Genome-Wide Diversity of MADS-Box Genes in Bread Wheat is Associated with its Rapid Global Adaptability

Qasim Raza1*, Awais Riaz1, Rana Muhammad Atif2,3,4, Babar Hussain5,6, Iqrar Ahmad Rana3,7, Zulfiquar Ali8, Hikmet Budak9 and Ibrahim A. Alaraidh10*

1Molecular Breeding Laboratory, Rice Research Institute, Kala Shah Kaku, Pakistan, 2Department of Plant Breeding and Genetics, University of Agriculture Faisalabad, Faisalabad, Pakistan, 3Centre for Advanced Studies in Agriculture and Food Security, University of Agriculture Faisalabad, Faisalabad, Pakistan, 4Precision Agriculture and Analytics Lab, National Centre for Big Data and Cloud Computing, University of Agriculture Faisalabad, Faisalabad, Pakistan, 5Department of Biological Sciences, Middle East Technical University, Ankara, Turkey, 6Department of Biotechnology, Faculty of Life Sciences, University of Central Punjab, Lahore, Pakistan, 7Centre of Agricultural Biochemistry and Biotechnology, University of Agriculture Faisalabad, Faisalabad, Pakistan, 8Institute of Plant Breeding and Biotechnology, Muhammad Nawaz Sharif University of Agriculture, Multan, Pakistan, 9Montana BioAgriculture, Inc., Missoula, MT, United States, 10Department of Botany and Microbiology, College of Science, King Saud University, Riyadh, Saudi Arabia

OPEN ACCESS

Edited by: Reyazul Rouf Mir, Sher-e-Kashmir University of Agricultural Sciences and Technology, India

Reviewed by: Brijesh Kumar, Patanjali Research Institute, India, Soleyman Dayani, Imam Khomeini International University, Iran

*Correspondence: Qasim Raza qasimnazami@gmail.com
Ibrahim A. Alaraidh ialaraidh@ksu.edu.sa

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

Received: 20 November 2021
Accepted: 21 December 2021
Published: 17 January 2022

Citation:
Raza Q, Riaz A, Atif RM, Hussain B, Rana IA, Ali Z, Budak H and Alaraidh IA (2022) Genome-Wide Diversity of MADS-Box Genes in Bread Wheat is Associated with its Rapid Global Adaptability. Front. Genet. 12:818880. doi: 10.3389/fgene.2021.818880

MADS-box gene family members play multifarious roles in regulating the growth and development of crop plants and hold enormous promise for bolstering grain yield potential under changing global environments. Bread wheat (Triticum aestivum L.) is a key stable food crop around the globe. Until now, the available information concerning MADS-box genes in the wheat genome has been insufficient. Here, a comprehensive genome-wide analysis identified 300 high confidence MADS-box genes from the publicly available reference genome of wheat. Comparative phylogenetic analyses with Arabidopsis and rice MADS-box genes classified the wheat genes into 16 distinct subfamilies. Gene duplications were mainly identified in subfamilies containing unbalanced homeologs, pointing towards a potential mechanism for gene family expansion. Moreover, a more rapid evolution was inferred for M-type genes, as compared with MIKC-type genes, indicating their significance in understanding the evolutionary history of the wheat genome. We speculate that subfamily-specific distal telomeric duplications in unbalanced homeologs facilitate the rapid adaptation of wheat to changing environments. Furthermore, our in-silico expression data strongly proposed MADS-box genes as active guardians of plants against pathogen insurgency and harsh environmental conditions. In conclusion, we provide an entire complement of MADS-box genes identified in the wheat genome that could accelerate functional genomics efforts and possibly facilitate bridging gaps between genotype-to-phenotype relationships through fine-tuning of agronomically important traits.

Keywords: genome-wide analysis, in-silico expression, TaMADS’s, transcription factors, Triticum aestivum, wheat adaptability
INTRODUCTION

In the fight for global food security and safety, bread wheat represents one of the largest contributing grain crops. It is one of the most produced, stored, and consumed food crops worldwide and a major source of energy and nutrients in the developing countries (Igrejas and Branlard, 2020). However, following the green revolution, improvement of wheat grain production has been hindered by various bottlenecks. This includes, but is not limited to, the non-availability of a reliable and fully annotated reference genome (Brenchley et al., 2012; Lukaszewski et al., 2014; Chapman et al., 2015; Clavijo et al., 2017; Zimin et al., 2017). The slow progress in harnessing a fully annotated reference genome was mainly due to the genome’s allohexaploid nature, which comprises three closely related, but independently maintained, sub-genomes, known as A, B, and D. This, in combination with a high frequency of repetitive sequences, particularly hindered progress. Recently, an alliance of geneticists combined resources over 13 years, to produce a fully annotated sequence of the wheat genome, which had a resultant size of ~17 Gbps. This is the largest known genome among crop plants (Appels et al., 2018). To add to this, the high-quality of the reference genome, together with large scale RNA-seq data and expression repositories (Borrill et al., 2016; Ramirez-González et al., 2018), provide a rich resource for studying evolutionary dynamics and functional characterization of important gene families, which in turn could facilitate crop improvement efforts.

