using Sigma Plot 12.5 software (Systat Software Inc., San Jose, CA, USA). TG expression levels in the AH population were used for QTT mapping and eQTL analysis after removing low quality data. The non-specific PCR amplification of *ACTIN* in each cDNA sample was regarded as the standard for estimating low quality data because the *ACTIN* primer pair consisted of cross-intron primers. To further evaluate the reliability of qRT-PCR data, all of the TG expression data was normalized using the following formula:

$$y = \frac{q - a}{SD}$$

in which "*y*" represents the normalized expression data of TG, "*q*" represents the TG expression level ($2^{-\Delta\Delta Ct}$) in each RIL of the AH population, "*a*" indicates the average of the TG expression levels in the AH population, and "*SD*" is the standard deviation of the TG expression levels in the AH population.

The scatter plot diagram of the normalized expression data of TGs was drawn using Adobe Photoshop CS6 v13.0 software (Adobe Systems Inc., San Jose, CA, USA). The qualified qRT-PCR data should be located in the interval ranging from −2 to 2.

Correlation Analysis, and QTT-Association Mapping for PDgr and TGs

The correlations of PDgr with the TG expression levels in the AH population were assessed using SPSS Statistics 19.0 software (Bivariate correlation, Pearson, $P < 0.05$). QTT-association mapping of PDgr and TGs expression levels in the AH population was performed based on a mixed linear model approach using the QTT functional module of the QTXNetwork software (Zhang et al., 2015). For the QTT analysis of PDgr, the 14 TGs expression levels were the genotypic data, while PDgr was the phenotypic data in each assay. The transcript locus regulating PDgr was called QTT to correspond with the TG. Subsequently, QTT mapping of TGs were performed, and the expression levels of the TGs regulating PDgr served as the phenotypic data, while the remaining TG expression levels served as the genotypic data. QTT regulating TG expression level was called tQTT to correspond with the TG. To the same analogy, QTT-association mapping of the tQTTs (TGs) regulating the corresponding TG expression levels was performed in sequence. The mapping order and permutation time were set to 3 and 1000, respectively. The superior x-Ome prediction was also included. The *P* threshold for declaring a QTT (tQTT) significant was set as 0.05 ($-LogP > 1.3$). The normalized expression data of TG was used for QTT analysis. For mapping transcripts in homozygote population, the dependent variables (y_{kh}) of the *k-th* subject in the *h-th* environment can be expressed by the following mixed linear model (Zhang et al., 2015):

$$y_{kh} = \mu + e_h + \sum_i q_i u_{ik} + \sum_{i<j} qq_{ij} u_{ijk} + \sum_i qe_{ih} u_{ikh} + \sum_{i<k} qqe_{ijh} u_{ijkh} + \varepsilon_{kh}$$

where μ represents the population mean; e_h represents the fixed effect of the *h-th* environment; q_i represents the *i-th* locus effect with coefficient u_{ik} (using expression values in QTT mapping); qq_{ij} represents the epistasis effect of locus $i \times$ locus j with coefficients u_{ijk} (using expression values $u_{ik} \times u_{jk}$ in QTT mapping); qe_{ih} represents the environment interaction effect of the *i-th* locus in the *h-th* environment with coefficient u_{ikh}; qqe_{ijh} represents the epistasis × environment interaction effect of locus $i \times$ locus j in the *h-th* environment with coefficient u_{ijkh}; and ε_{kh} represents the residual effect of the *k-th* individual in the *h-th* environment.

A QTT or tQTT with a heritability of at least 10% ($h^2 \geq 10\%$) was considered the major QTT or tQTT, while QTT or tQTT that was detected repeatedly in the two environments was considered a stable QTT or tQTT. Both are considered as the key QTTs or tQTTs.

Unconditional and Conditional eQTL Mapping of TGs

In our recent study (Wang et al., 2015), the AH genetic linkage map was constructed based on 2755 single-nucleotide polymorphism markers and 57 simple sequence repeats, and the QTLs for PDgr were been successfully detected. In this study, the TG expression levels in the AH population were regarded as phenotypic data for QTL linkage mapping, which was termed unconditional eQTL mapping. The software Windows QTL Cartographer 2.5 (Raleigh, NC, USA) was applied to perform the unconditional eQTL analysis (Wang et al., 2007). The composite interval mapping model was deployed for estimating putative eQTLs with additive effects (Zeng, 1994). The working speed and window size were set to 2, and 10 cM, respectively. The logarithm of odds threshold for detecting a significant eQTL ranged from 2.2 to 3.4 based on permutation test analyses (1,000 permutations, 5% overall error level) as described previously (Churchill and Doerge, 1994). Thus, the false discovery rate for eQTL analysis was 0.05. A conditional eQTL analysis was carried out as described by Zhu (1995). The key tQTTs were regarded as the conditional independent variables, and conditional expression levels (conditional dependent variables) of TGs were generated using the QGAstation software.

Construction of the Molecular Network Involved in Petal Development

Based on tQTTs and unconditional eQTLs, combined with our previous research (Wang et al., 2015; Yu et al., 2016), a regulatory network for the apetalous characteristic in "APL01" was constructed using Adobe Photoshop CS6 v13.0 software (Adobe Systems Inc).

RESULTS

Identification of TGs, and TG Expression Levels in the AH Population

In a previous study (Yu et al., 2016), 36 petal regulators and several candidate genes involved in the apetalous characteristic of line APL01 were obtained (**Table S2**). In this study, we determined that 13 petal regulators and 1 candidate gene *CG1* (candidate gene 1, *BnaC08g10840D*) showed the same expression patterns between "APL01" and "Holly" as between "APL01" and "PL01" as determined by qRT-PCR assays (**Figure 1**, **Table S2**). Thus, the 14 genes were regarded as TGs for the subsequent analyses. For these TGs, the expression levels of 3 genes increased at least 1.5-fold, while those of 11 decreased more than 1.6-fold in "APL01" compared with in "Holly" (**Table S2**).

To estimate the relative expression levels of TGs, the rapeseed *ACTIN* was used as the endogenous reference gene to determine the sample-to-sample variation in the amount of cDNA. As shown in **Figure 2**, the slopes of the curves for each TG are almost to the same as that of *ACTIN*, indicating that the amplification efficiency was the same for the 14 TGs and *ACTIN* (**Table S3**). Subsequently, the expression levels of 14 TGs in the AH population were generated from the two environments using qRT-PCR. After removing low quality data, a high-quality dataset derived from 174 RILs was obtained for the next experiment. The scatter plot diagram of the normalized expression data of TGs suggested that most of data were located in the interval from −2 to 2 (**Figure S1**), indicating that qRT-PCR data used in this study was reliable.

Correlation Analysis

Correlation analyses between two biological replicates of TG expression within an environment determined that the

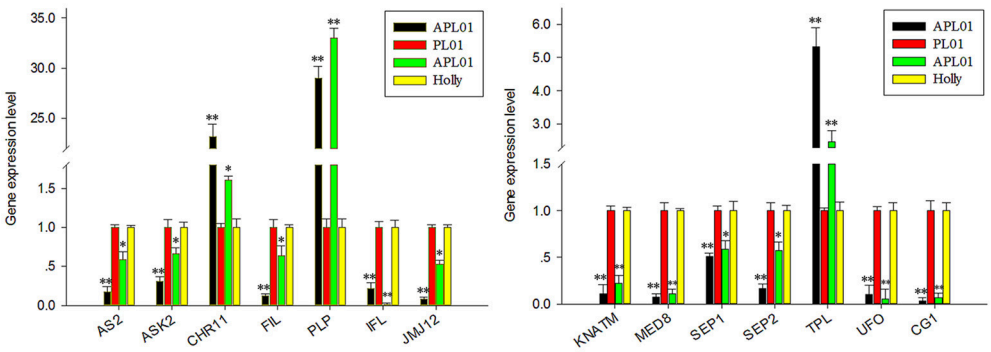

FIGURE 1 | Verification of TG expression patterns by using qRT-PCR**.** Fourteen putative petal regulators showed the same expression patterns between "APL01" and "PL01" (black and red bars, respectively) as between "APL01" and "Holly" (green and yellow bars, respectively). Rapeseed *ACTIN* was chosen as the internal control to normalize the expression data. Data are the mean with standard error (SE) from three independent experiments. Single asterisk indicates that the difference is significant (non-paired *t*-test, $P < 0.05$), double asterisks indicate that the difference is extremely significant (non-paired *t*-test, $P < 0.01$).

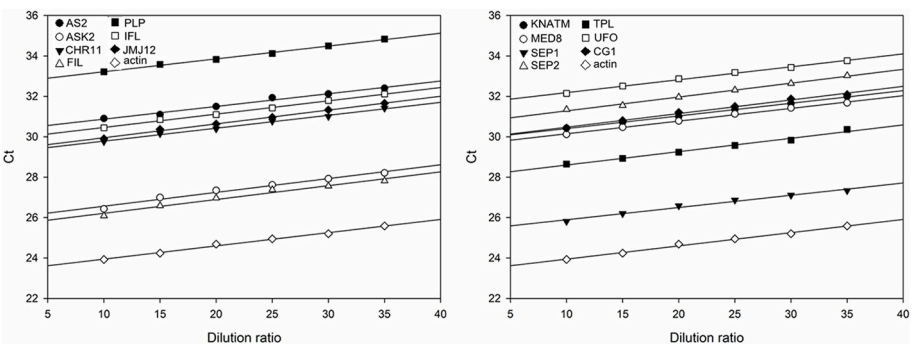

FIGURE 2 | Standard curves for the amplification of 14 TGs and the endogenous reference gene *ACTIN***.** The divisions on the horizontal axis represent the eight dilution ratio of standard cDNA, while the divisions on the vertical axis represent the threshold cycle values (C_t) of the amplification. The amplification reactions of the TGs are described by the corresponding regression formulae (**Table S3**). The slope of the curves reflects the amplification efficiency of the corresponding TGs.

