Journal of Advanced Research 42 (2022) 135–148

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Original Article

New insights into the evolution of CAF1 family and utilization of *TaCAF1Ia1* specificity to reveal the origin of the maternal progenitor for common wheat

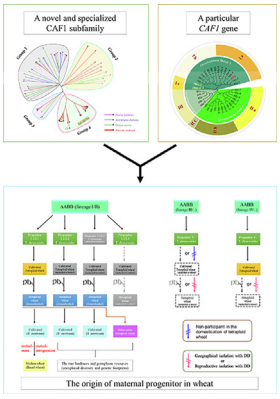
Longqing Sun [a,b,1], Ruilian Song [a,1], Yixiang Wang [a], Xiaofang Wang [a], Junhua Peng [c], Eviatar Nevo [d], Xifeng Ren [a,*], Dongfa Sun [a,*]

[a] Hubei Hongshan Laboratory, College of Plant Science & Technology, Huazhong Agricultural University, Wuhan, Hubei, China
[b] Food Crops Institute, Hubei Academy of Agricultural Sciences, Wuhan, Hubei, China
[c] Germplasm Enhancement Department, Huazhi Biotech Institute, Changsa, Hunan, China
[d] Institute of Evolution, University of Haifa, Mount Carmel, Haifa, Israel

HIGHLIGHTS

- The study presents a novel plant non-typical TaCAF1Ia subfamily.
- The conservation of TaCAF1Ia is useful for the phylogenetic analysis of Triticeae.
- Utilization of *TaCAF1Ia1* specificity reveals the new origin model for common wheat.
- TaCAF1Ia members may be novel factors for anther development in plant.
- The TaCAF1Ia proteins form CCR4-NOT complex in wheat by new interaction sites.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 14 November 2021
Revised 19 March 2022
Accepted 8 April 2022
Available online 13 April 2022

Keywords:
Wheat
CAF1
Progenitor
Phylogenetic
Duplication
Lineage

ABSTRACT

Introduction: Until now, the most likely direct maternal progenitor (AABB) for common wheat (AABBDD) has yet to be identified. Here, we try to solve this particular problem with the specificity of a novel gene family in wheat and by using large population of rare germplasm resources.
Objectives: Dissect the novelty of TaCAF1Ia subfamily in wheat. Exploit the conservative and specific characteristics of *TaCAF1Ia1* to reveal the origin of the maternal progenitor for common wheat.
Methods: Phylogenetic and collinear analysis of *TaCAF1* genes were performed to identify the evolutionary specificity of TaCAF1Ia subfamily. The large-scale expression patterns and interaction patterns analysis of CCR4-NOT complex were used to clarify the expressed and structural specificity of TaCAF1Ia subfamily in wheat. The population resequencing and phylogeny analysis of the *TaCAF1Ia1* were utilized for the traceability analysis to understand gene-pool exchanges during the transferring and subsequent development from tetraploid to hexaploidy wheat.
Results: TaCAF1Ia is a novel non-typical CAF1 subfamily without DEDD (Asp-Glu-Asp-Asp) domain, whose members were extensively duplicated in wheat genome. The replication events had started and

Peer review under responsibility of Cairo University.
* Corresponding authors.
E-mail addresses: renxifeng@mail.hzau.edu.cn (X. Ren), sundongfa1@mail.hzau.edu.cn (D. Sun).
[1] These authors contributed equally.

https://doi.org/10.1016/j.jare.2022.04.003

constantly evolved from ancestor species. Specifically, it was found that a key member *CAF1Ia1* was highly specialized and only existed in the subB genome and S genome. Unlike *CAF1s* reported in other plants, *TaCAF1Ia* genes may be new factors for anther development. These atypical TaCAF1s could also form CCR4-NOT complex in wheat but with new interaction sites. Utilizing the particular but conserved characteristics of the *TaCAF1Ia1* gene, the comparative analysis of haplotypes composition for *TaCAF1Ia1* were identified among wheat populations with different ploidy levels. Based on this, the dual-lineages origin model of maternal progenitor for common wheat and potential three-lineages domestication model for cultivated tetraploid wheat were proposed.

Conclusion: This study brings fresh insights for revealing the origin of wheat and the function of *CAF1* in plants.

Introduction

The genome sequences assembling for bread wheat by whole-genome shotgun sequencing started in 2012 [1], but it was not until 2018 when the world's first complete genome map of the hexaploid wheat was published under the effort of International Wheat Genome Sequencing Consortium (IWGSC) [2]. More recently, the 10 + Genome Project, as the most comprehensive atlas of wheat genome sequences ever documented, was unveiled to the world [3]. However, due to the complexity of the wheat genome and the incomplete germplasm collection of wild populations, the common wheat genome landscape and domestication history remain elusive. Common wheat is most likely derived from an ancestral crossing event between *Triticum durum* and a D genome close to *Aegilops tauschii* [4]. During the evolution, the intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat [5], suggesting that historic gene flow from wild relatives made a substantial contribution to the adaptive diversity of modern bread wheat [6]. It was also reported that the composite introgression from wild populations contributed to a substantial portion of the bread wheat genome [7]. Additionally, the genetic diversity of the A/B subgenomes in Chinese cultivars was suggested to be mainly contributed by the European landraces [8]. Recently, the genomic fingerprints of 44,624 wheat lines were generated and used for characterizing the role of selection in shaping patterns of allelic variation over time [9]. The large-scale genotyping and diversity analysis revealed endemic species with untapped diversity and genetic footprint [10]. In sum, although all these previous studies so far have provided valuable resources to the wheat community for use in enhancing wheat productivity and breeding, the information obtained is complex and scattered, making it difficult to be used for understanding the origin and evolution of wheat. A long-standing question of when, where and how many times wild emmer wheat has been domesticated remains unanswered. With the records of archaeological remains alone, it is difficult to determine where the process of domestication was first initiated and who was deemed to be the maternal progenitor for hexaploid wheat [11]. Also, whether tetraploid emmer wheat was domesticated independently in southeastern Turkey (northern Levant) and southern Levant or has a monophyletic origin combined with the subsequent hybridization among the populations has been an ongoing debate [12].

CCR4-NOT is a conserved multiprotein complex which regulates eukaryotic gene expression [13], including chromatin modification, transcription elongation, mRNA degradation, miRNA gene silencing, RNA nuclear surveillance and nuclear export [14]. It is a highly conserved regulatory biomacromolecule, widely existing in microorganisms and higher animals during the evolutionary process [15]. CAF1, a core factor in the CCR4-NOT complex, was shown to be crucial for deadenylation [14]. Loss of *CAF1* function in *C. elegans* resulted in sterility in both males and hermaphrodites due to blocked germ cell development at the pachytene stage of meiosis I [16]. *CAF1* is an essential gene for sperm production in mammals and has been well studied [17–19]. Although the CCR4-NOT complex also exists in plants, its function and the underlying molecular mechanism has been rarely studied. It was reported that OsCAF1 acts as a bridge between OsCCR4 and OsNOT1 in rice CCR4-NOT complex [20]. Moreover, the rice homologues were the same as those in yeast and human containing the conserved deadenylase subunits of DEDD domain [21]. More interestingly, most of the studies in *Arabidopsis*, rice, pepper, citrus and poplar have demonstrated that plant CAF1s are widely involved in both biotic and abiotic stress responses in plants [22–25]. The role CAF1 may play in plant reproductive development has not been verified.

Until now, the most likely direct maternal progenitor (AABB) for common wheat (AABBDD) has yet to be identified. The main reason for this may be that all the hexaploid wheat (AABBDD) were cultivated species, no wild hexaploid wheat existed. Moreover, frequent intra-species and inter-species introgression incessantly shapes the excessive complexity of genetic variation for common wheat, and the high genetic diversity of AB subgenome of common wheat adds additional complexity. It is important to mention that the A and D subgenomes of common wheat are derived from the diploid progenitors of *Triticum urartu* (AA) and *Aegilops tauschii* (DD) both having monophyletic mode of origin [26,27]. Therefore, the study on the origin of maternal progenitor (AABB) for common wheat is more dependent on the origin of BB subgenome. Molecular phylogenetic analysis is widely used to study the biological origin and evolution. Low-copy genes have been useful in inferring inter-species phylogenetic relationship in Triticeae [28,29]. However, due to the increased ploidy level, hybridization, infiltration, recombination and gene duplication, there are at least three copies for each one of those genes in common wheat genome, making them difficult to be amplified and clearly distinguished for phylogenetic studies in tetraploid and hexaploid wheat.