MADS-box transcription factors (TFs) are a well-documented group of genes known for playing vital roles in regulating the growth and development of several important plant species (Smaczniak et al., 2012; Ali et al., 2019). These genes influence diverse biological functions, including cell development, signal transduction, biotic and abiotic stress responses, vegetative organs development, control of flowering and anthesis time, formation of meristems and flower organs, ovule development, embryo development, dehiscence zone formation, and ripening of fruits and seeds (Colombo et al., 1995; Rounsley et al., 1995; Kuo et al., 1997; Riechmann and Meyerowitz, 1997; Alvarez-Buylla et al., 2000; Samach et al., 2000; Saedler et al., 2001; Moore et al., 2002; Messenguy and Dubois, 2003; Pařenová et al., 2003; Smaczniak et al., 2012; Ali et al., 2019). Given their importance, identification, and characterization of MADS-box genes in agriculturally important species are critical for crop improvement and fine-tuning of specific traits through genetic exploitations.

All MADS-box genes contain a highly conserved MADS (M) domain of approximately 58–60 amino acids. The M domain enables DNA binding and is in the N-terminal region of the protein (Yanofsky et al., 1990). Members of this gene family are classified into M-type I and MIKC-type II super clades, which were generated after an ancient gene duplication event that occurred before the separation of animal and plant lineages (Becker and Theißen, 2003). M-type genes have simple intron-exon structures (zero or one intron), only contain a single M domain, and encode Serum Response Factor (SRF)-like proteins. These can be subcategorized into four clades (Ma, Mc, Mδ, My, and Mδ). However, the Mδ clade genes closely resemble those within the MIKC* group, as previously reported (De Bodt et al., 2003).

In comparison, MIKC-type genes have four domains: 1) a highly conserved DNA binding M domain, 2) a less conserved Intervening (I) domain of ~30 AA involved in dimer formation, 3) a moderately conserved Keratin (K) domain of ~70 AA which regulates heterodimerization of MADS proteins, and lastly 4) a highly inconstant C-terminal region which contributes to transcriptional regulation and higher-order protein complex formation (Henschel et al., 2002; Díaz-Riquelme et al., 2009). MIKC-type genes encode Myocyte Enhancer Factor 2 (MEF2)-like proteins and are categorized into MIKC* and MIKC* clades (Kaufmann et al., 2005). MIKC*-type genes have duplicated K domains and relatively longer I domains than MIKC* -types (Duan et al., 2014). MIKC-type genes can be further categorized into several subclasses based on their phylogenetic relationships in flowering plants (Gramażow and Theißen, 2015).

Like what has been observed in other model and crop plant species, in wheat MADS-box genes are known to confer drought tolerance through regulation of drought tolerance genes and micro-RNAs (Budak et al., 2015). In addition to this, TaMADS51, TaMADS4, TaMADS5, TaMADS6, and TaMADS18 were showed up-regulation, where TaMADAGL17, TaMADAGL2, TaMADWM31c, and TaMADS14 were downregulated, under phosphorous (P) starvation, and further functional analysis confirmed their role in P-deficient stress responses (Shi et al., 2016). In another study, several MADS-box genes were activated and differentially expressed after inoculation of wheat spikes with fusarium head blight (FHB) (Kugler et al., 2013). Furthermore, a simple yet elegant ABCDE floral organ identity model explained the complex genetic interactions among MIKC-type wheat MADS-box genes for determining the fate of floral organs (Ali et al., 2019). These significant roles of MADS-box genes in biotic and abiotic stresses, fertilizer response, and flower development highlight the necessity of their comprehensive identification and characterization in the bread wheat genome.

Genome-wide identification and characterization of important gene families are becoming mainstream research approaches during recent years (Ahmed et al., 2021; Aleem et al., 2022). Previously, comprehensive genome-wide analyses of MADS-box genes have been carried out in diverse plant species. However, the available information about wheat MADS-box genes is comparatively sparse. For example, Ma et al. (2017) reported the first genome-wide analysis of MADS-box genes using only a draft version of the wheat genome (TGAcv1) which contains a lesser number of genes. Likewise, Schilling et al. (2020) studied only MIKC-type MADS-box genes from an updated genome version (Ref. Seq 1.0). Moreover, a significant number of low confidence and/or truncated protein-coding pseudogenes were also included in the before-mentioned studies.

To bridge the gaps in knowledge, in this study we comprehensively identify the entire complement of high confidence, full-length, protein-coding MADS-box genes present in the publicly available reference genome of bread wheat (IWGSC RefSeq v1.1). Through a combination of different search approaches, 300 high confidence, non-redundant, full-length MADS-box genes were identified.
Comparative phylogenetic analyses with model plant MADS-box genes further classified these into at least 16 Arabidopsis and/or grass specific subfamilies. Moreover, homologous and duplicated genes were identified to study probable gene family expansion and evolution mechanisms. Furthermore, expression patterns of identified MADS-box genes were studied under several biotic and abiotic stress conditions. In this way, we provide a comprehensive resource of wheat MADS-box genes which have the potential to facilitate molecular breeders in fine-tuning important traits for further improvement.