Pearson correlation coefficient was at least 0.601, which means that qRT-PCR data was repeatable (**Table S4**). Correlation analysis of PDgr determined that the Pearson correlation coefficient was 0.806 between the two environments (Bivariate correlation, $P = 2.01E\text{-}40$) (**Table 1**), which suggests that there was a slight difference in PDgr between two environments. The expression levels of the TGs in the AH population, except for *CHROMATIN-REMODELING PROTEIN 11* (*CHR11*), *SEP1,* and *TOPLESS* (*TPL*), showed highly significant correlations between the two environments, and the Pearson correlation coefficients ranged from 0.266 to 0.925 (**Table 1**), indicating that the TGs' expression levels were differentially affected by different environments. Furthermore, random errors have an obvious effect on the difference in TG expression between the two environments probably.

The correlation analyses between TG expression levels and PDgr indicated that only three TGs, *CHR11*, *PLP,* and *INTERFASCICULAR FIBERLESS* (*IFL*), were significantly and negatively correlated with PDgr in the first environment, while two (*PLP* and *TPL*) were significantly and negatively correlated to PDgr in the second environment (**Table 1**). Noticeably, based only on the correlation between TGs and PDgr, it is impossible to explain the molecular mechanism underlying the apetalous characteristic of rapeseed. In fact, the correlation analysis cannot determine the regulatory relationship between genotype and phenotype, because many genes usually participate in the regulation of phenotypic variation in an indirect manner.

QTT-Association Mapping for PDgr and TG Expression Levels

To study relationships between PDgr and the TGs, QTT-association analyses of both PDgr and TG expression levels in the AH population were performed in two environments.

In the first environment, QTT-association analysis of PDgr indicated that *PLP* was the only QTT ($-LogP = 9.86$, $h^2 = 18.62\%$) associated with PDgr that had an obvious and negative effect on PDgr. As shown in **Table 2**, the effect of *PLP* on PDgr was −6.88, meaning that PDgr will be decreased 6.88% when the expression level of *PLP* increases one unit in value. The transcript-association mapping of *PLP* expression levels showed that only *CHR11* ($-LogP = 9.08$, $h^2 = 17.28\%$) was associated with *PLP* expression, and the effect was 46.77, meaning that the expression level of *PLP* would be up-regulated 46.77 units in value when that of *CHR11* was up-regulated one unit in value (**Table 2**). Subsequently, the QTT analysis of *CHR11* expression levels detected two tQTTs regulating *CHR11* expression, *PLP* and *JUMONJI DOMAIN-CONTAINING PROTEIN 12* (*JMJ12*)×*SEP2*, and the transcript epistasis loci *JMJ12*×*SEP2* had a negative effect on *CHR11* (**Table 2**). By analogy, QTT-association mapping for *JMJ12*, *SYMMETRIC LEAVES 2* (*AS2*), *MEDIATOR SUBUNIT 8* (*MED8*), *CG1*, *ARABIDOPSIS SKP1 HOMOLOGUE 2* (*ASK2*), *KNOX ARABIDOPSIS THALIANA MEINOX* (*KNATM*), *UNUSUAL FLORAL ORGANS* (*UFO*), *SEP2*, *FILAMENTOUS FLOWER* (*FIL*), *TPL*, *SEP1,* and *IFL* expression levels suggested the existence of one to six tQTTs (**Table 2**, **Table S5**). In addition to *FIL* and *SEP1*, there was at least one major tQTT ($h^2 \geq 10\%$) for each TG. Furthermore,

TABLE 1 | Correlation analyses of both TGs and PDgr in the AH population.

Group A[a]	AS2_1[b] vs. 2[c]	ASK2_1 vs. 2	CHR11_1 vs. 2	FIL_1 vs. 2	PLP_1 vs. 2
r	0.876**	0.840**	0.029	0.468**	0.868**
	IFL_1 vs. 2	JMJ12_1 vs. 2	KNATM_1 vs. 2	MED8_1 vs. 2	SEP1_1 vs. 2
r	0.463**	0.925**	0.810**	0.266**	−0.012
	SEP2_1 vs. 2	TPL_1 vs. 2	UFO_1 vs. 2	CG1_1 vs. 2	PDgr_1 vs. 2
r	0.237**	0.028	0.564**	0.753**	0.806**
Group B[d]	AS2_1 vs. PDgr_1	ASK2_1 vs. PDgr_1	CHR11_1 vs. PDgr_1	FIL_1 vs. PDgr_1	PLP_1 vs. PDgr_1
r	−0.025	−0.076	−0.302**	−0.016	−0.442**
	IFL_1 vs. PDgr_1	JMJ12_1 vs. PDgr_1	KNATM_1 vs. PDgr_1	MED8_1 vs. PDgr_1	SEP1_1 vs. PDgr_1
r	−0.311**	−0.052	−0.029	−0.014	−0.032
	SEP2_1 vs. PDgr_1	TPL_1 vs. PDgr_1	UFO_1 vs. PDgr_1	CG1_1 vs. PDgr_1	
r	0.055	−0.028	−0.033	0.017	
Group C[e]	AS2_2 vs. PDgr_2	ASK2_2 vs. PDgr_2	CHR11_2 vs. PDgr_2	FIL_2 vs. PDgr_2	PLP_2 vs. PDgr_2
r	−0.105	−0.003	0.01	0.025	−0.400**
	IFL_2 vs. PDgr_2	JMJ12_2 vs. PDgr_2	KNATM_2 vs. PDgr_2	MED8_2 vs. PDgr_2	SEP1_2 vs. PDgr_2
r	−0.078	−0.084	−0.058	−0.018	0.072
	SEP2_2 vs. PDgr_2	TPL_2 vs. PDgr_2	UFO_2 vs. PDgr_2	CG1_2 vs. PDgr_2	
r	−0.066	−0.282**	−0.109	−0.133	

*TGs, target genes; PDgr, petalous degree. [a]Group A indicates the correlation analyses of TGs' expression patterns and PDgr in the AH population between two environments. [b]The expression levels of TGs in the first environment. [c]The expression levels of TGs in the second environment. r represents the Pearson correlation coefficient. [d]Group B indicates the correlation analyses between the TGs and PDgr in the first environment. [e]Group C indicates the correlation analyses between the TGs and PDgr in the second environment. Significance levels are as follows: **P < 0.01.*

TABLE 2 | The key QTTs and tQTTs for PDgr and TGs detected in the first environment.

Trait	QTT[a] (tQTT)[b]	Effect[c]	Predict[d]	*SE*	*−Logp*	h^2(%)	EC(A-H)[e]	PV[f]
Petalous degree	**PLP**	**q**	**−6.88**	**1.072**	**9.86**	**18.62**	**1464.34**	**−10079.21**
PLP expression	*CHR11*	*q*	*46.77*	*7.614*	*9.08*	*17.28*	*2.86*	*133.95*
CHR11 expression	*PLP*	*q*	*0.31*	*0.046*	*11.12*	*12.91*	*1464.34*	*457.9*
JMJ12 expression	**AS2**	**q**	**0.51**	**0.041**	**34.55**	**29.86**	**−1.09**	**−0.55**
	MED8	*q*	*0.34*	*0.041*	*16.13*	*13.57*	*−0.4*	*−0.14*
	CG1	*q*	*0.36*	*0.041*	*17.97*	*15.12*	*−73.77*	*−26.65*
AS2 expression	*ASK2*	*q*	*0.74*	*0.063*	*30.8*	*19.08*	*−2.68*	*−1.97*
	JMJ12	**q**	**0.87**	**0.063**	**42.21**	**26.42**	**−0.24**	**−0.21**
	CG1	*q*	*0.75*	*0.063*	*31.84*	*19.75*	*−73.77*	*−55.28*
MED8 expression	*JMJ12*	*q*	*0.49*	*0.046*	*26.01*	*32.35*	*−0.24*	*−0.12*
CG1 expression	*AS2*	*q*	*51.28*	*6.725*	*13.59*	*13.58*	*−1.09*	*−55.99*
	JMJ12	**q**	**91.27**	**6.725**	**41.04**	**43.02**	**−0.24**	**−21.67**
	UFO	**q**	**18.72**	**6.725**	**2.27**	**1.81**	**−21.97**	**−411.33**
ASK2 expression	**KNATM**	**q**	**2.11**	**0.185**	**29.43**	**36.26**	**−1.74**	**−3.68**
KNATM expression	**ASK2**	**q**	**1.33**	**0.125**	**25.75**	**35.47**	**−2.68**	**−3.56**
UFO expression	**JMJ12**	**q**	**32.67**	**2.47**	**39.05**	**37.32**	**−0.24**	**−7.76**
	SEP2	*q*	*25.74*	*2.484*	*24.36*	*23.16*	*−84.05*	*−2163.26*
SEP2 expression	**UFO**	**q**	**255.74**	**0.001**	**300**	**12.75**	**−21.97**	**−5619.2**
	FIL×TPL	*qq*	*−668.94*	*0.001*	*300*	*87.25*	*115.04*	*−76952.77*
FIL expression	**ASK2**	**q**	**0.26**	**0.084**	**2.75**	**4.56**	**−2.68**	**−0.7**
TPL expression	*ASK2*	*q*	*38.47*	*2.478*	*53.33*	*34.03*	*−2.68*	*−103.09*
	ASK2×SEP2	*qq*	*39.91*	*5.992*	*10.55*	*36.61*	*−1088.97*	*−43455.73*
SEP1 expression	FIL	q	0.12	0.028	5.02	9.58	−0.83	−0.1
IFL expression	**KNATM**	**q**	**0.85**	**0.137**	**9.21**	**16.08**	**−1.74**	**−1.48**