In search for a new gene specialized but conserved in the subB genome for using to reveal the origin of the maternal progenitor in common wheat, we identified a novel plant *CAF1* gene *TaCAF1Ia1*, which was highly specialized, existing only in subB genome and S genome. With this gene, we conducted traceability analysis using a largest ever population of wild emmer wheat (AABB) from the Fertile Crescent where the AABB genome originated and a worldwide collection of tetraploid and hexaploid wheat populations. The results obtained provided new insights to reveal the origin of maternal progenitor for common wheat. Functional analysis also implied that TaCAF1Ia may play an important role in anther development.

Materials and methods

Plant materials

A total of 1506 wheat accessions were used in this study, of which 864 were from three natural populations of wheat, including

294 wild emmer accessions from the Fertile Crescent and its peripheries, 238 landraces and cultivars of the durum wheat accessions from 44 countries and regions, 301 landraces and cultivars of the common wheat from Chinese mini-core collections, and 31 accessions of several other cultivated tetraploid varieties from 8 countries. Those materials were provided by the University of Haifa, the USDA (United States Department of Agriculture) and Huazhong Agricultural University wheat germplasm collection. Information about geographical origins of individuals used in this investigation is listed in **Table S1**. Seedling tissues of the above accessions were used for DNA isolation. The rest of the 642 hexaploid wheat accessions were from resequencing studies published since 2019 [3,5,7,30,31].

Identification of TaCAF1 genes

The hidden Markov model (HMMs) of the CAF1 domain (PF04857) was downloaded from the Pfam database for scanning *CAF1* genes in the IWGSC (International Wheat Genome Sequencing Consortium) v1.0 of common wheat genome database. BLAST search was also conducted for large-scale identification of wheat CAF1s using *Arabidopsis* and rice CAF1 protein sequences as queries with a cutoff E-value of $\leq 10^{-5}$. All candidate *TaCAF1* genes obtained by above methods were further entered into NCBI conserved domain database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) to confirm the conserved CAF1 domain. A total of 28 *TaCAF1* genes with annotated information were identified. BLAST search was conducted to extensively identify wheat TaCAF1Ia subfamily genes using *TaCAF1Ia1* sequence. Then, the method of the in-depth alignment scanning was performed for the deep excavation of the subfamily members in wheat genome database. Through the genome-wide identification, there are altogether 46 members identified for TaCAF1Ia subfamily in this study, in which only six genes were annotated in common wheat genome. The other 40 members were not found and annotated in wheat genome before. All of the 46 members had more than 70% identity of nucleotide sequence with TaCAF1Ia1.

Sequence analysis

ExPASy (http://web.expasy.org/protparam/) proteomics server was used to predict the molecular weight (Mw), isoelectric point (pI), aliphatic index (AI), instability index (II) and grand average of hydropathicity (GRAVY) of the CAF1 family proteins. The MEME tool (http://meme-suite.org/) was used to predict the conserved motifs of each CAF1 protein sequence. The exon-intron structure and the chromosome physical location of the CAF1 genes were displayed by TBtools software [32]. Furthermore, the homologous gene clusters were identified based on their chromosome physical positions and their sequence similarities. We constructed x/y/z/m sites for defining the structures of gene clusters formed by TaCAF1Ia subfamily members on wheat genome, and any two sites from the same chromosome were considered as the old tandem duplicated events. Then, we defined adjacent homologous genes with the similar structure and CDS in each site as new tandem duplicated genes, most of which have a distance of less than 50 kb. Each of the identified *CAF1* genes was assigned with a unique name based on its cluster type of classification and chromosome location. The time of divergence between duplicated genes was estimated using a rate of 6.5×10^{-9} synonymous substitutions per synonymous site per year [33].

Interspecific and intraspecific collinearity analyses

The Multiple Collinearity Scan toolkit was adopted to analyze gene duplication events (MCScanX: https://chibba.pgml.uga.edu/mcscan2/). To exhibit the duplicated pairs and orthologous pairs of *CAF1* genes between common wheat and its ancestral species, the synteny relationship was built using dual synteny plot toolkit (https://github.com/CJ-Chen/TBtools). The visual presentation of collinearity analysis was completed using TBtools software [32]. For homologous genes identification and collinearity analysis in Triticeae species, the *CAF1* nucleotide sequences were used as query for Blast search against public databases with sequences identity >95%, then comparing the alignment results one by one for validating the order of x, y, z and m sites in each chromosome group. For AA, DD, AABB and AABBDD genomes, the Ensembl Plants database was used (https://plants.ensembl.org/index.html). For SS, $S^{sh}S^{sh}$, EE, HH, and RR genomes, the WheatOmics 1.0 database was used for identifying the different types and sites of TaCAF1Ia members on different chromosome groups [34]. Comparison of syntenic *CAF1Ia* physical maps in different genomes was used for inferring the evolutionary relationship of *Triticae* species.

Phylogenetic analysis of TaCAF1 genes

Multiple sequence alignments were performed using MEGA 7.0 [35]. Nucleotide diversity and Tests of neutral evolution were performed using the DnaSP version 5.0 software [36]. Arlequin software was used to compute genetic indices, grouping the identical sequences into haplotypes. The network of haplotype cluster analysis was done with MEGA 7.0 and Arlequin software [35,37], and loaded into PopART software to visualize [38]. The phylogenetic tree was constructed using MEGA 7.0 with the maximum likelihood (ML) method under the JTT + G amino acid substitution model and 1000 times bootstrapping test, and the neighbor joining (NJ) method under the pairwise deletion model with 1000 replications. The phylogenetic tree was visualized by plotting it using the Figtree software (http://tree.bio.ed.ac.uk/software/figtree/).

Expression analysis of TaCAF1 genes

The expression data of 28 wheat *TaCAF1* genes were downloaded from Wheat Expression Browser (www.wheat-expression.com). We picked up the expression values of these genes in the major tissues including grain, root, leaf and spike at the different developmental stages referring to 81 samples for further analysis. The transcripts per million (TPM) value of each gene in every sample was normalized by $\log_{10}$TPM and used for the heatmap generation. The expressional values of *CAF1* family members in different tissues of *Arabidopsis thaliana* and rice were collected from the online data (https://bar.utoronto.ca/eFP-Seq_Browser/). The fragments per kilobase of transcript per million mapped reads (FPKM) value of each gene in every sample was normalized by $\log_{10}$FPKM and used for the heatmap generation. The expressional values of *CAF1* family genes in different tissues of human (BioProject: PRJEB4337) and mouse (BioProject: PRJNA66167) were downloaded from the RNA-seq database in National Center for Biotechnology Information (NCBI). The FPKM value of each gene in every sample was normalized by $\log_{10}$FPKM and used for the heatmap generation.

Natural variation of TaCAF1Ia1 in different wheat populations

Genomic DNA was prepared using Plant Genomic DNA Kit (TIANGEN Biotech, Beijing, China). PCR amplifications were conducted using the 2 × Hieff® PCR Master Mix (No Dye) (Yeasen, Shanghai, China) following the manufacturer's protocol. Briefly, the PCR conditions were set as follows: denaturation at 94 °C for 5 min, followed by 32 cycles of 94 °C for 30 s, Tm for 30 s and 72 °C for varied timing (set as 30 s per 1 kb length of amplicon, see the protocol). Finally, the PCR mixtures were kept at 72 °C

for 5 min. The PCR products were directly sequenced to acquire the full allelic information of *TaCAF1Ia1* from each wheat accession.