MATERIAL AND METHODS

Identification of MADS-Box Genes
The publicly available genome version of wheat (IWGSC Ref. Seq v1.1) was accessed through the ensemble plants database (Bolser et al., 2017) and searched using protein family database (Pfam) identifiers of MADS (PF00319) and K (PF01486) domains. A total of 281 MADS-domain and 131 K-domain encoding genes were identified. Additionally, a query-based search using “MADS” yielded 300 genes. Together, these searches identified a total of 712 genes (Supplementary Table S1), among which 300 were found to be high confidence and non-redundant and considered for further analyses. Detailed information of these 300 genes was retrieved from the ensemble plants database and is provided in Supplementary Table S2. Moreover, amino acid sequences for the identified genes were uploaded into the NCBI conserved domain database (Marchler-Bauer et al., 2015) for validation of putative protein domains (Supplementary Table S3). Gene identifiers were used as gene names for subsequent analyses; however, alternative names used in previous studies (Schilling et al., 2020) were also provided in Supplementary Table S4.

Comparative Phylogenetic Analyses and Subfamily Classifications
Differentiation between type I (M-type) and type II (MIKC-type) MADS-box proteins were achieved by separate MAFFT alignments using only the MADS domain (L-INS-i algorithm) (Katoh and Standley, 2013; Katoh et al., 2018) between Arabidopsis (Pařenícová et al., 2003) and wheat, and rice (Arora et al., 2007) and wheat MADS-box proteins. Subsequently, maximum likelihood (ML) phylogenies were inferred using IQ-TREE (Nguyen et al., 2015) by choosing JTT substitution model according to the Bayesian information criterion (BIC) (Kalyaanamoorthy et al., 2017). The consistency of the ML trees was validated by setting an Ultrafast bootstrap value of 1,000 (Minh et al., 2013; Hoang et al., 2018). The final phylogenetic trees were visualized with MEGA7 (Kumar et al., 2016).

Subfamily classifications were accomplished by separate MAFFT alignments within M-type (L-INS-i algorithm) and MIKC-type (E-INS-i algorithm) MADS-box protein sequences of Arabidopsis, rice, and wheat (Katoh and Standley, 2013; Katoh et al., 2018). The full-length alignments were subjected to the GapStrip/Squeeze v2.1.0 tool (www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html) for masking the individual residues by removing the gaps with default parameters. Then, masked alignments of M-type and MIKC-type proteins were independently subjected to the IQ-TREE software (Nguyen et al., 2015) for generating ML phylogenetic trees as described above. Subfamily names were given by following subfamily classifications in Arabidopsis and/or major grass species (Gramzow and Theißen, 2015) (Supplementary Table S2 and Supplementary Table S4).

Homeologs Identification
Putative homeologs were recognized based on strong phylogenetic relationships (Ultrafast bootstrap value >90) within different sub-families. Classifications reported in previous studies were also considered (Appels et al., 2018). The homeologs status of thirty-one genes could not be determined due to lower Ultrafast bootstrap values (Supplementary Table S5).

Gene Duplication and Evolution Analyses
Coding sequences (CDS) of all wheat MADS-box genes were retrieved from the ensemble plants database and blasted against each other using Sequence Demarcation Tool V1.2 (Muhire et al., 2014) for the identification of sequence identities. Gene pairs with ≥ 90% identity (E value < 1e−10) and non-homeologous status were considered as duplicated (Ning et al., 2017) (Supplementary Table S6). If the duplicated homologous gene pair was located on the same chromosome it was defined as tandem duplication. Otherwise, when homologous gene pairs were located on different chromosomes it was defined as segmental duplication. The CDS sequences of duplicated genes were MAFFT aligned, masked, and subjected to the Synonymous Non-Synonymous Analysis Program V2.1.1 (www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html) to compute the synonymous (Ks) and non-synonymous (Ka) substitution rates. To find out which type of codon selection operated during evolution, the ratio of Ka/Ks was also calculated. The approximate divergence time between duplicated gene pairs was calculated by using formulae T = Ks/2r × 10−6 assuming a substitution rate (r) of 6.5 × 10−9 substitutions/synonymous site/year (El Baidouri et al., 2017) (Supplementary Table S6).

In-Silico Expression Analysis
Expression data under all available abiotic and biotic stress conditions were retrieved from the expVIP Wheat Expression Browser (Borrill et al., 2016; Ramírez-González et al., 2018) as Log2 TPM (processed expression value in transcripts per million) obtained via RNA-seq analysis. Detailed information about expression levels of individual genes in tested tissues/growth stages and stresses/diseases are provided in the supplementary information (Supplementary Table S7). TTools (Chen et al., 2020) was used to generate a heatmap from the obtained expression data. Expression based clustering of genes was achieved by following the K-means clustering method (K = 10, iterations = 1,000, runs = 5) (Supplementary Table S8).
In this study, conserved domains and query search-based approaches identified a total of 300 high confidence and non-redundant MADS-box genes from the publicly available wheat genome (Figure 1A, Supplementary Table S2). The NCBI-CDD batch search further revealed that 167 (~55.7%) encode MADS-box proteins, 125 (~41.7%) encode MADS and K-box proteins, and only 8 (~2.6%) encode K-box domain-containing proteins (Supplementary Table S3). The protein lengths, molecular weight, and isoelectric points of MADS domain-containing proteins ranged from 58 to 450 amino acids, 6.483–48.304 kD, and 4.420–12.123 pI, respectively (Supplementary Table S2). This data suggests that different MADS-box genes may function within different environments.