[a] QTT, quantitative trait transcript associated with PDgr. [b] tQTT, QTT regulating TGs expression. [c] q indicates the individual transcript loci, and qq indicates the additive by additive effects. [d] Predicted effect of QTT or tQTT for the target trait. SE, standard error. −Logp, the minus log of the P-value for detecting a significant QTT or tQTT. h^2, the heritability of QTT or tQTT. [e] Expression change of QTT or tQTT, the incremental expression level of QTT or tQTT in "APL01" compared with in "Holly." [f] Phenotypic variation of target trait, the incremental phenotype variation in "APL01" compared with in "Holly." The bold QTTs or tQTTs are detected repeatedly in all two environments. The italic tQTTs are detected only in the first environment.

there was always one stable tQTT (repeatedly detected in the two environments) for eight TGs except for *CHR11*, *TPL,* and *SEP1*, while there are two for *CG1*. Specifically, *AS2* was a stable tQTT with a positive effect for *JMJ12*, and *JMJ12* acted as the positive and stable tQTT regulating *AS2*, *CG1,* and *UFO* expression. *UFO*, as a stable tQTT, played a positive role in the regulation of *CG1* and *SEP2* expressions. *KNATM* served as a stable tQTT positively regulating *ASK2* and *IFL* expressions. *ASK2*, as the stable tQTT, had positive effects on *KNATM* and *FIL* expression levels. In addition, there was at least one transcript epistasis loci for TG expression apart from *PLP*, *CG1*, *FIL*, *SEP1,* and *IFL* (**Table S5**).

In the second environment, QTT-association mapping for PDgr still showed that only *PLP* ($-LogP = 7.79$, $h^2 = 15.11\%$) was negatively associated with PDgr, and the effect was −6.13 (**Table 3**). However, the QTT analysis of *PLP* expression levels did not detect any tQTT associated with *PLP*, implying that genes other than the 14 TGs in the present study regulate *PLP* expression. In the same way, the QTT mapping of *AS2*, *JMJ12*, *UFO*, *CG1*, *FIL*, *TPL*, *IFL*, *SEP2*, *ASK2,* and *KNATM* expression levels suggested the existence of one to five tQTTs (**Table 3**, **Table S6**). Compared with the first environment, in addition to the 10 stable tQTTs, there was also at least one major tQTT ($h^2 \geq 10\%$) that was only detected in the second environment for the six TGs, including *PLP*, *JMJ12*, *UFO*, *CG1*, *TPL*, and *ASK2* (**Table 3**, **Table S6**). In particular, *UFO* was the major tQTT positively regulating *JMJ12* expression. *CG1*, as a major tQTT, had a positive effect on *UFO* expression. The major transcript epistasis loci, *FIL×TPL* and *IFL×SEP2*, played positive roles in the regulation of *CG1* expression. For the two major tQTTs regulating *TPL* expression, *IFL×SEP2* served as a positive regulator, while *KNATM×SEP2* acted as a negative regulator. Another major tQTT (*FIL×KNATM*) for *ASK2* showed a positive effect on the regulation of *ASK2* expression. Furthermore, just like in the first environment, there was a universal transcript epistatic effect among most of TGs (**Table 3**, **Table S6**), suggesting that the epistatic effect between TGs was vital regulator of TG expression. In addition, QTT analyses of *CHR11*, *SEP1,* and *MED8* did not detect any tQTT.

Unconditional and Conditional eQTL Mapping of TGs

In addition to aforementioned tQTTs, the genomic region is another key factor influencing TG expression levels. In our previous work (Wang et al., 2015), the AH map, a high-density genetic linkage map of 2,027.53 cM with an average marker interval of 0.72 cM, has been constructed and used to identify QTLs for PDgr. An eQTL analysis for TGs was performed based on the AH map.

TABLE 3 | The key QTTs and tQTTs for PDgr and TGs detected in the second environment.

Trait	QTT[a] (tQTT)[b]	Effect[c]	Predict[d]	*SE*	*−Logp*	h^2(%)	EC(A-H)[e]	PV[f]
Petalous degree	**PLP**	**q**	**−6.13**	**1.084**	**7.79**	**15.11**	**1182.48**	**−7243.16**
AS2 expression	**JMJ12**	**q**	**4.76**	**0.134**	**268.05**	**72.04**	**−0.41**	**−1.94**
JMJ12 expression	**AS2**	**q**	**1.77**	**0.044**	**300**	**71.86**	**−1.87**	**−3.32**
	UFO	*q*	*0.79*	*0.044*	*72.05*	*14.28*	*−30.95*	*−24.48*
UFO expression	**JMJ12**	**q**	**22.7**	**1.324**	**64.64**	**42.25**	**−0.41**	**−9.24**
	CG1	*q*	*16.72*	*1.321*	*35.88*	*22.94*	*−90.88*	*−1519.87*
CG1 expression	**JMJ12**	**q**	**189.96**	**4.464**	**300**	**41.14**	**−0.41**	**−77.35**
	UFO	**q**	**30.07**	**4.452**	**10.84**	**1.03**	**−30.95**	**−930.82**
	FIL×TPL	*qq*	*179.13*	*11.35*	*55.04*	*36.58*	*69.61*	*12468.74*
	IFL×SEP2	*qq*	*120.76*	*4.267*	*172.11*	*16.63*	*−154.73*	*−18686.17*
FIL expression	**ASK2**	**q**	**1.58**	**0.139**	**29.27**	**42.27**	**−2.08**	**−3.29**
TPL expression	*IFL×SEP2*	*qq*	*1515.97*	*0*	*300*	*80.2*	*−154.73*	*−234570.01*
	KNATM×SEP2	*qq*	*−753.16*	*0*	*300*	*19.8*	*−25.6*	*19279.63*
IFL expression	**KNATM**	**q**	**41.38**	**4.531**	**19.14**	**28.74**	**−1.49**	**−61.48**
	CG1	q	24.36	4.518	7.15	9.96	−90.88	−2213.99
SEP2 expression	**UFO**	**q**	**30.13**	**4.383**	**11.19**	**21.27**	**−30.95**	**−932.64**
ASK2 expression	**KNATM**	**q**	**6.63**	**0.309**	**100.36**	**62.65**	**−1.49**	**−9.84**
	FIL×KNATM	*qq*	*2.69*	*0.137*	*84.23*	*10.35*	*−8.07*	*−21.74*
KNATM expression	**ASK2**	**q**	**7.67**	**0.223**	**252.36**	**85.14**	**−2.08**	**−15.95**

The definitions of a–f are the same as in ***Table 2****. The bold QTTs or tQTTs are detected repeatedly in all two environments. The italic tQTTs are detected only in the second environment.*

In the current study, unconditional eQTL linkage mapping of 14 TG expression levels in the first environment suggested the existence of one to three eQTLs (**Figure 3**, **Table 4**, **Table S7**), and *uqCHR11C4-2*, *uqSEP1A5-1,* and *uqSEP1A5-1* explained 11.17, 10.76, and 10.11%, respectively, of the estimated phenotypic variation, while the remaining eQTLs explain less than 10% (**Table 4**, **Table S7**). Further analyses of the eQTLs determined that *uqJMJ12A3* (43.5–44.4 cM) shared the same single-nucleotide polymorphism marker (Bn-A03-p15435174, 44.42 cM) with *uqMED8A3* (43.6–44.4 cM), and the two eQTLs were close to *qPD.A3* (46.9–49.5 cM) for PDgr (Wang et al., 2015), and may be regarded as pleiotropic effects caused by the same locus. However, none of the unconditional eQTLs colocalized with the QTLs identified in the previous study for PDgr (**Figure 3**, **Table S7**). Furthermore, all of the unconditional eQTLs mapped to chromosomes different from the corresponding TGs, which means that these eQTLs are trans-acting factors based on the classification rules of eQTL (Kliebenstein, 2008; Sasayama et al., 2012). To evaluate the reliability of QTT analysis results in the first environment, a conditional eQTL analysis was carried out as described by Zhu (1995). Because there is almost one key tQTT ($h^2 \geq 10\%$ or repeatedly detected in the two environments) for each TG, their conditional expression levels for the key tQTT can be generated using the QGAstation software. Conditional eQTL mapping suggested that only four conditional eQTLs, *cqIFLA8*, *cqKNATMA6*, *cqKNATMC2,* and *cqTPLC7*, were obtained (**Table 4**), and they were different from the unconditional eQTLs (**Figure 3**). The result suggested that the four conditional eQTLs were suppressed by the corresponding conditional independent variables, *ASK2*, *IFL,* and *ASK2×SEP2*, under the unconditional situation. Furthermore, the four conditional eQTLs had negative effects on the corresponding TGs expression, which implied that *ASK2*, *IFL,* and *ASK2×SEP2* could act as the positive regulator of *IFL*, *KNATM,* and *TPL* expression. Interestingly, the results were consistent with the results of QTT analyses. Thus, conditional eQTL analyses further confirm the validity of QTT-association mapping for TGs expression levels.