GUS staining assay

In order to understand the relationship between tissue differential expression patterns of *TaCAF1Ia* members and their functions, the promoters of 17 *TaCAF1Ia* genes were respectively constructed into the GUS fusion expression vectors and transformed into *Arabidopsis* for GUS histochemical assays. The promoter sequences of the *TaCAF1Ia* genes were PCR-amplified from wheat genomic DNA using the 2x Prime STAR Max Premix (TaKaRa) following the manufacturer's instruction. The primers for amplification are listed in **Table S1**. The amplified sequences were first inserted into pGEM-T easy vector (Promega) for sequence verification, and then recombined into the expression vector pGWB433 containing a GUS reporter gene using Gateway cloning system (Invitrogen). We selected transgenic plants among offspring by growing them on MS media containing kanamycin and eventually obtained homozygous transgenic lines.

Yeast two-hybrid (Y2H) assay

To explore and validate the hypothesis that TaCAF1Ia proteins also form CCR4-NOT complex in wheat, the protein interaction patterns were identified by Y2H assay. Y2H assays were performed using the GAL4-based two-hybrid system (Clontech, USA). The pGADT7 and pGBKT7 vectors were digested with the two endonucleases EcoRI and BamHI respectively, and the purified PCR products were connected to the yeast expression vectors by homologous recombination according to the operating instructions of the ClonExpress® II One Step Cloning kit (Vazyme, Nanjing, China). Briefly, to verify the protein interaction relationships among CCR4, NOT1 and CAF1, the coding sequence (CDS) of *CCR4* was divided into five fragments. The full-length of *CCR4* and each of the five fragments and the two domains of *NOT1* were cloned into the pGBKT7 vector to generate pGBK-baits (Clontech, USA), respectively. The coding regions of the 10 *TaCAF1* genes were cloned into the pGADT7 vector to generate pGAD-preys, respectively. To verity the interactions between CCR4 and NOT1, the two domains of NOT1 were separately recombined into the pGADT7 plasmid. After that, the different combinations of the recombinant plasmids pGAD-preys and pGBK-baits were respectively co-transformed into the yeast strain Y2HGold according to the manufactor's protocol. Primers used in this experiment are listed in **Table S1**.

BiFC assays

For verifying the protein interactions from the Y2H assays, we performed a BiFC assay. 2300-YC and 1300-YN vectors and the co-transformation vector 35S: P19 were used for BiFC assays. The CDS of *CCR4* and *TaNOT1* (TTP domain) without stop codons were inserted into the 2300-YC plasmid between the SmaI and SpeI sites, containing DNA encoding the C-terminal regions of YFP. Similarly, the *CAF1s* sequences without the start codons were cloned into the 1300-YN plasmid between the SmaI and SalI sites containing DNA encoding the N-terminal regions of YFP. The recombinant plasmids were transformed into *Agrobacterium tumefaciens* GV3101, respectively. *Agrobacteria* containing different constructs were incubated, harvested, and resuspended in infiltration buffer (10 mM MgCl2, 0.2 mM acetosyringone) at a final concentration of OD600 = 0.4–0.5, then mixed in equal volumes to obtain an infection solution and incubated at room temperature for 2 h. The mixture containing two plasmids were infiltrated into leaves of 5- to 6-week-old *Nicotiana benthamiana* plants. After 48 h, YFP fluorescence signals were detected by the confocal microscope (Leica; Carl Zeiss, Heidenheim, Germany). The primers used for plasmid construction are presented in **Table S1**.

Protein expression and Pull-Down assay

The Pull-Down assay was employed to further confirm the direct physical interactions between TaCAF1Ia proteins and TaCCR4 or TaNOT1 (TTP domain). The full-length sequences of the 10 *TaCAF1* genes amplified and sequencing confirmed were inserted into the GST-tag vector pGEX-4T-1 through the Gateway reaction. The CDSs of the *CCR4* and TTP domain were amplified and cloned separately into the BamHI/HindIII sites of the MBP-tag vector pMAL-c2X (Genecreate, Wuhan, China). The resulting plasmids were introduced into the BL21 (DE3) competent cells (TransGen, Beijing, China). Positive clones were cultured in Luria-Bertani (LB) medium containing 50 mg/ml ampicillin at 37 °C to OD600 value of between 0.4 and 0.6. Expression of the fusion protein was then induced using 0.3 mM isopropyl b-D-thiogalactopyranoside (IPTG) for 13 h at 18 °C or 4 h at 37 °C. The cultured cells were harvested and resuspended using GST pull down buffer (20 mm Tris-cl, pH8.0, 200 mm NaCl, 0.5% NP-40, 1 mm EDTA, pH8.0) containing 0.2 mg/ml lysozyme and protease inhibitor cocktail, and lysed by sonication. The suspension after sonication was centrifuged at 4 °C to collect the supernatant. For the subsequent GST pull down assay, approximately 500 µl supernatant containing GST or GST-fusion protein was incubated with MBP-tagged fusion protein for more than 4 h at 4 °C on a rotating platform. The 300 µl MagneGST™ Glutathione Particles were washed 2 times with 1 mL GST pull down buffer, then incubated with 1 mL mixed protein lysates rocked more than 2 h at 4 °C. The protein complex was washed 5 times with 1 mL GST pull down buffer, and then eluted with GST elution buffer (1 M Tris-cl, pH8.0, containing 10 mM reduced glutathione). The bound proteins were analyzed by 10% SDS-PAGE gel electrophoresis and transferred to a PVDF membrane (Yeasen, Shanghai, China). The membrane was blocked using 5% (w/v) skim milk powder for 1 h and incubated overnight at 4 °C in 1 × Tris-buffered saline containing Tween 20 (TBST; 1 × TBS with 0.1% [v/v] Tween 20) and anti-MBP or anti-GST antibodies (diluted 1:5,000). After washing three times with 1 × TBST for 10 min each time, the membrane was incubated in 1 × TBST solution containing the secondary antibody (diluted 1:10,000) for 1 h at room temperature, followed by three washes (10 min each) with 1 × TBST, and then used for detection by super ECL detection reagent (Yeasen, Shanghai, China). The signals were detected using a Chemiluminescence Imaging system (tanon-5200). Three experimental replicates were performed, and the results of protein-protein interactions were consistent. Primers used in this experiment are listed in **Table S1.**

Results

Identification of CAF1 proteins in common wheat

CCR4-associated factor 1 (CAF1) proteins are highly conserved deadenylases in eukaryotic cells [39], acting as a core factor in the CCR4–NOT complex. They could be well distinguished among regnum, division, classes, orders, families, genera and species in phylogenetic analysis (**Fig. S1**), suggesting that the *CAF1* gene family has been well conserved through evolution from microbe to animals and plants. In microbe, there was only a single copy of the gene *CAF1*, while in fish and mammals there were two homologous genes of *CAF1* and *POP2* [40]. However, in model plants, there were 11 and 18 homologous genes of *CAF1* for *Arabidopsis* and rice, respectively [21] (**Fig. S2a,b**). Our previous studies have

shown that *TaCAF1Ia1* (TraesCS3B02G424000) is the new target gene of tae-miR2275 with conserved CAF1-domain (**Fig. S2c**), which plays an important role in anther development in wheat [41]. Nevertheless, the phylogenetic tree constructed by all CAF1 proteins from *Arabidopsis* and rice together with TaCAF1Ia1 showed that TaCAF1Ia1 does not belong to any of the main clades (**Fig. S2d**), suggesting a specialization event occurred in wheat.