Comparative phylogenetic analyses between *Arabidopsis* and wheat MADS-box genes (Supplementary Figure S1), as well as between rice and wheat (Supplementary Figure S2), distinguished wheat genes into M- and MIKC-types (Supplementary Table S2). In wheat, 128 (~43%) MADS-box genes exhibited high sequence similarity with *Arabidopsis* and rice M-type (type-I) genes, whereas 128 (~57%) revealed more homology with MIKC-type (type-II) genes. In general, MADS-box genes were equally distributed among the 21 wheat chromosomes, apart from the three homoeologous chromosomes (7A, 7B and 7D), which harboured a significantly higher number of genes and displayed peak gene density values (Figure 1A). This observation could be explained by the higher prevalence of duplicated gene pairs on these chromosomes (Appels et al., 2018). However, M-type genes were randomly distributed among chromosomes and were predominantly located on three homoeologous chromosomes of 3rd, 6th, and 5th linkage groups (Figure 1A and Supplementary Figure S3). In comparison, MIKC-type genes were equally distributed on all chromosomes. Furthermore, the percent contributions of A, B, and D genomes were also comparable (Figure 1B). As expected, MIKC-type genes were the largest group of MADS-box genes in wheat.

### Subfamily Diversity in MADS-Box Genes

Separate ML phylogenies among *Arabidopsis*, rice, and wheat MADS-box genes exposed 14 MIKC-type and 3 M-type major subfamilies. The 172 MIKC-type genes were unevenly dispersed into AP1 (9), AP3 (6), PI (6), AG/SIK (12), SEP (28), AGL6 (3), AGL12 (6), AGL17 (31), Bsister (19), MIKC* (27), OsMADS32 (3), SOC1 (13), and SVP (9) subfamilies (Figure 2A, Supplementary Table S2). As expected, monocot (OsMADS32) and eudicot (*FLC*) specific gene subfamilies were also witnessed. Likewise, 128 M-type genes were randomly distributed into *Ma* (53), *M* (28), and *My* (47) subfamilies (Figure 2B).

In general, *Arabidopsis*, rice and wheat MADS-box gene subfamilies roughly followed the species-specific phylogenetic clades. Triads of wheat homeologous genes exhibited close relationships with one or more rice genes, with *Arabidopsis* genes representing a sister group association to grass genes (e.g., the AP1, AP3, PI, AG/SIK, AGL6, AGL12, OsMADS32, SOC1, and SVP subfamilies; Figure 2A). Whereas the subfamily phylogenies were more complex in the case of AGL17, Bsister, MIKC*, SEP, and all M-type MADS-box genes, probably due to multiple duplication events during the polyploidization of wheat genome.

### Previously Classified FLC-Like Genes Grouped With MIKC*-Like Genes

*FLOWERING LOCUS C (FLC)-like* genes have been confirmed to regulate flowering time and vernalization responses in plants (Distelfeld et al., 2009; Andrés and Coupland, 2012) and their wheat and rice counterparts have been reported based on sequence homology and phylogenetic tree reconstructions (Ruelens et al., 2013; Schilling et al., 2020). However, in this study, despite employing updated resources and tools (see Materials and Method section), none of the wheat and rice MADS-box genes fell into the *Arabidopsis* specific FLC-clade.
All the wheat genes which were previously classified as FLC-like (Supplementary Table S4) were grouped with MIKC*-like MADS-box genes. Furthermore, we compared the amino acid sequences of Arabidopsis specific FLC genes and found that these were significantly different from MIKC*-like genes (Figure 3). The most significant differences were detected in MADS domain region at 30th, 34th, and 50th positions where glutamic acid (E), glutamine (Q), and alanine/glycine/serine (A/G/S) residues of FLC genes were substituted with lysine (K), glutamic acid (E), and proline (P), respectively. Collectively, these results indicate that wheat and rice genomes might lack FLC clade.

**Subfamily-Specific Gene Duplications in MADS-Box Genes**

Overall, MADS-box genes were equally located in the interstitial and proximal regions (R2a, R2b and C) and distal telomeric (R1 and R3) ends of chromosomes (49 and 51%, respectively) (Supplementary Table S2). However, substantial differences were observed among gene locations of M-type and MIKC-type genes, as well as among subfamilies. Most of the M-type genes were in distal telomeric segments (62%), whereas MIKC-type genes were more prevalent in central chromosomal segments (57%). Generally, a larger portion of genes belonging to significantly expended subfamilies tend to be in distal telomeric ends, whereas genes of smaller subfamilies were more clustered in central chromosomal segments (Supplementary Table S2).