In the second environment, unconditional eQTL analyses of the 10 TGs showed that only 11 unconditional eQTLs for 7 TGs were detected (**Figure 3**, **Table 5**, and **Table S8**). All of the unconditional eQTLs were distinguishable from those detected in the first environment (**Figure 3**). Comparing to QTLs identified for PDgr in a previous study, the confidence interval of *uqSEP1A9* (59.4–62.2 cM) overlapped that of *qPD.A9-1* (59.66–74.36 cM) (**Figure 3**, **Table S8**), suggested that *qPD.A9-1* participates in the petal development of line APL01 by regulating *SEP1* expression. In the relationship between the unconditional eQTL and the corresponding TG, *uqTPLA7* is a cis-acting factor (within 5 Mb), while the remaining 10 unconditional eQTLs are trans-acting factors (on different chromosomes). In addition, just as in the first environment, the conditional expression levels of the TGs were obtained using the key tQTTs. The conditional eQTL mapping of TGs showed that 13 conditional eQTLs for 6 TGs were obtained (**Figure 3**, **Table 5**). The conditional eQTL *uqCG1A8* (73.1–74.7 cM) is the same as the unconditional eQTL *cqCG1A8-2* (73.1–74.7 cM), while the remaining conditional eQTLs are novel compared with the unconditional eQTLs (**Figure 3**, **Table 5**). Over half conditional eQTLs had negative effects on the corresponding TGs, which was consistent with the QTT mapping results. More detailed information

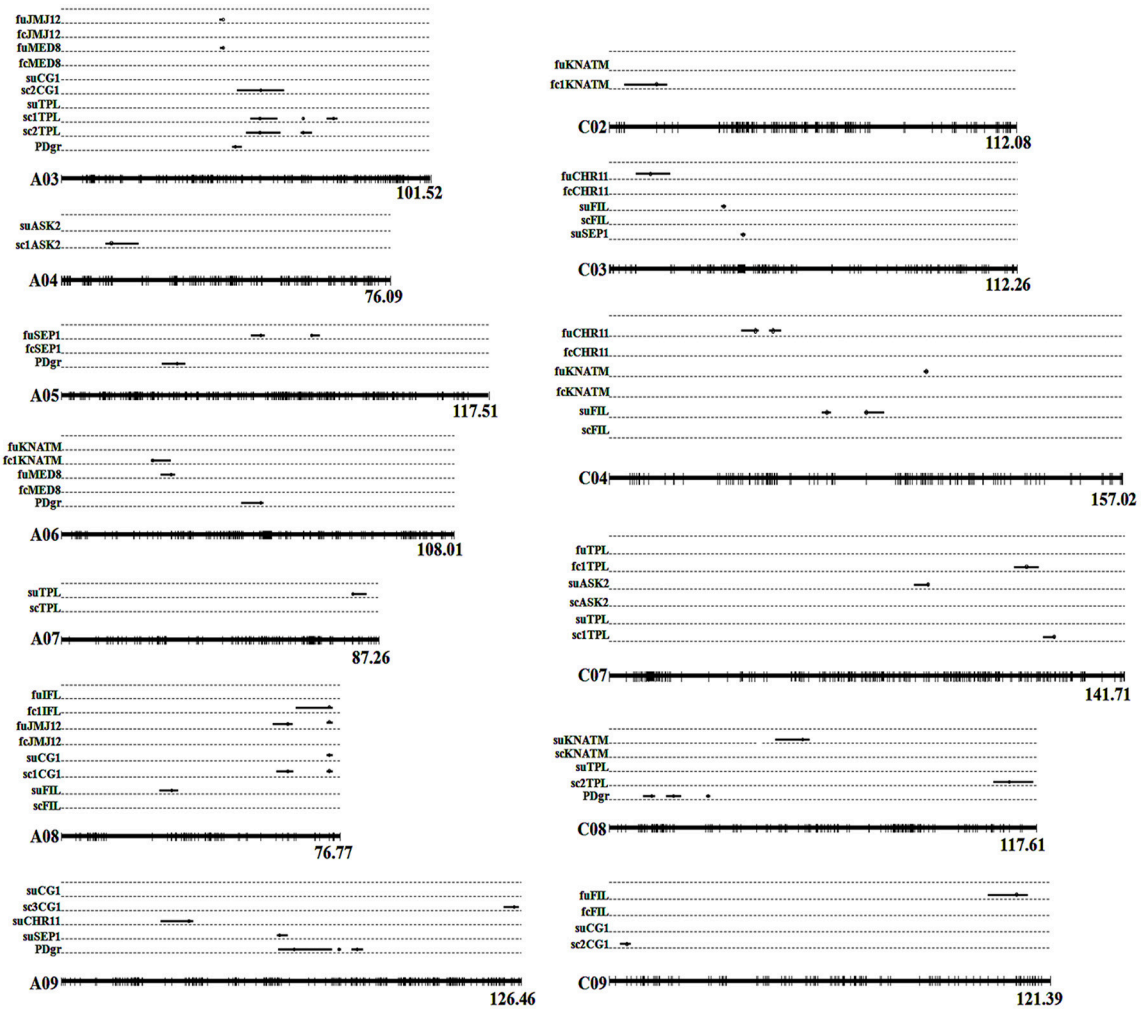

FIGURE 3 | Alignments between unconditional and conditional eQTLs of TG expression levels in two environments**.** Whole linkage groups are represented by *black lines* labeled with molecular markers (*short vertical bars*) on the bottom. The *Arabic numerals* listed on the right side indicate the lengths of the linkage groups. The TGs' unconditional and conditional expression levels are listed on the left side. "fu" represents the TG's unconditional expression level, while "fc" represents the TG's conditional expression level in the first environment. "su" represents the TG's unconditional expression level, while "sc" represents the TG's conditional expression level in the second environment. The black lines on the linkage groups show the QTL confidence interval and the circles indicate the peak position. Detailed information of eQTLs is shown in **Tables 4**, **5**. PDgr is the acronym of petalous degree. fcCHR11: *CHR11|PLP*, *CHR11|MED8*; fcFIL: *FIL|ASK2*, *FIL|SEP1*; fc1IFL: *IFL|ASK2*; fcKNATM: *KNATM|ASK2*, *KNATM|IFL*; fcJMJ12: *JMJ12|AS2*, *JMJ12|MED8*, *JMJ12|CG1*; fc1KNATM: *KNATM|IFL*; fcMED8: *MED8|CHR11*, *MED8|JMJ12*; fcSEP1: *SEP1|FIL*; fc1TPL: *TPL|ASK2*×*SEP2*. scASK2: *ASK2|KANT*, *ASK2|FILxKNATM*; sc1ASK2: *ASK2|FIL*×*KNATM*; scCG1: *CG1|JMJ12*, *CG1|UFO*, *CG1|FIL*×*TPL*, *CG1|IFL*×*SEP2*; sc1CG1: *CG1|FIL*×*TPL*; sc2CG1: *CG1|IFL*×*SEP2*; sc3CG1: *CG1|UFO*; scFIL: *FIL|ASK2*; scKNATM: *KNATM|ASK2*; scTPL: *TPL|IFL*×*SEP2*, *TPL|KNATM*×*SEP2*; sc1TPL: *TPL|IFL*×*SEP2*; sc2TPL: *TPL|KNATM*×*SEP2*.

on the conditional eQTLs was provided in **Table 5** and **Table S8**.

TGs Regulate Petal Development through *CHR11-PLP* Pathway

Based on the QTTs and unconditional eQTLs in this study, together with our previous works (Wang et al., 2015; Yu et al., 2016), a hypothetical regulatory network involved in petal development of "APL01" was constructed. As shown in **Figure 4**, the 14 petal regulators potentially regulate the petal development of "APL01" through the *CHR11-PLP* pathway. *PLP* acts as the terminal signal integrator negatively regulating petal development in the *CHR11-PLP* pathway. In addition, *PLP* expression level may be negatively regulated by *AS2* in other manners as well.

The *CHR11-PLP* pathway consists of 29 tQTTs and 12 unconditional eQTLs (**Figure 4**). *PLP* directly and negatively regulates petal development of line APL01 in the *CHR11-PLP* pathway. *CHR11* acts as the main promoter of *PLP* expression, while *CHR11* is positively regulated by *PLP* as well. The transcripts of the epistatic loci *JMJ12*×*SEP2* are key negative regulator of *CHR11*. Three unconditional eQTLs with negative effects, *uqCHR11C3*, *uqCHR11C4-1*, and *uqCHR11C4-2*, also participate in the regulation of *CHR11*

TABLE 4 | The eQTLs for TGs unconditional and conditional expression levels in the first environment.