In order to obtain more insight on *CAF1* gene family, the genome sequence of the 'Chinese Spring' was analyzed to identify the *TaCAF1* genes in wheat. A total of 28 *TaCAF1* genes were identified in this study. Comparison of TaCAF1 protein sequences revealed that the protein similarity ranges from 10.16 to 97.48%, suggesting that most of TaCAF1s may be different from each other and present great divergence in wheat (**Fig. S3a**). Further, chromosome position analysis showed that the 28 *TaCAF1* genes were unevenly distributed in 17 out of 21 wheat chromosomes (**Fig. S3b**). The interchromosomal synteny was employed to identify the duplication events occurring at the *TaCAF1* family members. Most collinear gene pairs happened within the same chromosome group. However, no genes generated collinear pairs with *TaCAF1Ia1* (**Fig. S3c**). The exon-intron organization analysis for *TaCAF1s* showed that the wheat *CAF1s* were intron-less genes, 16 of which had no intron in their genome sequences, with the *TaCAF1Ia1* containing only one exon with the shortest ORF sequence (**Fig. S3d**).

CAF1 belongs to the RNase D group in the DEDD family [42]. Here, we found that only 12 of 28 *TaCAF1* members from wheat own conservative DEDD motifs, and the rest belong to the non-typical TaCAF1s without DEDD motifs. Obviously, the *TaCAF1Ia1* is a non-typical *CAF1* gene (Fig. 1**a**). Phylogenetic tree showed that all TaCAF1s of wheat were clustered into 6 clades, and most of them in the same clade were from the same chromosome group. Both clades I and II could be further divided into two subgroups (Fig. 1**b**). However, it is obvious that TraesCS6A02G358700, TraesCS6B02G391400 and TraesCS6D02G341500 exhibit the most closely-related relationship with the representative CAF1s, whereas *TaCAF1Ia1* has rather distant phylogenetic relationship with them (**Fig. S4**).

To gain a more in-depth insight to this gene family, we analyzed the *TaCAF1s* gene expression from the wheat-expression datasets. As shown in Fig. 1**b**, the expression patterns of different *TaCAF1s* in the same evolutionary clade were similar, and many of them showed high expression level or specificity. The expression of the *TaCAF1* members in clade I is relatively low in most of the tissues, but was mainly detected in spikes. The members in clade IIa were dominantly expressed in leaves and roots, except TraesCS2D02G597200, which highly expressed specifically in leaves. All the members except TraesCS7A02G082100 in clade IIb were ubiquitously expressed in almost all of the samples and tissues. The expression patterns of the *TaCAF1s* in clades III and Ⅵ are very similar. The *TaCAF1* genes in clades IV and Ⅴ were expressed in all tissues with a relatively low level, except TraesCS3A02G447900 with a slightly higher expression than others. Remarkably, *TaCAF1Ia1* was found to be specifically expressed in the reproductive organ of spike. However, the three closely-related TaCAF1s, TraesCS6A02G358700, TraesCS6B02G391400 and TraesCS6D02G341500, were expressed in all tissue samples (Fig. 1**b**), consistent with the observation in animals that *CAF1s* were also widely expressed in various tissues (**Fig. S5a,b**). Even more interestingly, the *CAF1* homoeologous genes in *Arabidopsis* and rice still exhibit wide expression in all detected growth and development tissues (**Fig. S5c,d**). These results revealed that in model species, *CAF1s* have a broad expression in different tissues. However, the wheat-specialized *TaCAF1Ia1*, which is specifically expressed in spike, may be have the potential significance in molecular evolution of *CAF1* family.

In-depth exploration and classification of the TaCAF1Ia subfamily members from different chromosome groups of common wheat

To further explore the origin of this organ-specific *TaCAF1Ia1*, synteny of the *CAF1* genes between common wheat and ancestor species was constructed for tracing phylogenetic history of the *CAF1* family in donor genome. For most of the *TaCAF1s*, a corresponding member can be identified in their ancestral species. Interestingly, no colinear gene pairs of *TaCAF1Ia1* were found between *T. dicoccoides* and *T. aestivum* genomes (**Fig. S6a**). However, after scanning deeply into the *T. dicoccoides* genome, we identified an orthologous gene from *T. dicoccoides* with 99% identity in nucleotide sequence to *TaCAF1Ia1* (**Fig. S6b**), which had not been found and annotated before, together with other highly homologous sequences. Similarly, we conducted a genome-wide deep scanning with sequential excavation method for the identification of *TaCAF1Ia1* homologues in common wheat, and found many unannotated sequences (**Table S2**). Finally, a total of 46 homologous genes were identified in chromosome groups 3, 4 and 6, all of which comprised the *TaCAF1Ia* subfamily in wheat (Fig. 2**a**). Multiple alignment of DNA sequences showed more than 70% identity among these 46 subfamily members. In particular, in chromosome groups 4 or 6, most of the sequence pairs had more than 90% intra-group identity (Fig. 2**b**). Statistically, 10 homologous members had large fragments deletions leading to the loss of a start codon. The other 36 members have putative ORF sequences, and like *TaCAF1Ia1*, each member has only one exon. Analysis of the amino acid sequences alignment revealed that 4 of the 36 sequences had frameshift mutations (Fig. 2**a**, **Table S2**) due to the nucleotide base mutations or fragment deletions. Finally, according to the similarities between sequences and the physical distances of these members on the chromosomes, we constructed x/y/z/m loci for defining the structures of gene clusters formed by *TaCAF1Ia* subfamily members on wheat genome. We then divided *TaCAF1Ia* members into four gene types of Y-like/Y/Z/M according to the four sites of x/y/z/m, respectively (Fig. 2**a,b**). After that, the phylogenetic tree for all wheat TaCAF1 proteins of this study was reconstructed together with all CAF1s from *Pharus latifolius* which is one of the basal taxa of the Poaceae [43], *Oryza sativa*, and *Arabidopsis thaliana*. The results showed that all *Arabidopsis thaliana* and *Pharus latifolius* members were only present in Group 1, hence, *CAF1s* in group 1 were likely the core members maintaining the conservation (Fig. 2**c**). Interestingly, the group 4 only comprised members from the whole *TaCAF1Ia* subfamily (Fig. 2**c**), suggesting *TaCAF1Ia* subfamily should be specialized in wheat.

Furthermore, all *TaCAF1Ia* members can be divided into three clades. Members in each clade are all from the same chromosome group, except the 4Dy, which is grouped together with the members from chromosome group 3 (Fig. 2**d**). The gene members on chromosome group 3 could be divided into three sub-clades, in which the members from each of the Y, Z and M types were individually aggregated together to form Ia, Ib and Ic groups, respectively. The amino acid sequences alignment also showed that there were large structural variations among Y, Z and M members on chromosome group 3 (**Fig. S7a**). It coincides with the result of motif analysis which revealed the structural diversity of the members from chromosome group 3 (**Fig. S7b**). Interestingly, the *TaCAF1Ia* members on chromosome groups 4 and 6 have no prominent changes in amino acid sequences (**Fig. S7c**). Furthermore, the motifs of each member from chromosome groups 4 and 6 except 4Dy were the same as those of Z type members from chromosome group 3 (**Fig. S7b**). Thus, the members from x and y sites in chromosome groups 4 and 6 were defined as Z-like type genes. Similarly, the truncated genes of the 3Ax site, which shared fairly high homology with the genes of y sites, were defined as Y-like type members (Fig. 2**a,b**). Without exception, all *TaCAF1Ia*