Gene duplications were identified through sequence similarities in coding sequences of all MADS-box genes. A total of 201 duplicated gene pairs with ≥90% sequence homology were identified, which corresponded to 123 non-redundant genes (Figure 4, Supplementary Table S6). Two genes (TraesCSU02G209900, TraesCSU02G235300) with unknown chromosomal location information were also recognized to be duplicated. However, TraesCSU02G235300 showed duplications with genes located on chromosome 3B only, strongly suggesting that it was also located on the 3B chromosome. The MIKC-type MADS-box genes were also found to be more duplicated than M-type genes (60 vs 40% of duplicated gene pairs, respectively), particularly due to expended subfamilies (e.g., AGL17, Bsister, MIKC*, SEP, and SOC1). Remarkably, duplicated gene pairs were subfamily-specific and particularly recognized in subfamilies containing unbalanced homeologs, except for the AP3 subfamily (Table 1, Supplementary Table S5 and Supplementary Table S6). Among subfamilies, Ma, AGL17, MIKC*, and SEP contained 25, 21, 14, and 12% of the total duplicated gene pairs, respectively, and the majority of these were in distal telomeric and sub-telomeric (one gene located on the telomeric segment and other on the central segment) chromosomal regions. We also observed that the majority of the duplicated gene pairs (~51%) were in distal telomeric segments, whereas only 26 and 23% of the duplicated gene pairs were in proximal and sub-telomeric segments of chromosomes. These results could be explained by higher gene density in distal vs central chromosomal
segments (Appels et al., 2018). Furthermore, segmental duplications were more prevalent than tandem duplications (61 and 37%) in 197 of the duplicated gene pairs with available chromosomal information (Supplementary Table S6). Interestingly, >37% of all tandem duplications were identified on chromosome 3B, consistent with IWGSC findings (Appels et al., 2018). Collectively, these results strongly suggest that unbalanced homeologs in distal telomeric regions derive MADS-box subfamilies expansion through segmental duplications.

**Rapid Evolution of M-Type MADS-Box Genes**

To investigate evolution rates, we estimated substitutions in coding sequences and computed approximate divergence time
between duplicated gene pairs (Supplementary Table S6). In M-type genes, 1st and 3rd quartiles of synonymous substitutions (Ks) were considerably narrower than MIKC-type genes, whereas non-synonymous to synonymous substitution ratios (Ka/Ks) of M-type genes were significantly higher than 1st and 3rd quartiles of MIKC-type genes (Figure 5). Furthermore, the percentage of duplication events because of positive/Darwinian selection (Ka/Ks ratio >1) was almost doubled (23%) in M-type genes as compared with MIKC-type genes (12%) (Supplementary Table S6). Additionally, the estimated divergence time of M-type genes was significantly narrower than the MIKC-type genes. The mean divergence time of M-type genes was nearly half of the mean of MIKC-type genes (Figure 5). Taken together, these
data strongly indicate a rapid evolution of M-type MADS-box genes.

Expression Patterns of MADS-Box Genes Under Abiotic and Biotic Stresses

Extensive investigations have been carried out to study the expression patterns of MADS-box genes during the growth and development of crop plants. However, their transcriptional regulation under stressful conditions is somewhat obscure. Therefore, we analysed RNA-seq based expression data of 300 MADS-box genes under all available biotic and abiotic stress conditions in the exVIP wheat expression browser (Borrill et al., 2016; Ramírez-González et al., 2018) (Figure 6, Supplementary Table S7). Out of the total 300 genes, nearly 57% were expressed (log2 TPM 0.20–7.89) during at least one developmental stage of one or more of the stresses included in this study. Whereas the remaining 43% of genes showed no or very low expression (log2 TPM < 0.0) and were subsequently considered as not expressed.
FIGURE 6 | Expression patterns of wheat MADS-box genes under stressful conditions. RNA-seq based expression data of all MADS-box genes were retrieved from the exVIP wheat expression browser (Borrill et al., 2016; Ramírez-González et al., 2018) and a heatmap was generated with TBtools (Chen et al., 2020). Processed expression levels of all genes under different abiotic (I, spikes with water stress; II, drought and heat-stressed seedlings; III, seedling treated with PEG to induce drought; IV, shoots after 2 weeks of cold stress; V, phosphate starvation in roots/shoots/leaves) and biotic stresses (VI, coleoptile infection with Fusarium pseudograminearum/crown root; VII, FHB infected spikelets (0–48 h); VIII, FHB infected spikelets (30–50 h); IX, spikes inoculated with FHB and ABA/GA; X, CS spikes inoculated with FHB; XI, leaves naturally infected with Magnaporthe oryzae; XII, PAMP inoculation of seedlings; XIII, stripe rust infected seedlings; XIV, stripe rust and powdery mildew infection in seedlings; XV, Septoria tritici infected seedlings; XVI, Zymoseptoria tritici infected seedlings) (columns) and in different subfamilies (rows) are presented as Log2 transcripts per million (Log2 TPM). Detailed information about expression levels of individual genes in tested tissues/growth stages and stresses/diseases are provided in Supplementary Table S7.
Overall, MIKC-type MADS-box genes were highly expressed under all studied stresses, whereas nearly 75% of the non-expressed genes belonged to the three M-type subfamilies. Interestingly, the remaining 25% of the non-expressed genes belonged to MIK-subfamilies containing duplicated genes, e.g., AGL17, MIKC*, and Bsister. As most of the duplicated genes were in distal telomeric and sub-telomeric segments of chromosomes, it could be expected that promoters of these genes had undergone H3K27me3 (trimethylated histone H3 lysine 27) hypermethylation which resulted in their lower or non-existent expression (Ramírez-González et al., 2018).