Trait	eQTL	Chr[a]	Peak	Marker[b]	CI[c]	*LOD*	R^2(%)	*Add*[d]	Env.[e]	Acting[f]
CHR11 expression	*uqCHR11C3*[g]	C3	11.41	Bn-scaff_16614_1-p546020 (7.382)	7.4–16.8	4.08	9.61	−0.26	2015a	trans-eQTL
	uqCHR11C4-1	C4	44.61	Bn-scaff_15908_1-p289000 (44.593)	40.3–45.5	3.70	7.63	−0.23	2015a	trans-eQTL
	uqCHR11C4-2	C4	50.01	Bn-scaff_19248_1-p207144 (49.936)	48.9–52.5	5.59	11.17	−0.28	2015a	trans-eQTL
FIL expression	*uqFILC9*	C9	112.01	Bn-scaff_17750_1-p587349 (111.994)	104.2–115.1	3.62	5.66	−0.30	2015a	trans-eQTL
JMJ12 expression	***uqJMJ12A3***	**A3**	**44.41**	**Bn-A03-p15435174 (44.42)**	**43.5–44.4**	**3.02**	**6.02**	**0.41**	**2015a**	**trans-eQTL**
	uqJMJ12A8-1	A8	62.31	Bn-A08-p19642040 (62.239)	58.2–63.5	3.11	6.81	0.54	2015a	trans-eQTL
	uqJMJ12A8-2	A8	73.71	Bn-A08-p20855631 (73.676)	73.1–74.7	3.74	7.59	−0.57	2015a	trans-eQTL
KNATM expression	*uqKNATMC4*	C4	96.91	BRAS021 (96.859)	96.2–97.5	3.99	8.84	−0.83	2015a	trans-eQTL
MED8 expression	***uqMED8A3***	**A3**	**44.41**	**Bn-A03-p15435174 (44.42)**	**43.6–44.4**	**2.94**	**6.43**	**0.27**	**2015a**	**trans-eQTL**
	uqMED8A6	A6	30.21	Bn-A06-p4933932 (30.205)	27.4–31.3	3.34	7.39	0.27	2015a	trans-eQTL
SEP1 expression	*uqSEP1A5-1*	A5	54.81	BnGMS294 (56.076)	52.1–55.8	3.75	10.76	−0.16	2015a	trans-eQTL
	uqSEP1A5-2	A5	68.81	Bn-A05-p17758084 (68.739)	68.5–71.0	4.46	10.11	0.16	2015a	trans-eQTL
IFL|ASK2 expression	*cqIFLA8*[h]	A8	73.71	Bn-A08-p20855631 (73.676)	64.6–74.7	3.24	7.01	−0.52	2015a	trans-eQTL
KNATM|IFL expression	*cqKNATMA6*	A6	25.11	Bn-A06-p4056446 (25.131)	24.9–30.1	3.75	8.10	−0.65	2015a	trans-eQTL
	cqKNATMC2	C2	13.01	Bn-scaff_15714_1-p1091353 (12.977)	4.2–15.9	3.58	7.66	−0.63	2015a	trans-eQTL
TPL|ASK2×SEP2 expression	*cqTPLC7*	C7	114.71	Bn-scaff_16069_1-p4085872 (114.718)	111.3–118.0	3.29	7.28	−12.01	2015a	trans-eQTL

[a]*Chromosome.* [b]*The closest marker and the marker position in the AH map.* [c]*The 2-LOD confidence interval of eQTLs.* [d]*Additive effects.* [e]*2015a represents the first environment in which the eQTLs were detected.* [f]*The classification of eQTLs, cis-eQTL is mapped to the same genomic location like an expressed gene (within 5 Mb) while trans-eQTL is mapped to a different genomic location from an expressed gene (>5 Mb or on different chromosomes).* [g]*The unconditional eQTLs of TGs expression levels.* [h]*The conditional eQTLs of TGs conditional expression levels. The bold eQTLs are at the same position with each other.*

TABLE 5 | The eQTLs for TGs unconditional and conditional expression levels in the second environment.

Trait	eQTL	Chr[a]	Peak	Marker[b]	CI[c]	*LOD*	R^2(%)	*Add*[d]	Env.[e]	Acting[f]
ASK2 expression	*qASK2C7*[g]	C7	87.61	Bn-scaff_15705_1-p2283255 (87.581)	83.8–87.9	2.56	5.74	−3.46	2015b	trans-eQTL
CHR11 expression	*qCHR11A9*	A9	35.11	Bn-A09-p17415894 (35.144)	27.4–36.2	2.75	7.01	−4.05	2015b	trans-eQTL
FIL expression	*qFILA8*	A8	30.41	Bn-A08-p15723645 (30.366)	27.0–32.1	3.68	7.34	−0.68	2015b	trans-eQTL
FIL expression	*qFILC3*	C3	31.61	Bn-scaff_22728_1-p1065288 (31.571)	30.9–32.1	3.42	6.98	0.66	2015b	trans-eQTL
FIL expression	*qFILC4-1*	C4	66.51	Bn-scaff_20042_1-p1582 (66.539)	65.0–67.7	7.70	16.48	1.29	2015b	trans-eQTL
FIL expression	*qFILC4-2*	C4	78.61	Bn-scaff_26946_1-p121318 (78.628)	78.0–84.0	4.77	10.15	−1.00	2015b	trans-eQTL
KNATM expression	*qKNATMC8*	C8	53.21	Bn-scaff_16770_1-p3966893 (53.829)	45.7–55.0	2.97	6.70	−2.20	2015b	trans-eQTL
SEP1 expression	***qSEP1A9***[i]	**A9**	**60.01**	**Bn-A09-p25821544 (59.96)**	**59.4–62.2**	**3.01**	**7.03**	**−1.10**	**2015b**	**trans-eQTL**
SEP1 expression	*qSEP1C3*	C3	36.81	Bn-scaff_18322_1-p2155092 (36.875)	36.5–37.0	2.63	5.78	0.75	2015b	trans-eQTL
TPL expression	*qTPLA7*	A7	80.11	Bn-scaff_25466_1-p15589 (80.094)	79.9–83.7	2.54	3.97	334.00	2015b	cis-eQTL
CG1 expression	***qCG1A8***[j]	***A8***	***73.71***	***Bn-A08-p20855631 (73.676)***	***73.1–74.7***	***3.37***	***7.02***	***–80.35***	***2015b***	***trans-eQTL***
ASK2|FIL×KNATM expression	*cqASK2A4*[h]	A4	11.61	Bn-A04-p3820381 (11.625)	10.2–17.8	3.07	7.17	1.23	2015b	trans-eQTL
CG1|FIL×TPL expression	*cqCG1A8-1*	A8	62.31	Bn-A08-p19642040 (62.239)	59.3–63.8	3.31	7.44	80.14	2015b	trans-eQTL
	qCG1A8-2[j]	***A8***	***73.71***	***Bn-A08-p20855631 (73.676)***	***73.1–74.7***	***3.45***	***7.23***	***–79.27***	***2015b***	***trans-eQTL***
CG1|IFL×SEP2 expression	*cqCG1A3*	A3	54.81	Bn-scaff_17298_1-p705887 (54.794)	48.2–61.0	3.24	8.37	63.50	2015b	trans-eQTL
	cqCG1C9	C9	4.81	CB10103 (4.82)	3.0–5.8	4.58	12.26	79.61	2015b	trans-eQTL
CG1|UFO expression	*cqCG1A9*	A9	124.51	Bn-scaff_16389_1-p578073 (124.527)	121.7–125.8	3.65	8.32	37.00	2015b	trans-eQTL
TPL|IFL×SEP2 expression	*cqTPLA3-1*	A3	54.51	Bn-A03-p16431100 (54.482)	52.0–59.3	6.12	16.99	−477.35	2015b	trans-eQTL
	cqTPLC7	C7	122.31	Bn-scaff_16110_1-p3700752 (122.401)	119.3–122.6	3.10	6.97	−228.88	2015b	trans-eQTL
TPL|KNATM×SEP2 expression	*cqTPLA3-1*	A3	54.51	Bn-A03-p16431100 (54.482)	50.7–60.0	5.27	14.28	−556.66	2015b	trans-eQTL

The definitions of a–d and f-h are the same as in ***Table 4****.* [e]*2015b represents the second environment in which the eQTLs were detected.* [i]*The eQTL is at the same approximate position with qPD.A9-1 identified in the previous study for PDgr.* [j]*The eQTLs are at the same position with each other.*

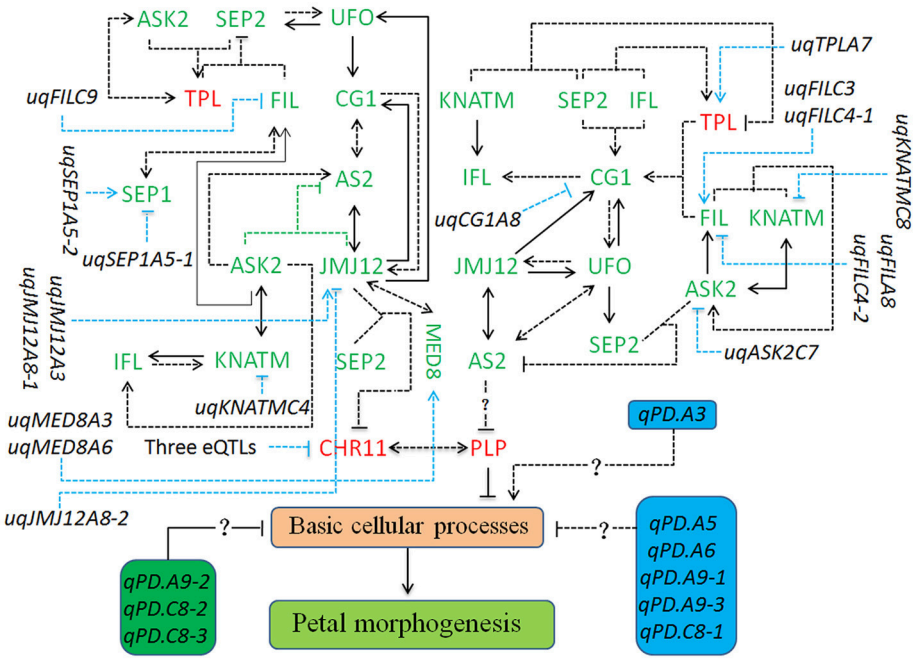

FIGURE 4 | The regulatory network involved in the petal development of apetalous "APL01." The regulatory network mainly consists of the *CHR11-PLP* pathway, 12 unconditional eQTLs of TGs, and nine QTLs for PDgr. The *CHR11-PLP* pathway contains 29 tQTTs and 12 unconditional eQTLs, representing 41 kinds of regulatory relationships. The three eQTLs negatively regulating *CHR11* expression are *uqCHR11C3*, *uqCHR11C4-1,* and *uqCHR11C4-2*. Genes marked in red are up-regulated, while genes marked in green are down-regulated in "APL01" compared with those in "Holly." Arrows represent the positive regulation of tQTTs for the downstream TGs, while blunted lines represent the negative regulation of tQTTs for the downstream TGs. Arrows or blunted solid lines marked indicate the regulatory relationships repeatedly detected in all two environments, while arrows or blunted dotted lines indicate the regulatory relationships only detected in one environment. In addition, there may be the *AS2-PLP* pathway regulating *PLP* expression, and this pathway consists of 21 tQTTs and 8 unconditional eQTLs, representing 29 kinds of regulatory relationships.