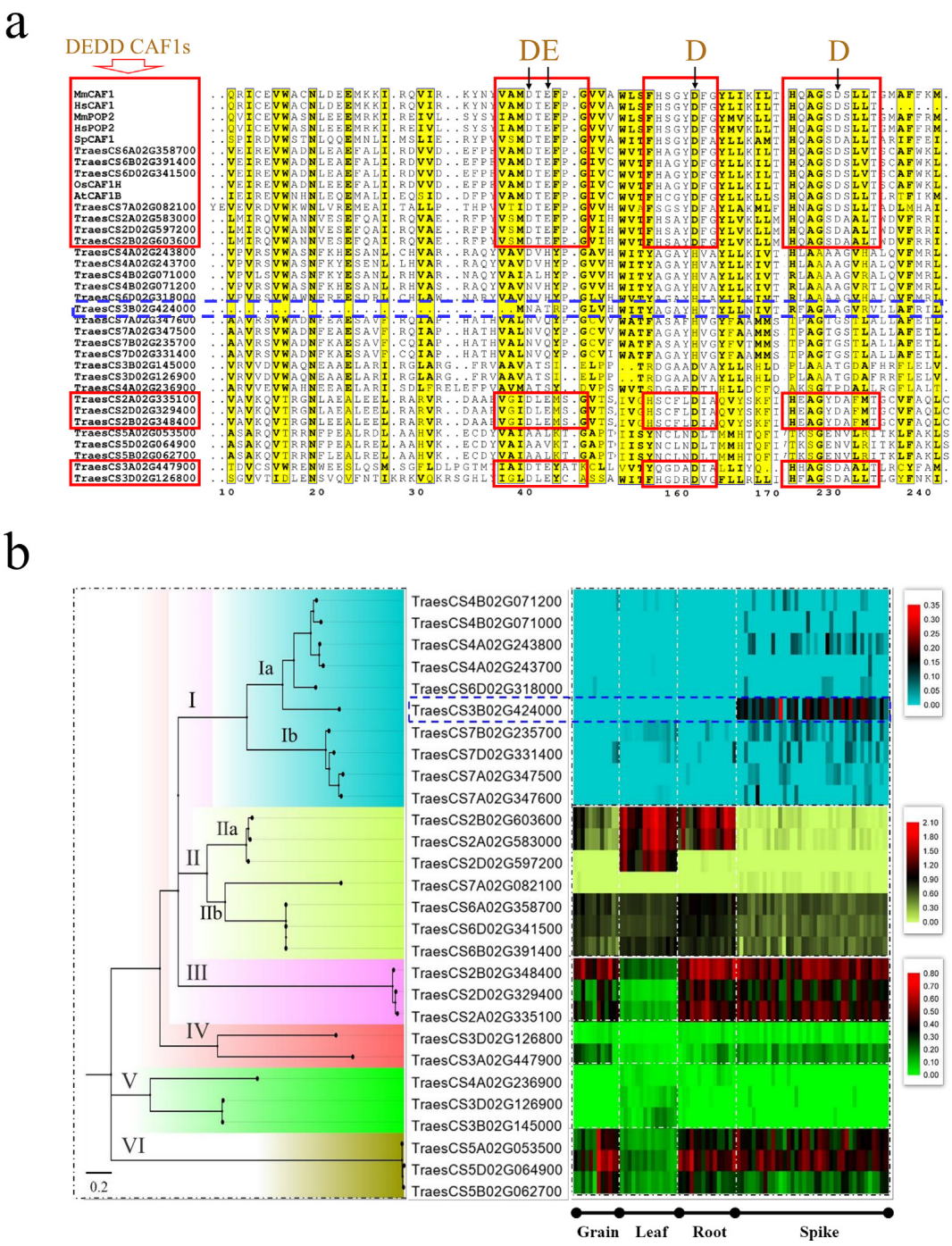


Fig. 1. Phylogeny and identification of the *CAF1* family. (a) Amino acid sequences alignment of the 28 TaCAF1 family members with representative CAF1s from other model species. The 12 TaCAF1 members with three conserved DEDD motifs are indicated with red boxes. The rest of the family members are non-typical TaCAF1s without DEDD motifs in their sequences. The *TaCAF1Ia1* (TraesCS3B02G424000) marked with blue box is a non-typical *CAF1* gene. (b) Phylogenetic relationship and expression pattern analysis of the CAF1-domain proteins from common wheat. Maximum likelihood phylogeny of the 28 *TaCAF1* family members was constructed based on the full-length protein sequences using MEGA 7.0 software. The JTT + G (5 categories) amino acid substitution model was used with 1000 bootstrap replicates to assess tree reliability. Expression of *TaCAF1* genes in the major organs of common wheat. The expression data of the 28 wheat *TaCAF1* genes were downloaded from Wheat Expression Browser (wheat-expression.com). We specifically chose the expression profiles of these genes in the major tissues including grain, root, leaf and spike at the different developmental stages referring to a total of 81 samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

members do not contain DEDD motifs belonging to non-typical *CAF1s* (Fig. S7**d**).

Characterization of CAF1Ia subfamily members in wheat progenitor species

The analysis of divergence time showed that the expansion of wheat *TaCAF1Ia* subfamily was attributed to the multiple gene duplication events since 21 Mya (Fig. S8**a**). Comparative analysis revealed that the gene cluster occurred by tandem gene duplications. For better distinguishing the *TaCAF1Ia* members on chromosome group 3, we defined the members among the four sites of y/z/m as old tandem duplications and the inner members in each one site belonging to new tandem duplications (Fig. S8**b-d**). One of the replicative pairs 3Ay-1 and 3Ay-2 sharing 99.4% similarity in nucleotide sequences was the most recent duplication event that