In general, genes of all MIKC-type subfamilies (except AGL17) were ubiquitously expressed in fusarium head blight infected spikelets/spikes (Figure 6). Nevertheless, floral homeotic genes (AP1, AP3, PI, AG/STK, AGL6, and SEP) showed the highest expression patterns. SVP, MIKC*, AGL12, and SOCI-like genes were also ubiquitously expressed in all studied abiotic and biotic stresses, strongly indicating that genes belonging to these subfamilies are active guardians of the wheat genome under stressful environments. Remarkably, all AP1-like genes demonstrated high expression in Magnaporthe oryzae infected leaves, suggesting disease-specific expression, as well as the involvement of these important subfamily genes in diverse aspects of wheat growth and development. Similarly, AGL17-like genes were expressed only in seedlings under drought, heat, and phosphate starvation conditions (Figure 6).

We also calculated expression based hierarchical clusters to analyze diversification in expression patterns present within subfamilies (Supplementary Figure S4, Supplementary Table S8). The MIKC* and SOCI-like genes were grouped into seven and six random clusters, respectively. Similarly, Bsister and SEP-like genes were separately grouped into five random clusters. By contrast, AP1 and My-like genes exhibited no variation in their expression patterns, as all genes were grouped into a single random cluster. Likewise, AGL6, AGL17, AP3, OsMADS32, Ma, and Mβ-like genes displayed little diversification in expression, and genes of each subfamily were grouped into two random clusters. In comparison, many of the genes with no or very low expression (~58%) were grouped into cluster 10 and all belonged to subfamilies with duplicated genes (Supplementary Table S8).

**DISCUSSION**

**Recent Significant Advancements in Identification and Characterization of MADS-Box Family Members in Wheat**

MADS-box gene family members have been identified across all groups of eukaryotes and are known to confer diverse biological functions. In plants, these play a central role during growth and development and are thus very important targets for crop improvement (Ali et al., 2019). Recently, significant advancements have been made in the identification and characterization of MADS-box family members in wheat, which could facilitate crop breeding efforts and aid in the development of more resilient and higher-yielding cultivars. Ma et al. (2017) reported the first genome-wide analysis of MADS-box family members and identified a total of 180 MADS-box genes in an earlier genome version (Łukaszewski et al., 2014). Their in-silico expression data provided insights into the stress associated functions of MADS-box genes. To add to this, Schilling et al. (2020) conducted a comprehensive analysis of MIKC-type MADS-box genes and identified a total of 201 genes from an updated genome version (Ref. Seq 1.0) (Appels et al., 2018). They speculated that pervasive duplications, functional conservation, and putative neo-functionalization may have contributed to the adaptation of wheat to diverse environments. These advancements, along with results presented in this study, have begun to elucidate the broad landscape of wheat MADS-box genes and provide new insights into the phylogenomics, evolution, and stress associated functions of MADS-box family members.

**MADS-Box Genes Are Underestimated in the Wheat Genome**

Bread wheat is a hexaploid species (AABBDD, 2n = 6x = 42), that originated from a series of naturally occurring hybridization events among three closely related and independently maintained sub-genomes. Its full genome size is ~17 Gbps, among which ~14.5 Gbps (85%) is sequenced and contains ~107,891 coding genes (Appels et al., 2018). In this study, through a comprehensive genome-wide analysis, we identified a total of 300 high confidence MADS-box genes (Figure 1, Supplementary Table S2). To date, this is the second-highest number of MADS-box genes identified in a crop plant, closely following Brassica napus (Wu et al., 2018), which contains 307 full-lengths and/or incomplete (pseudo) MADS-box genes. Possible explanations for the abundance found in wheat may encompass the larger genome size, higher gene number, high rate of homeolog retention, and hexaploid nature of bread wheat (Appels et al., 2018). Moreover, at least 35 low confidence MIKC-type MADS-box genes were recently reported in the IWGSC reference genome (Schilling et al., 2020). If these low confidence genes are also included with the 300 MADS-box genes identified in this study, then the final number will be the highest identified in a plant genome. Furthermore, a more recent chromosome-scale assembly of the bread wheat genome revealed hundreds of new genes and thousands of additional gene copies (Alonge et al., 2020) which were substantially missing from the IWGSC reference genome explored in the current study. Collectively, all these observations strongly suggest that MADS-box genes are underestimated in wheat and their exploration might help in understanding the rapid global adaptability of wheat.