expression. For the *JMJ12* expression level, there are two positive closed regulatory circuits, in which *JMJ12–CG1–AS2–JMJ12* is a bidirectional circuit while *JMJ12–UFO–CG1–AS2–JMJ12* is a unidirectional circuit. Moreover, two unconditional eQTLs (*uqJMJ12A3* and *uqJMJ12A8-1*) with positive effects and the repressive *uqJMJ12A8-2* also participated in the regulation of *JMJ12* expression. Additionally, *JMJ12* positively regulates *MED8*. In the *JMJ12–CG1–AS2–JMJ12* circuit, *AS2* was also regulated by the promoter *ASK2*, and the transcript epistatic loci (*ASK2*×*JMJ12*) had a negative effect. In addition, *ASK2* was positively regulated by *ASK2–KNATM–IFL–ASK2*, a unidirectional circuit. In the *JMJ12–UFO–CG1–AS2–JMJ12* circuit, the *UFO* expression level was also regulated by the promoter *SEP2*, while *SEP2* expression level was attributed to the integrated regulation of the promoter *UFO* and the transcripts of the epistatic loci (*FIL*×*TPL*) had a negative effect. Furthermore, *FIL* was regulated by both activators (*ASK2* and *SEP1*) and the repressive *uqFILC9*, while *TPL* was positively regulated by both the activator *ASK2* and the transcript epistatic loci (*ASK2*×*SEP2*), which had a positive effect. Finally, the regulatory effects of the tQTTs and unconditional eQTLs were integrated into the expression level of *PLP* and then prevented the basic cellular processes responsible for petal morphogenesis by up-regulating *PLP* (**Figure 4**).

In addition to *CHR11*, *PLP* expression level may be also regulated by the suppressor *AS2* (**Figure 4**). However, the regulation of *PLP* by *AS2* probably requires gene other than the above 14 petal regulators. The *AS2* expression level was attributed to the integrated regulation of multi-factors containing 21 tQTTs and 8 unconditional eQTLs (**Figure 4**).

DISCUSSION

There are a large number of upstream regulators involved in petal development in *Arabidopsis* (Zik and Irish, 2003; Kaufmann et al., 2009, 2010; Wuest et al., 2012). In a previous study (Yu et al., 2016), 36 petal regulators and several candidate genes involved in the regulation of the apetalous trait in *B. napus* were identified. However, how these genes collaboratively regulate petal development in both *Arabidopsis* and *B. napus* is unclear. In this study, we determined that 14 TGs participate in the regulation of apetalous characteristic in "APL01" by analyzing the expression patterns of 37 petal regulators in "APL01," "PL01," and "Holly." The same slopes of the standard curvesof 14 TGs and the endogenous reference gene *ACTIN* indicated the same amplification efficiency. Thus, the use of qRT-PCR in the AH population is dependable (Yin et al., 2010).

From the Pearson correlation coefficients, the similarity level of PDgr in the AH population is high between the two

environments ($r = 0.806$) but not completely the same, which can probably be ascribed to unknown environmental effects (Wang et al., 2015). The similarities of the TG expression patterns in the AH population are poor between the two environments ($r < 0.8$), except for five TGs, which congruously explains the variation in PDgr between the two environments. The correlation analyses between the 14 TGs and PDgr determined that only a few TGs were significantly correlated ($P < 0.05$) with PDgr in the two environments, implying that only a few genes were directly related to petal development. In fact, several previous researches have suggested that many transcriptional regulators indirectly regulate petal development in one way or another (Zik and Irish, 2003; Kaufmann et al., 2009, 2010; Wuest et al., 2012). However, a linear correlation analysis failed to discover the intricate relationships between genes and petal morphogenesis.

QTT-association mapping, based on a mixed linear model, is mainly used to analyze complex traits (Zhang et al., 2015). A QTT analysis of PDgr determined that *PLP* acts as the major negative QTT of PDgr in the two environments, indicating that *PLP* negatively regulates petal development in *B. napus*. In *Arabidopsis*, *PLP* encodes the alpha-subunit shared between protein farnesyltransferase and protein geranylgeranyltransferase-I (Running et al., 2004). *plp* mutant leads to dramatically enlarged meristems and increased floral organ number (Running et al., 2004). Based on the high degree of chromosomal colinearity between *B. napus* and *Arabidopsis* (Chalhoub et al., 2014), it is very likely that *BnPLP* plays the same role in regulating petal development as *AtPLP*. Except for *PLP,* the remaining TGs were not significantly associated with PDgr, suggesting that these TGs potentially participate in petal development of rapeseed by regulating *PLP* expression.

The QTT mapping of *PLP* expression levels showed that *CHR11* was positively associated with *PLP* in the first environment, indicating that *CHR11* acts as a positive regulator of *PLP* expression. However, we can not detect the effect of *CHR11* on *PLP* in the second environment, implying that *CHR11* regulates *PLP* expression in an environment dependent way. Previous reports suggested that *CHR11* encoded a SWI2/SNF2 chromatin remodeling protein belonging to the ISWI family that was involved in the epigenetic regulation of eukaryotic genes (Li et al., 2012, 2015). In the second environment, the effect of *CHR11* on *PLP* may be too weak to detect by QTT-association mapping because of some unqualified environmental conditions. By analogy, QTT mapping for the remaining TGs detected 38 tQTTs, associated with 13 TGs, and 31 tQTTs, associated with 10 TGs in the first and second environment, respectively. A total of 10 tQTTs can be repeatedly detected in the two environments, implying that these regulatory relationships may occur *in vivo*, as well as being required for petal development in *B. napus*. In addition, the detection of some tQTTs in one environment might be the result of the different expression patterns of TGs between two environments. Meanwhile, these tQTTs may act as the decisive factors that give rise to variable PDgr between the two environments because gene expression' diversity is a vital mechanism underlying phenotypic diversity among individuals (Yin et al., 2010). Thus, the different tQTTs between the two environments are also required for petal development.

For the molecular functions of QTT or tQTT, *PLP*, *CHR11*, and *FIL*×*TPL*, respectively, acted as a repressor of PDgr, an activator of *PLP*, and a repressor of *SEP2* in the first environment, which echoes previous studies in *Arabidopsis* that suggested that *PLP* (Running et al., 2004), *CHR11* (Li et al., 2012) and *TPL* (Krogan et al., 2012) acted as repressors regulating petal development. There are mostly positive regulatory relationships between the remaining 10 TGs in the first environment, which supports our recent inference that the 10 TGs play positive roles in petal development in *B. napus* (Yu et al., 2016). In the second environment, the regulatory signals of the tQTTs are finally integrated into the expression level of *AS2* and may have then negatively regulated *PLP* expression by regulating some intermediate regulators (**Figure 4**); however, we cannot detect the negative effect of *AS2* on *PLP* because only a limited number of genes are included in the present study. Although the regulatory relationships among TGs presented in this study need to be verified through more molecular experiments, these relationships are logically possible. For example, *UFO*, as an essential component of the SCF complex that is a key ubiquitin E3 ligase (Skowyra et al., 1997), is involved in both floral meristem and floral organ development in *Arabidopsis* (Levin and Meyerowitz, 1995). In this study, *UFO* probably regulates the expression of *SEP2*, *CG1,* and *JMJ12* in a LEAFY-dependent manner, just like it regulates *AP3* transcription in *Arabidopsis* (Chae et al., 2008). Moreover, *CG1* as a candidate gene in the CI of *qPD.C8-2* regulating the apetalous trait in line APL01 functions upstream of the *CHR11-PLP* pathway, implying that *qPD.C8-2* potentially regulates the petal development of line APL01 through the *CHR11-PLP* pathway.

Unconditional eQTL mapping of TG expression levels in the AH population determined that only a few unconditional eQTLs were obtained for the TGs in two environments, and that all of the unconditional eQTLs were minor QTLs ($R^2 < 20\%$) (Shi et al., 2009). Thus, the strength of TG expression levels was mainly ascribed to effects of tQTTs. Based on the description for trans-eQTLs (Kliebenstein, 2008; Sasayama et al., 2012), all of the unconditional eQTLs presently identified are trans-eQTL, except for *uqTPLA7*, which indicates that most of the unconditional eQTLs act as transcription factors or transcriptional coactivators of the corresponding TGs. *uqSEP1A9*, a trans-eQTL identified in the second environment, overlapped a QTL (*qPD.A9-1*) for PDgr (Wang et al., 2015), indicating that the PDgr and *SEP1* expression were causally related (Thumma et al., 2001) and that the *qPD.A9-1* potentially participated in the regulation of PDgr by regulating *SEP1* expression. Furthermore, the colocalization of TG expression levels may reflect the pleiotropism of a genomic region (QTL), such as *JMJ12* and *MED8* in the first assay.