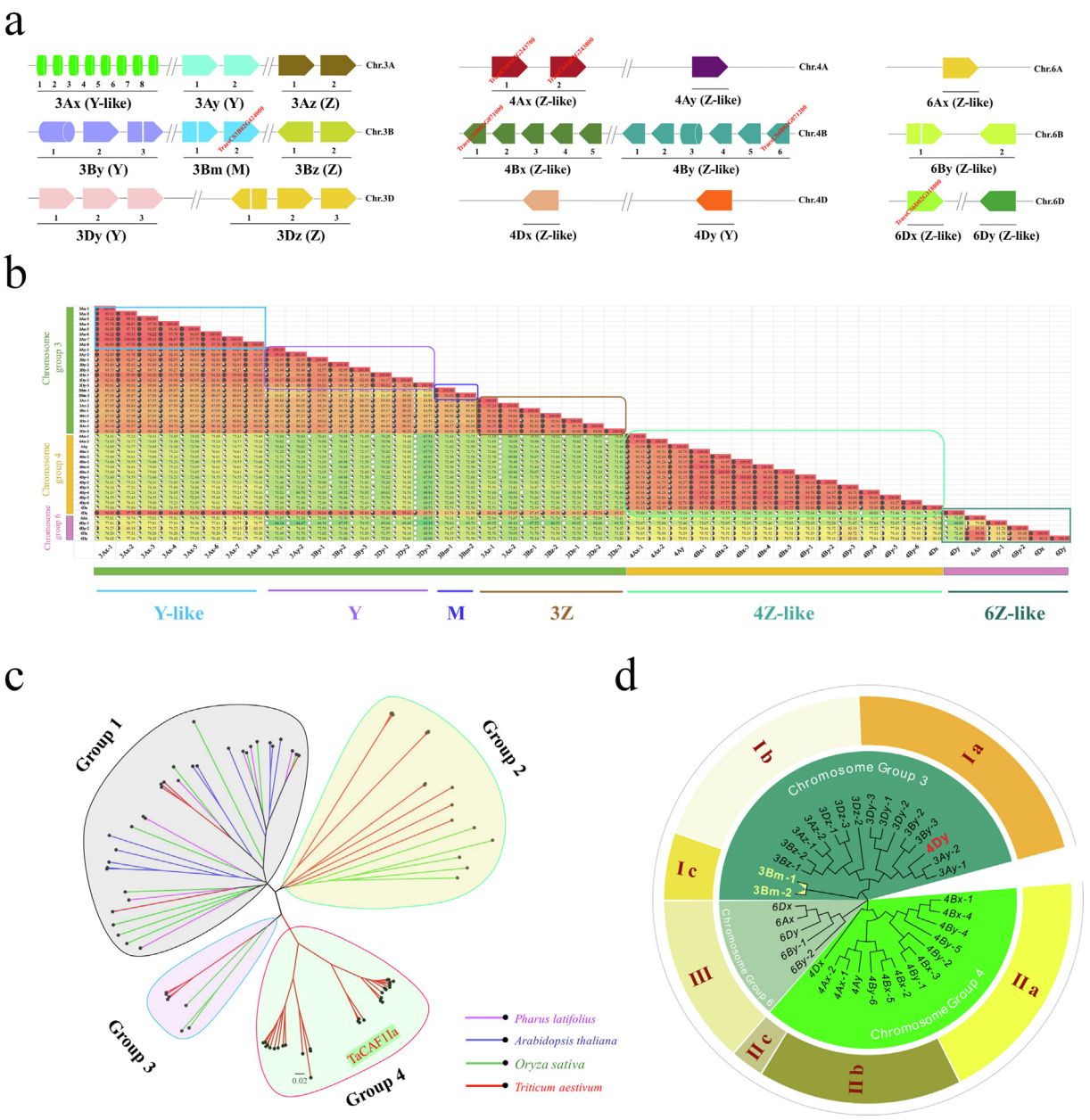


Fig. 2. Identification of the *TaCAF1Ia* subfamily members in common wheat (Chinese Spring) genome and construction of the collinear and phylogenetic relationships of the *TaCAF1Ia* genes. (a) Identification and classification of the *TaCAF1Ia* members on chromosome groups 3, 4 and 6 in common wheat. The cylindrical form indicating the *TaCAF1* gene had a DNA fragment deletion leading to the loss of the start codon. The cusp denotes the direction of the encoding sequence on the chromosome. The vertical line inside the pentagonal arrow indicates a frameshift mutation leading to a premature termination. (b) The percent identity matrix of the *TaCAF1Ia* subfamily members in common wheat. The similarity of the *TaCAF1Ia* nucleotide sequences ranges from approximately 70% to 100%. (c) Phylogenetic relationship between *Triticum aestivum* and other plant species. Phylogenetic tree of the TaCAF1 proteins with all CAF1s from *Oryza sativa*, *Arabidopsis thaliana*, and *Pharus latifolius* was constructed using the NJ method with MEGA 7.0. All members of the *TaCAF1Ia* subfamily of wheat formed a separate clade of the group 4. (d) Phylogenetic relationship among *TaCAF1Ia* subfamily genes from different chromosome groups in common wheat. The phylogenetic tree of the 36 *TaCAF1Ia* subfamily members was constructed based on the full-length nucleotide sequences by the neighbor-joining method using MEGA 7.0 software. *TaCAF1Ia* subfamily genes were grouped into three clades represented by different colors.

occurred 0.6 Mya ago (Fig. S8**b**). In addition, the collinear comparison with the ancestor species of common wheat also showed that the structural variation and gene duplication of the *CAF1a* family had been continuously occurring in ancestor species (Fig. S9). On chromosome group 3, the x/y/z sites in common wheat were generated early in each diploid species of AA and DD (Fig. S9**a**). There were few changes of duplication in subAB genomes between tetraploid and hexaploid (Fig. S9**b**), no change between diploid DD and hexaploid subD genomes (Fig. S9**a,c**). Interestingly, the changes of m sites among AABB and AABBDD genomes are complicated, including gene deletion, fragment deficiency and base mutation of frameshift. Surprisingly, 3Bm-2 (TraesCS3B02G424000/*TaCAF1Ia1*) is quite conservative. Most of the *CAF1Ia* members on chromosome groups 4 and 6 also had already existed in the diploid ancestor species, but there was almost no variation in their gene structures, most of which belonged to Z-like type genes (Fig. S9**d,e**). Interestingly, it was found that 4Dy in common wheat may have derived from the 3Ay site, which shared 99% identity with its ortholog on 4D chromosome of the diploid DD genome (Fig. S10).

Collinear relationships of CAF1Ia among the homologous chromosome groups in different distantly related species

In order to further trace the replication events and structural variation, we identified *CAF1Ia* members in Triticeae species, such as *Thinopyrum elongatum* (EE), *Hordeum vulgare* (HH), *Secale cereale* (RR), *Aegilops sharonensis* ($S^{sh}S^{sh}$) and *Aegilops speltoides* (SS) (**Fig. S11**). The results showed that z sites all existed on homologous chromosome group 3 of these five species, but x, y and m sites on homologous chromosome group 3 were different among these species. Interestingly, the m site was only identified on chromosome 3 of SS genome (**Fig. S11a**). It was assumed that m site should be specially evolved in the common ancestor species of BB and SS genomes. It is indeed a specialized locus in the chromosome group 3. All sites of *CAF1Ia* on chromosome groups 4 and 6 of these species belonged to Z-like type genes (**Fig. S11b,c**). Furthermore, 3Bm and 3Sm genes are uniquely clustered into a distinct branch in isolation (**Fig. S11d**). Thus, this demonstrated that *Aegilops speltoides* is indeed the closest relative of BB progenitor of common wheat.

Expression pattern analysis of the TaCAF1Ia subfamily members

To further understand the mechanism of this novel *TaCAF1Ia* subfamily and how plant deals with functional redundancy when so many homologous genes integrated into the hexaploidy, we first selected *CAF1* members from different sites in different chromosome groups for the large-scale analysis of gene expression patterns. A total of 17 members have been verified using GUS histochemical assays after genetic transformation of *TaCAF1Ia* promoters with the *GUS* reporter gene to *Arabidopsis*. Only 4 members, pro3Bm-1, pro3Bm-2, pro4Ax-2 and pro6By-1, were found to be specifically and highly expressed in anthers (Fig. 3, S12-S15). Interestingly, the expression pattern of 3Bm-2 basically coincided with that of tae-miR2275 in anthers, and both no longer expressed after the 10th stage (Fig. 3**a,b**). However, the expression of 4Ax-2 started from the 10th stage, complementary to that of 3Bm-2 (Fig. 3**c**). In addition, the promoters of 3Bm-1 and 6By-1 drived high level of *GUS* expression (**Fig. S15**). However, even if these two genes were expressed, there would be no functional protein product due to the mutations in their coding sequences (Fig. 2**a, Table S2**).

Reestablishing the interaction pattern of CCR4-NOT complex with non-typical CAF1s in wheat

Our data revealed that the TaCAF1Ia members were all non-typical CAF1s without DEDD motifs (**Fig. S7d**). To find out if these non-typical CAF1s also could form CCR4-NOT complex, the interaction relationships between TaCAF1Ia members, TaCCR4 and TaNOT1 were examined (Fig. 4**a, Fig. S16-S17**). It was found that the Y and Z types proteins had interaction activity with the N-terminal (1–216) of the CCR4 protein, while Z-like proteins interacted with the C-terminal (217–605) of CCR4 (**Fig. S16a**). Taking the protein TraesCS6A02G358700 (TaCAF1H) with DEDD domain and sharing 91% protein sequence identity with the rice OsCAF1H as the positive control, it was shown that TaCAF1H could also interact with the N-terminal of TaCCR4 at the amino acids 63–120, similar to the result of OsCAF1H-OsCCR4 interaction in rice [20]. We also found that the other regions of the amino acids 121–216 of OsCCR4a were able to interact with these Y/Z/M types of TaCAF1a members without DEDD domain, rather than the regions of the amino acids 63–120 (**Fig. S16b**). In addition, it is assumed that the four motifs of 2, 3, 5 and 7 in TaCAF1Ia members are critical for TaCAF1Ia and TaCCR4 to be able to interact (**Fig. S16b**). Likewise, except for TaCAF1H which is like the OsCAF1H in rice interacting with MIF4G domain of OsNOT1, all checked TaCAF1Ia proteins do not interact with MIF4G (Fig. 4**a**). However, 3Ay-2, 3By-2, 3Bm-2 and 3Dy-2 can interact with TTP domain of OsNOT1 (**Fig. S17**). In addition, TaCCR4 could not interact with TaNOT1 protein (**Fig. S16c**). Interestingly, 3Ay-2 and 3Bm-2 also interacted with the C-terminal of CCR4 (**Fig. 16a**). Therefore, all our results confirmed that Y and M types of TaCAF1Ia could form the interaction model of CCR4-CAF1-NOT1, but the Z and Z-like types of TaCAF1Ia may only form the interaction model of CCR4-CAF1. Most notably, 3Dy-1, lacking motif 7 (**Fig. S7b**), did not interact with TaCCR4 and TaNOT1. Thus, the motif 7 of non-typical CAF1s likely plays a vital role in forming wheat CCR4-NOT complex. The BiFC and pull-down assay also supported the interaction results (Fig. 4**b,c and S17b,c**).