**Cereal Genomes Probably Lack FLC-Like Genes**

FLCs are key genes conferring vernalization requirement, which act as flowering repressors in Arabidopsis (Whittaker and Dean, 2017). High expression of FLCs repress other flowering activator genes and causes delayed flowering. In winter varieties of
temperate cereals, including barley, Brachypodium, and wheat, vernalization is regulated through VERNALIZATION genes (VRN1, VRN2, VRN3) (Greenup et al., 2009). To date, strong disagreement exists in the published literature regarding the presence or absence of FLC-like genes in grasses (Paolacci et al., 2007; Zhao et al., 2011; Ruelens et al., 2013; Wei et al., 2014; Fatima et al., 2020; Schilling et al., 2020). Previously, grass genes were classified into FLC-like based on synteny and phylogenetic reconstruction methods (Ruelens et al., 2013; Schilling et al., 2020). However, despite using a more recent reference genome sequence and sophisticated phylogenetic tree reconstruction methods, we could not find FLC-clade in wheat and rice genomes (Figure 2A). Interestingly, all those genes which were previously classified as FLC-like (Ruelens et al., 2013; Schilling et al., 2020), as well as OsMADS37 and OsMADS51/65, were grouped with MIKC*-like genes (Figure 2A, Supplementary Table S4). In comparison, all other FLC homologs were grouped into a separate distinct Arabidopsis-specific FLC clade. Moreover, we also compared the multiple sequence amino acid alignments of FLC- and MIKC*-like genes and observed that these were significantly dissimilar and none of the grass genes shared sequence homology with Arabidopsis FLC paralogs (Figure 3). Altogether, these results indicate that FLC-like genes might have been lost in cereal genomes after they diverged from eudicots. Several other studies are also in agreement with current results and were unable to identify FLC-like genes in grass genomes (Arora et al., 2007; Paolacci et al., 2007; Zhao et al., 2011; Wei et al., 2014; Fatima et al., 2020). These observations suggest that vernalization pathways in monocots and dicots evolved independently and are regulated through a completely different set of genes. However, earlier reports on the identification of FLC-like genes in grasses (Ruelens et al., 2013; Sharma et al., 2017; Schilling et al., 2020) and functional conservation of ODDSOC2 with FLC in the regulation of vernalization (Greenup et al., 2010), suggest that evolution of vernalization pathways between monocots and dicots may not be fully independent. Thus, the biological question of the presence or absence of FLC-like genes in grass genomes still needs to be explored. Further research efforts in this direction might address this fundamental question, and propitious outcomes could help in breeding efforts to develop winter varieties of temperate cereals which are adapted to changing environments.

### Subfamily-Specific Distal Telomeric Duplications in Unbalanced Homeologs Facilitate Rapid Adaptation to Changing Environments

Distal telomeric chromosomal segments are evolutionary hotspots for frequent recombination events and give rise to fast-evolving genes (Glover et al., 2015; Chen N. W. G. et al., 2018). Many adaptability traits related genes that are induced in response to external stimuli are found to be predominately located in distal chromosomal segments. In comparison, genes related to housekeeping and conserved developmental functions are positioned in central chromosomal segments (Appels et al., 2018; Ramirez-González et al., 2018). In this study, the majority of the identified duplicated gene pairs were in distal telomeric (~51%) and sub-telomeric (23%) chromosomal segments (Figure 4, Supplementary Table S6). Remarkably, all these duplications were subfamily-specific and primarily identified in larger subfamilies containing unbalanced homeologs (Ma, Mb, My, AGL17, Bsister, MIKC*, SEP, and SOCI) (Table 1). Interestingly, genes belonging to these subfamilies have been reported to regulate plants adaptability in changing environments. For example, OsMADS57 (AGL17-like gene in rice) is induced by abscisic acid, chilling, drought, and salinity stresses (Arora et al., 2007), promotes cold tolerance by interacting with a defence gene (OsWRKY94) (Chen L. et al., 2018), and modulates root to shoot nitrate translocation under deprived nitrate conditions (Huang et al., 2019). Similarly, downregulation of a SEP clade gene in pepper (CaMADS) caused more sensitivity to cold, salt, and osmotic stresses, whereas its overexpression in Arabidopsis conferred higher tolerance against these stresses (Chen et al., 2019). By contrast, almost negligible duplications were identified in a single smaller subfamily (AP3) with completely balanced homeologs (Table 1). These results strongly indicate that unbalanced homeologs of expanded subfamilies have undergone subfamily-specific distal telomeric duplications, thus facilitating the rapid adaptation of bread wheat to diverse global environments. Contrarily, completely balanced homeologs of functionally conserved smaller subfamilies might have some evolutionary advantage in minimizing the developmentally detrimental gene copy number variations, due to their localization in central chromosomal segments.

### M-Type MADS-Box Genes Are Evolving at a Faster Rate

Unlike MIKC-type MADS-box genes, the M-type genes have not been extensively studied and little information is available about their evolutionary origin in crop plants (Masiero et al., 2011). In this study, we observed that M-type genes were predominately located in distal telomeric chromosomal segments (Supplementary Table S2) and are evolving at a faster rate, probably due to the higher frequency of tandem gene duplications and stronger purifying selections (Figure 5, Supplementary Table S6). Moreover, lesser synonymous substitutions and higher non-synonymous to synonymous substitution ratios of M-type duplicated genes exposed their more rapid evolution as compared with MIKC-type genes (Figure 5, Supplementary Table S6). These results agree with Nam et al. (2004), who also reported faster birth-and-death evolution of M-type genes in angiosperms. This data might indicate the significance of M-type MADS-box genes in understanding the evolutionary history of the bread wheat genome.