In addition, conditional eQTL mapping of TGs determined that the unconditional eQTLs were lost, except for *uqCG1A8* (*cqCG1A8-2*), in the two environments, implying that the effects of those unconditional eQTLs were completely attributed to the upstream tQTTs regarded as the conditional independent variables (Zhu, 1995). In other words, the effects of those unconditional eQTLs were passed from tQTTs to the corresponding downstream TGs, indicating that there is a relationship between the tQTT and the corresponding

downstream TG, indirectly verifying the likelihood of tQTTs regulating TGs' expression levels (Zhu, 1995). Compared with the unconditional eQTLs, almost all of the conditional eQTLs are novel, indicating that these conditional eQTLs are generally suppressed by the corresponding upstream tQTTs under an unconditional situation (Zhu, 1995). That is, the upstream tQTTs participate in the positive regulation of TG expression by repressing the corresponding conditional eQTLs (Zhu, 1995). Thus, the conditional eQTLs should have negative effects on the TGs' expression, which is consistent with the results of conditional eQTL mapping in this study. Unconditional eQTL coupled with conditional QTL mapping indirectly verifies that the tQTTs detected in this study are valid.

The relationships among TGs and PDgr are presented in **Figure 4**. The apetalous characteristic of "APL01" is not only attributed to the regulators identified in this study, but it is possible that the aforementioned TGs participate in the petal development of "APL01" in the manner described in **Figure 4**. Although these regulatory relationships need to be further verified, our findings provided a basis for solving the puzzle of petal development in *B. napus*.

AUTHOR CONTRIBUTIONS

KY and XW co-wrote the first draft of the manuscript. JZ and RG designed the project, acquired funding, and finalized the manuscript. KY and SC collected the young inflorescences of the AH population. KY and XW performed the qRT-PCR assays. QP and FC performed total RNA exaction. HL and WZ performed the first strand cDNA synthesis. SF and MH collected and processed the data used in this study. WL and PC assisted and analyzed the data. All authors have reviewed and approved the final version of the manuscript and therefore are equally responsible for the integrity and accuracy of its content.

ACKNOWLEDGMENTS

The work was supported by National Key Research and Development Program of China (2016YFD0100202), National Natural Science Foundation of China (31371660, 31601334), the Industry Technology System of Rapeseed in China (CARS-12), Natural Science Foundation of Jiangsu Province (BK20160578), Jiangsu Agriculture Science and Technology Innovation Fund (CX(14)5011), and Jiangsu Collaborative Innovation Center for Modern Crop Production.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00089/full#supplementary-material

Figure S1 | The scatter plot diagram of the normalized expression levels of TGs across the AH population in two environments.

Table S1 | List of the primers used for qRT-PCR assays in this study.

Table S2 | The expression patterns of 37 petal regulators in RNA-seq and qRT-PCR assays.

Table S3 | The regression formulas of 14 TGs' amplification reactions in the standard cDNA.

Table S4 | Correlation analysis between two biological replicates of TG expression within an environment.

Table S5 | Highly significant QTTs and tQTTs for PDgr and TGs detected in the first environment.

Table S6 | Highly significant QTTs and tQTTs for PDgr and TGs detected in the second environment.

Table S7 | Comparison between QTLs for PDgr and eQTLs for TGs unconditional and conditional expression levels in the first environment.

Table S8 | Comparison between QTLs for PDgr and eQTLs for TGs unconditional and conditional expression levels in the second environment.

REFERENCES

Battle, A., and Montgomery, S. B. (2014). Determining causality and consequence of expression quantitative trait loci. *Hum. Genet.* 133, 727–735. doi: 10.1007/s00439-014-1446-0

Bowman, J. L., Smyth, D. R., and Meyerowitz, E. M. (1991). Genetic interactions among floral homeotic genes of *Arabidopsis*. *Development* 112, 1–20.

Chae, E., Tan, Q. K., Hill, T. A., and Irish, V. F. (2008). An *Arabidopsis* F-box protein acts as a transcriptional co-factor to regulate floral development. *Development* 135, 1235–1245. doi: 10.1242/dev.015842

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

Chapman, J. F., Daniels, R. W., and Scarisbrick, D. H. (1984). Field studies on ^{14}C assimilate fixation and movement in oil-seed rape (*B. napus*). *J. Agric. Sci.* 102, 23–31.

Chen, G., Zhang, F., Xue, W., Wu, R., Xu, H., Wang, K., et al. (2016). An association study revealed substantial effects of dominance, epistasis and substance dependence co-morbidity on alcohol dependence symptom count. *Addict. Biol.* 22, 1475–1485. doi: 10.1111/adb.12402

Cheung, V. G., Conlin, L. K., Weber, T. M., Arcaro, M., Jen, K., Morley, M., et al. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* 33, 422–425. doi: 10.1038/ng1094

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963–971.

Cui, F., Li, J., Ding, A., Zhao, C., Wang, L., Wang, X., et al. (2011). Conditional QTL mapping for plant height with respect to the length of the spike and internode in two mapping populations of wheat. *Theor. Appl. Genet.* 122, 1517–1536. doi: 10.1007/s00122-011-1551-6

Damerval, C., and Becker, A. (2017). Genetics of flower development in *Ranunculales*–a new, basal eudicot model order for studying flower evolution. *New Phytol.* 216, 361–366. doi: 10.1111/nph.14401

de Martino, G., Pan, I., Emmanuel, E., Levy, A., and Irish, V. F. (2006). Functional analyses of two tomato *APETALA3* genes demonstrate diversification in their roles in regulating floral development. *Plant Cell* 18, 1833–1845. doi: 10.1105/tpc.106.042978

DeCook, R., Lall, S., Nettleton, D., and Howell, S. H. (2005). Genetic regulation of gene expression during shoot development in *Arabidopsis*. *Genetics* 172, 1155–1164. doi: 10.1534/genetics.105.042275

Ditta, G., Pinyopich, A., Robles, P., Pelaz, S., and Yanofsky, M. F. (2004). The *SEP4* gene of *Arabidopsis thaliana* functions in floral organ and meristem identity. *Curr. Biol.* 14, 1935–1940. doi: 10.1016/j.cub.2004.10.028

Drea, S., Hileman, L. C., de Martino, G., and Irish, V. F. (2007). Functional analyses of genetic pathways controlling petal specification in poppy. *Development* 134, 4157–4166. doi: 10.1242/dev.013136

Fray, M. J., Puangsomlee, P., and Goodrich, J. (1997). The genetics of stamenoid petal production in oilseed rape (*Brassica napus*) and equivalent variation in *Arabidopsis thaliana*. *Theor. Appl. Genet.* 94, 731–736.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., et al. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* 2:e130. doi: 10.1371/1journal.pgen.0020130

Göring, H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. A., et al. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* 39, 1208–1216. doi: 10.1038/ng2119

Hileman, L. C., and Irish, V. F. (2009). More is better: the uses of developmental genetic data to reconstruct perianth evolution. *Am. J. Bot.* 96, 83–95. doi: 10.3732/ajb.0800066

Hirano, H. Y., Tanaka, W., and Toriba, T. (2014). "Grass flower development," in *Flower Development*, eds J. L. Riechmann and F. Wellmer (New York, NY: Springer Science+Business Media), 57–84.

Jack, T. (2001). Relearning our ABCs: new twists on an old model. *Trends Plant Sci.* 6, 310–316. doi: 10.1016/S1360-1385(01)01987-2

Jamaux, D. I., and Spire, D. (1999). Comparison of responses of ascospores and mycelium by ELISA with anti-mycelium and anti-ascospore antisera for the development of a method to detect *Sclerotinia sclerotiorum* on petals of oilseed rape. *Ann. Appl. Biol.* 134, 171–179. doi: 10.1111/j.1744-7348.1999.tb05253.x

Jansen, R. C., and Nap, J. P. (2001). Genetical genomics: the added value from segregation. *Trends Genet.* 17, 388–391. doi: 10.1016/S0168-9525(01)02310-1

Jordan, M. C., Somers, D. J., and Banks, T. W. (2007). Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. *Plant Biotechnol. J.* 5, 442–453. doi: 10.1111/j.1467-7652.2007.00253.x

Kaufmann, K., Muiño, J. M., Jauregui, R., Airoldi, C. A., Smaczniak, C., Krajewski, P., et al. (2009). Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the *Arabidopsis* flower. *PLoS Biol.* 7:e1000090. doi: 10.1371/journal.pbio.1000090

Kaufmann, K., Wellmer, F., Muiño, J. M., Ferrier, T., Wuest, S. E., Kumar, V., et al. (2010). Orchestration of floral initiation by *APETALA1*. *Science* 328, 85–89. doi: 10.1126/science.1185244

Kelly, A., Fray, M., Arthur, E. A., Lydiate, D. J., and Evans, E. J. (1995). "The genetic control of petalless flowers and upright pods," in *Proc. 9th Int. Rapeseed Congr.* (Cambridge), 4–8.

Kim, S., Yoo, M. J., Albert, V. A., Farris, J. S., Soltis, P. S., and Soltis, D. E. (2004). Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication. *Am. J. Bot.* 12, 2102–2018. doi: 10.3732/ajb.91.12.2102

Kliebenstein, D. (2008). Quantitative Genomics: analyzing intraspecific variation using global gene expression polymorphisms or eQTL. *Annu. Rev. Plant Biol.* 60, 93–114. doi: 10.1146/annurev.arplant.043008.092114

Kramer, E. M., Dorit, R. L., and Irish, V. F. (1998). Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-Box gene lineages. *Genetics* 2, 765–783.