Comparison of haplotype composition with the conserved TaCAF1Ia1 gene among wheat populations of different ploidy levels

The sequences of the *TaCAF1Ia1* gene in hexaploid and tetraploid wheat were found to be conserved with little difference (**Fig. S6**). This prompted us to further explore the origin and the evolution history of *TaCAF1Ia1*. A large number of wild emmer, durum wheat and common wheat were used for genetic diversity analysis based on PCR method. Because the alleles of *TaCAF1Ia1* obtained were significantly different from 3Bm-1 gene in SNP loci of polymorphism (**Fig. S18**), the amplification of *TaCAF1Ia1* was surely accurate. A total of 16 single nucleotide polymorphisms (SNPs) and 16 haplotypes were identified from all 833 wheat accessions (Fig. 5**a**). Haplotypes were compared and shown in **Table S3**. Ten haplotypes were detected across 294 wild emmer accessions. Of these, 7 haplotypes were identified from 233 (79.25%) accessions in Israel, 5 haplotypes from 22 (7.48%) accessions in Lebanon, 4 haplotypes from 24 (8.16%) accessions in Syria, 4 haplotypes from 15 (5.10%) accessions in Turkey (**Table S4**). Notably, the Hap2 was present in 211 (71.77%) wild emmer accessions (**Table S5**). Five haplotypes were detected across 238 durum wheat accessions. Of these, Hap4 was the most dominant haplotype observed in 174 (73.1%) accessions. Hap2 and Hap8 were the two main haplotypes in common wheat, present in 150 (49.8%) and 142 (47.2%) accessions, respectively (**Table S5;** Fig. 5**b**).

Bayesian evolutionary tree showed that 16 *TaCAF1Ia1* haplotypes were clustered into four distinct clades, and the earliest ancestral genotype was differentiated around ~0.8 Mya years ago (Fig. 5**b**). Almost all the haplotypes of clades I and II were present in common wheat. Particularly, clade III only appeared in wild emmer and clade IV was not detected in common wheat. Interestingly, all the haplotypes of clade I were absent in wild emmer from Israel, thus it was most likely to have evolved independently as a non-Israeli originated clade (Fig. 5**b**). Consistently, these results were further supported by the haplotype network analysis, which showed that the haplotypes of three populations were distinctly separated into four groups (Fig. 3**c**). Hap2, Hap4 and Hap8 are three largest haplotypes among three natural populations. Hap2 and Hap8 were the shared haplotypes for all these three natural populations of wheat (Fig. 5**b,c**). According to the result of phylogenetic relationship and origin region of wild emmer accessions, the four clades were regarded as four lineages of AABB progenitors in this study (Fig. 5**b**). In addition, the genetic diversity was different among these three populations. The lowest level of diversity in both haplotype and nucleotide was found to be in durum population (**Table S6**). The NJ tree showed that the wild emmer was more closely related to common wheat (**Fig. S19**). These results implied that other cultivated tetraploid wheat may be involved in the origin of wheat.

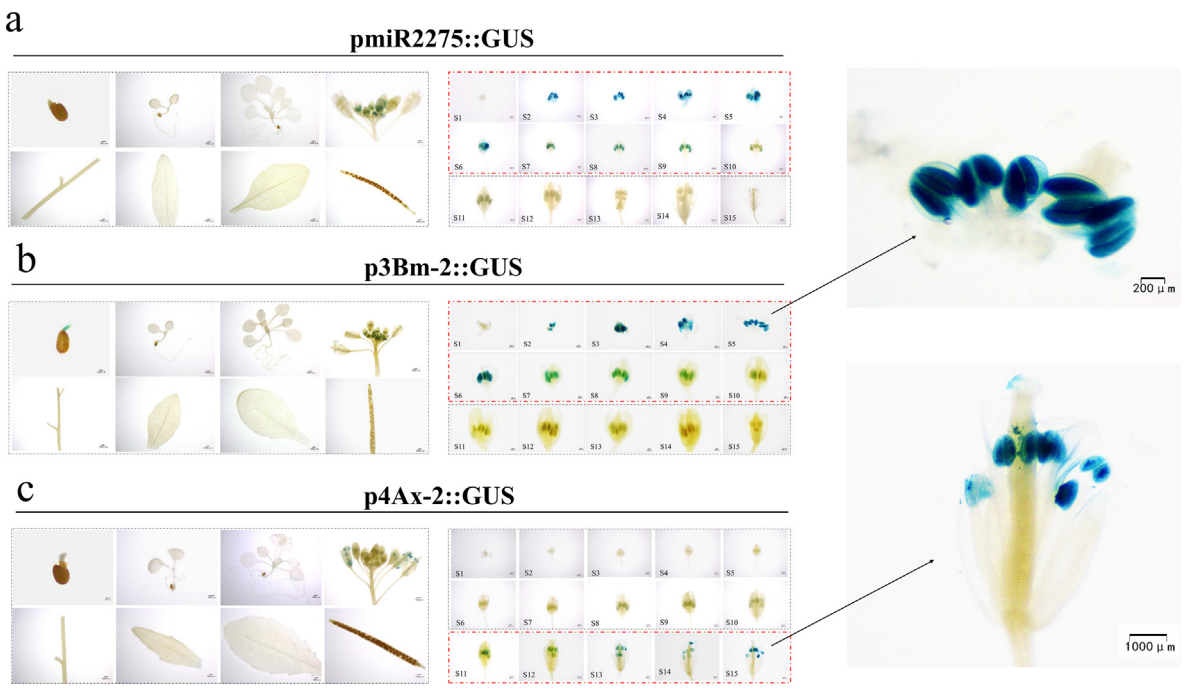


Fig. 3. Gene expression profiling by GUS assay. The promoter::GUS fusion construction was used to reveal expression patterns of the tae-miR2275 and the *TaCAF1la* subfamily members in transgenic *Arabidopsis*. The tissues where GUS reporter gene expressed turned blue in the GUS staining buff. The buds from the same bouquet of an inflorescence were sampled and then arranged in the order of their sizes from the smallest to the largest, representing developmental stages of S1-S15. Bars = 200 μm, 500 μm, and 1000 μm, respectively. (a) As revealed by GUS staining, tae-miR2275 was undetectable in the germinated seed, young seedling, stem, rosette leaf, cauline leaf and silique, but highly induced in buds from S2-S10. (b) As revealed by GUS staining, 3Bm-2 (*TaCAF1la1*) was undetectable in the young seedling, stem, rosette leaf, cauline leaf and silique, but highly expressed in buds from S2 to S10. (c) As revealed by GUS staining, 4Ax-2, was undetectable in the germinating seed, young seedling, stem, rosette leaf, cauline leaf and silique, but highly induced in buds from S11 to S15. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