### MADS-Box Genes as Active Guardians of Plants Against Pathogen Insurgency and Harsh Environments

MADS-box genes regulate diverse developmental processes, and their functions are well studied in plant morpho-
organogenesis (Smaczniak et al., 2012; Ali et al., 2019). However, several members of the MADS-box gene family are reported to be involved in the regulation of biotic and abiotic stress responses (Zhang et al., 2016; Wang Q. et al., 2018; Castelán-Muñoz et al., 2019), which point towards possible dynamic roles in stress response. In this study, we analysed the complete atlas of their expression profiles in all publicly available wheat transcriptomic data under biotic and abiotic stresses (Figure 6 and Supplementary Figure S4, Supplementary Table S7 and Supplementary Table S8). Except for AGL17, genes of all other MIKC-type subfamilies were ubiquitously expressed in FHB infected spikes/spikelets (Figure 6) mimicking their defensive roles against Fusarium infection. Previously, Yang et al. (2015) functionally characterized FgMcm1 in the causal agent (Fusarium graminearum) of barley and wheat head blight disease and demonstrated that FgMcm1 played crucial roles in cell identity and fungal development. More recently, Xu et al. (2020) also reported a close relationship between anther extrusion and field FHB resistance and pointed out that Rht and Vrn genes might have pleiotropic effects on these traits. Since FHB is a floral disease and floral organ identity is controlled by MIKC-type MADS-box genes (Ali et al., 2019), this indicates a cooperative relationship for their involvement in the pathogen response network. In future, it would be interesting to investigate how MADS-box genes are involved in fighting FHB infection.

Similarly, SVP, MIKC*, AGL12, and SOC1 subfamily genes were ubiquitously expressed under all studied stresses (Figure 6, Supplementary Table S7), pointing towards their more critical role in stress associated functions. In comparison, members of the largest MIKC group (AGL17) were expressed under phosphate starvation, and heat and drought stress conditions. Several functionally characterized genes belonging to all these subfamilies have also been reported to regulate different abiotic and/or biotic stresses in plants (Khong et al., 2015; Shi et al., 2016; Wang Z. et al., 2018; Hwang et al., 2019; Li P. et al., 2020). Interestingly, few M-type genes, especially Ma-like, were weakly expressed under different biotic stresses (Supplementary Table S7) and these expression patterns agree with Guo et al. (2013), who observed differential expression of an M-type gene in response to stripe rust infection in wheat. Collectively, these results strongly highlight MADS-box genes, especially MIKC-type, as key members of gene regulatory networks (Castelán-Muñoz et al., 2019) and their involvement in wheat response to possible pathogen insurgency and harsh environments.

Towards Bridging the Gaps Between Genotype and Phenotype Using Genetic and Genomic Resources

Understanding the functions of candidate genes controlling agronomically important traits is critical for the acceleration of crop improvement efforts. Following the unveiling of recent wheat genetic and genomic resources (Avni et al., 2017; Luo et al., 2017; Maccalferrì et al., 2019; Adamski et al., 2020), rapid testing of previous model plant discoveries and their cross-application in staple crops is paving the way towards bridging the gap between genotype and phenotype. During recent years, several wheat genes related to flowering, yield, quality, disease resistance, male sterility, and nutrient use efficiency traits have been cloned and functionally characterized [briefly reviewed by Li J. et al. (2020)], which exposed the unprecedented potential of these genetic and functional genomic resources in establishing genotype-to-phenotype relationships. These resources could also facilitate in elucidating the regulatory roles of neglected MADS-box subfamilies, such as Ma, Mβ, My, and MIKC*, during wheat growth and development, and subsequently help in deciphering whether MADS-box genes confer pathogen resistance and/or abiotic stress tolerance. Furthermore, many other fundamental biological questions about their phylogenomics, evolutionary origin, and stress associated functions, could be addressed by exploiting these paramount resources.

CONCLUSION

MADS-box genes are critically important for wheat growth and development and hold enormous promise for bolstering yield potential under changing global environments. Our results speculate that the abundance of MADS-box genes in the wheat genome might be associated with its global adaptability and that subfamily-specific distal telomeric duplications in unbalanced homeologs facilitate its rapid adaptation. In addition, through comprehensive genome-wide and comparative analyses, we demonstrated that wheat and rice genomes might lack FLC-like genes, which could help molecular breeders in the identification of alternative target genes for fine-tuning of winter wheat varieties. Moreover, our in-silico expression data strongly indicated the possibility of protective roles of MADS-box genes against pathogen attacks and harsh climatic conditions. In this way, we provided an entire complement of MADS-box genes identified in the wheat genome that could accelerate functional genomics efforts and possibly facilitate in bridging genotype-to-phenotype relationships through fine-tuning of agronomically important traits.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

QR conceived and designed the study. QR, AR, RA, BH and IR retrieved the data. QR, AR and RA curated the data and performed final analyses. QR, AR and BH drafted the manuscript. QR, IR, ZA, HB and IA revised and finalized the manuscript. IA secured funding for research publication. All authors read and approved the final manuscript.
ACKNOWLEDGMENTS

All authors are grateful to Dr Yuling Jiao, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences and the reviewers for their comments and suggestions on the manuscript. The inspiration and support from Precision Agriculture and Analytics Lab (PAAL), an affiliate of National Centre in Big Data and Cloud Computing (NCBC) is also gratefully acknowledged. A preprint version of the manuscript is publicly available at BioRxiv (Raza et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found on the link: [https://www.frontiersin.org/articles/10.3389/fgen.2021.818880/full#supplementary-material]