Kramer, E. M., Holappa, L., Gould, B., Jaramillo, M. A., Setnikov, D., and Santiago, P. M. (2007). Elaboration of B gene function to include the identity of novel floral organs in the lower eudicot *Aquilegia*. *Plant Cell* 19, 750–766. doi: 10.1105/tpc.107.050385

Krogan, N. T., Hogan, K., and Long, J. A. (2012). APETALA2 negatively regulates multiple floral organ identity genes in *Arabidopsis* by recruiting the co-repressor TOPLESS and the histone deacetylase HDA19. *Development* 139, 4180–4190. doi: 10.1242/dev.085407

Lamb, R. S., and Irish, V. F. (2003). Functional divergence within the *APETALA3/PISTILLATA* floral homeotic gene lineages. *Proc. Natl. Acad. Sci. U.S.A.* 100, 6558–6563. doi: 10.1073/pnas.0631708100

Levin, J. Z., and Meyerowitz, E. M. (1995). *UFO*: an *Arabidopsis* gene lnvolved in both floral meristem and floral organ development. *Plant Cell* 7, 529–548.

Li, C., Chen, C., Gao, L., Yang, S., Nguyen, V., Shi, X., et al. (2015). The *Arabidopsis* SWI2/SNF2 chromatin remodeler BRAHMA regulates polycomb function during vegetative development and directly activates the flowering repressor gene SVP. *PLoS Genet.* 11:e1004944. doi: 10.1371/journal.pgen.1004944

Li, G., Zhang, J., Li, J., Yang, Z., Huang, H., and Xu, L. (2012). ISWI chromatin remodeling factors and their interacting RINGLET proteins act together in controlling the plant vegetative phase in *Arabidopsis*. *Plant J.* 72, 261–270. doi: 10.1111/j.1365-313X.2012.05074.x

Li, H., Liang, W., Yin, C., Zhu, L., and Zhang, D. (2011). Genetic interaction of OsMADS3, DROOPING LEAF, and OsMADS13 in specifying rice floral organ identities and meristem determinacy. *Plant Physiol.* 156, 263–274. doi: 10.1104/pp.111.172080

Li, Y., Alvarez, O. A., Gutteling, E. W., Tijsterman, M., Fu, J., Riksen, J. A., et al. (2006). Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet.* 2:e222. doi: 10.1371/journal.pgen.0020222

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

Mendham, N. J., and Rao, M. (1991). "The apetalous flower character as a component of a high yielding ideotype," in *Proceedings of the 8th International Rapeseed Congress*. (Saskatoon, SK).

Morrall, C. (1996). Immunofluorescent staining of sclerotinia ascospores on canola petals. *Can. J. Plant Pathol.* 18, 237–241. doi: 10.1080/07060669609500618

Pelaz, S., Ditta, G. S., and Baumann, E. (2000). B and C floral organ identity functions require *SEPALLATA* MADS-box genes. *Nature* 6783, 200–203. doi: 10.1038/35012103

Polowick, P. L., and Sawhney, V. K. (1986). A scanning electron microscopic study on the initiation and development of floral organs of *Brassica napus* (Cv. *Westar). Am. J. Bot.* 2, 254–263. doi: 10.2307/2444180

Rao, M. S. S., Mendham, N. J., and Buzza, G. C. (1991). Effect of the apetalous flower characteristic on radiation distribution in the crop canopy, yield and its components in oilseed rape (*Brassica napus* L.). *J. Agric. Sci.* 117, 189–196.

Rijpkema, A. S., Royaert, S., and Zethof, J. (2006). Analysis of the Petunia TM6 MADS box gene reveals functional divergence within the DEF/AP3 lineage. *Plant Cell* 18, 1819–1832. doi: 10.1105/tpc.106.042937

Running, M. P., Lavy, M., Sternberg, H., Galichet, A., Gruissem, W., Hake, S., et al. (2004). Enlarged meristems and delayed growth in plp mutants result from lack of CaaX prenyltransferases. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7815–7820. doi: 10.1073/pnas.0402385101

Sasayama, D., Hori, H., Nakamura, S., Miyata, R., Teraishi, T., Hattori, K., et al. (2012). Identification of single nucleotide polymorphisms regulating peripheral blood mRNA expression with Genome-Wide Significance: an eQTL study in the Japanese population. *PLoS ONE* 8:e54967. doi: 10.1371/journal.pone.0054967

Schwarz-Sommer, Z., Huijser, P., and Nacken, W. (1990). Genetic control of flower development by homeotic genes in *Antirrhinum majus*. *Science* 4983, 931–936. doi: 10.1126/science.250.4983.931

Shi, J., Li, R., Qiu, D., Jiang, C., Long, Y., Morgan, C., et al. (2009). Unraveling the complex trait of crop yield with quantitative trait loci mapping in *Brassica napus*. *Genetics* 182, 851–861. doi: 10.1534/genetics.109.101642

Skowyra, D., Craig, K. L., Tyers, M., Elledge, S. J., and Harper, J. W. (1997). F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell* 91, 209–219. doi: 10.1016/S0092-8674(00)80403-1

Smaczniak, C., Immink, R. G. H., Muiño, J. M., Blanvillain, R., Busscher, M., Busscher-Lange, J., et al. (2012). Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1560–1565. doi: 10.1073/pnas.1112871109

Smyth, D. R., Bowman, J. L., and Meyerowitz, E. M. (1990). Early flower development in *Arabidopsis*. *Plant Cell* 2, 755–767.

Sun, W., Bennett, V. C., Eggins, S. M., Kamenetsky, V. S., and Arculus, R. J. (2003). Enhanced mantle-to-crust rhenium transfer in undegassed arc magmas. *Nature* 422, 294–297. doi: 10.1038/nature01482

Theissen, G., and Saedler, H. (2001). Plant biology. Floral quartets. *Nature* 6819, 469–471. doi: 10.1038/35054172

Thumma, B. R., Naidu, B. P., Chandra, A., Cameron, D. F., Bahnisch, L. M., and Liu, C. (2001). Identification of causal relationships among traits related to drought resistance in *Stylosanthes scabra* using QTL analysis. *J. Exp. Bot.* 52, 203–214. doi: 10.1093/jexbot/52.355.203

van der Krol, A. R., and Chua, N. (1993). Flower development in *Petunia*. *Plant Cell* 5, 1195–1203. doi: 10.1105/tpc.5.10.1195

Vandenbussche, M., Zethof, J., Royaert, S., Weterings, K., and Gerats, T. (2004). The duplicated B-class heterodimer model: whorl-specific effects and complex

genetic interactions in *Petunia* hybrida flower development. *Plant Cell* 16, 741–754. doi: 10.1105/tpc.019166

Wang, S., Basten, C. J., and Zeng, Z. B. (2007). *Windows QTL Cartographer 2.5*. Department of Statistics, North Carolina State University, Raleigh, NC.

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

Wang, Y., Han, Y., Teng, W., Zhao, X., Li, Y., Wu, L., et al. (2014). Expression quantitative trait loci infer the regulation of isoflavone accumulation in soybean (*Glycine max* L. *Merr.)* seed. *BMC Genomics* 15:680. doi: 10.1186/1471-2164-15-680

Wuest, S. E., O'Maoileidigh, D. S., Rae, L., Kwasniewska, K., Raganelli, A., Hanczaryk, K., et al. (2012). Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. *Proc. Natl. Acad. Sci. U.S.A.* 109, 13452–13457. doi: 10.1073/pnas.1207075109

Yates, D. J., and Steven, M. D. (1987). Reflexion and absorption of solar radiation by flowering canopies of oil-seed rape (*Brassica napus* L.). *J. Agric. Sci.* 109, 495–502.

Yin, Z., Meng, F., Song, H., Wang, X., Xu, X., and Yu, D. (2010). Expression quantitative trait loci analysis of two genes encoding rubisco activase in soybean. *Plant Physiol.* 153, 1625–1637. doi: 10.1104/pp.109.148312

Yu, K., Wang, X., Chen, F., Chen, S., Peng, Q., Li, H., et al. (2016). Genome-wide transcriptomic analysis uncovers the molecular basis underlying early flowering and apetalous characteristic in *Brassica napus* L. *Sci. Rep.* 6:30576. doi: 10.1038/srep30576

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* 136, 1457–1468.

Zhang, F. T., Zhu, Z. H., Tong, X. R., Zhu, Z. X., Qi, T., and Zhu, J. (2015). Mixed linear model approaches of association mapping for complex traits based on omics variants. *Sci. Rep.* 5:0298. doi: 10.1038/srep10298

Zhang, Y., Wang, X., Zhang, W., Yu, F., Tian, J., Li, D., et al. (2011). Functional analysis of the two *Brassica AP3* genes involved in apetalous and stamen carpelloid phenotypes. *PLoS ONE* 6:e20930. doi: 10.1371/journal.pone.0020930

Zhang, Z., Liu, Z., Cui, Z., Hu, Y., Wang, B., and Tang, J. (2013). Genetic analysis of grain filling rate using conditional QTL mapping in maize. *PLoS ONE* 8:e56344. doi: 10.1371/journal.pone.0056344

Zhao, J., Becker, H. C., Zhang, D., Zhang, Y., and Ecke, W. (2006). Conditional QTL mapping of oil content in rapeseed with respect to protein content and traits related to plant development and grain yield. *Theor. Appl. Genet.* 113, 33–38. doi: 10.1007/s00122-006-0267-5

Zhou, L., Liu, S., Wu, W., Chen, D., Zhan, X., Zhu, A., et al. (2016). Dissection of genetic architecture of rice plant height and heading date by multiple-strategy-based association studies. *Sci. Rep.* 6:29718. doi: 10.1038/srep29718

Zhu, J. (1995). Analysis of conditional genetic effects and variance components in developmental genetics. *Genetics* 141, 1633–1639.

Zik, M., and Irish, V. F. (2003). Global identification of target genes regulated by APETALA3 and PISTILLATA floral homeotic gene action. *Plant Cell* 15, 207–222. doi: 10.1105/tpc.006353

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