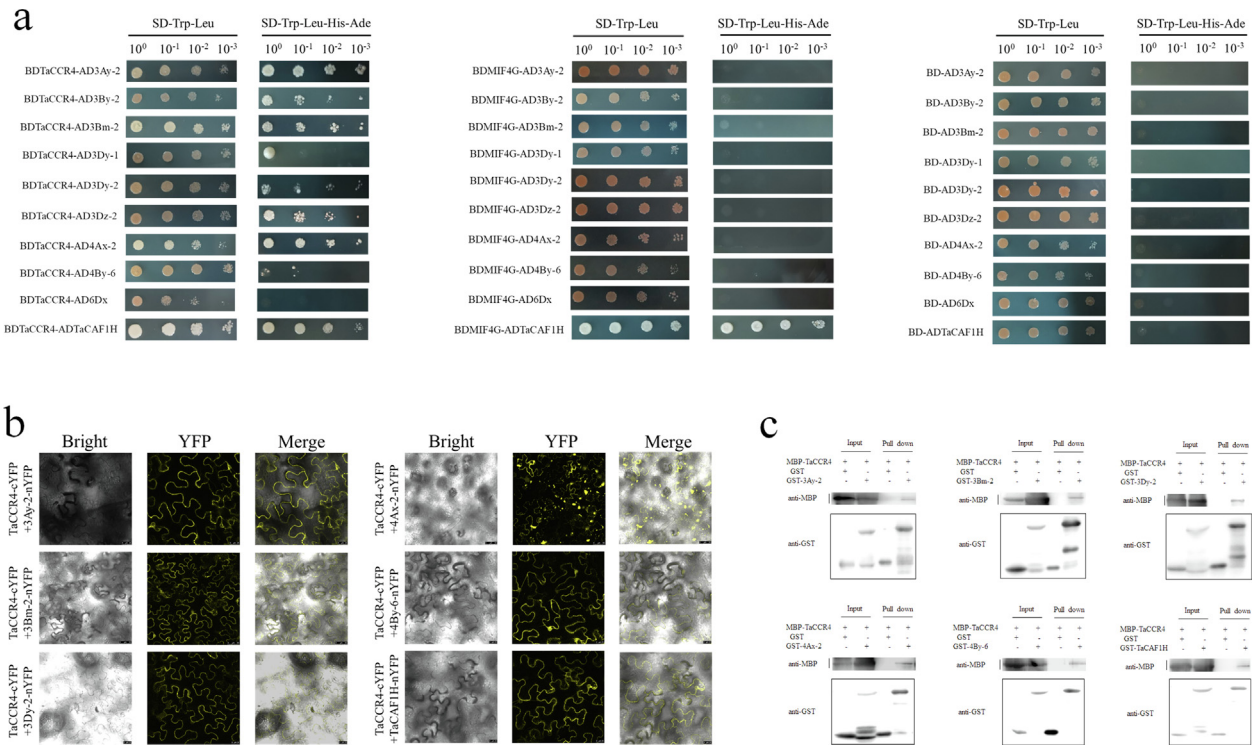


Fig. 4. The protein–protein interactions among TaCCR4, TaCAF1s and TaNOT1. (a) Analyses of the interactions of TaCCR4 or TaNOT1 (MIF4G domain) with the different TaCAF1la members by yeast two-hybrid assay. BDTaCCR4-ADTaCAF1H is a positive control. TaCAF1H is the TraesCS6A02G358700 which has the high homology with OsCAF1H. BD, pGBK-T7; AD, pGADT7. (b) TaCCR4 interacted with different TaCAF1s members in BiFC assays using tobacco leaf epidermis. Scale bar = 50 μm. (c) Confirmation of the interaction between the full length TaCCR4 and TaCAF1s using Pull-Down assays.

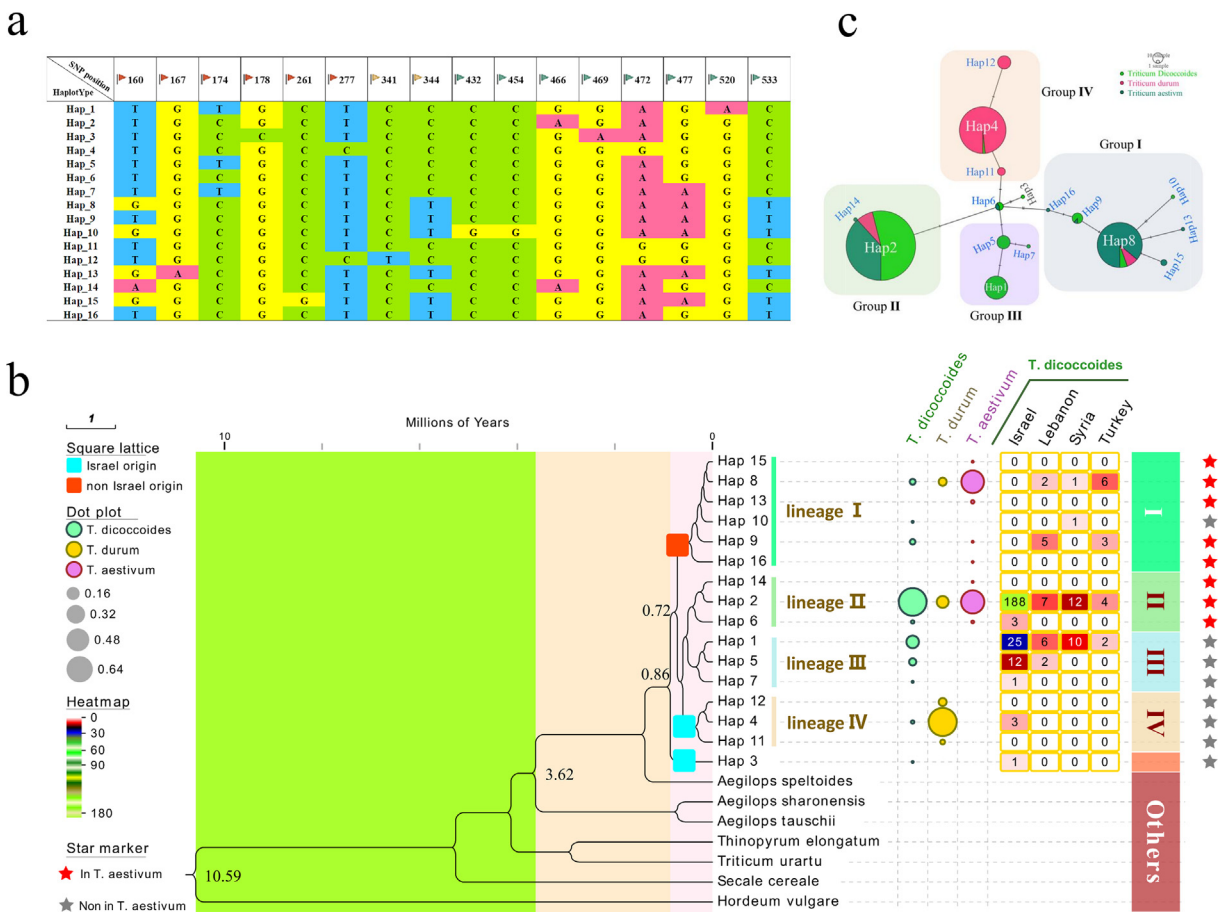


Fig. 5. Identification and phylogeny analysis of the haplotypes of *TaCAF1Ia1* using different wheat species. (a) Distribution of polymorphic SNPs across the sixteen *TaCAF1Ia1* haplotypes. The colours represent the different bases. (b) Bayesian evolutionary tree was constructed using all haplotypes of the *TaCAF1Ia1* gene from different wheat species, with homologous *CAF1Ia1* gene from different Triticeae species as the outgroup. The light-blue lattice represents the clade derived from Israel and the red lattice represents the clade derived from non-Israel regions. The size of the dot plot indicates the frequency of the haplotype. The numbers of different wild emmer accessions for each haplotype from Israel, Lebanon, Syria and Turkey were used to create the heatmap. The 16 haplotypes of *TaCAF1Ia1* were clustered into four different clades which were defined as four distinct lineages, and Hap3 was differentiated as the earliest ancestral genotype. (c) The haplotype network construction for *TaCAF1Ia1* gene. Each circle represents a single haplotype. The size of the circle indicates the frequency of the haplotype. All 16 haplotypes are grouped into four groups. The group **I** and **II** are the two largest groups containing the three wheat species. The group **III** only includes *T. dicoccoides*. The group IV includes *T. dicoccoides* and *T. durum*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Discussion

Evolution of the CAF1 family

Ciona intestinalis, the lowest group of chordates, has one *CAF1* member in their genomes. Even the highest group of mammals has only two members of *CAF1* and *POP2*. Similarly, there is only one single gene family member in the single-celled algae, such as *Chlamydomonas* (**Table S7**). Further, only two members were found in *Physcomitrella patens* known as the early evolved land plant [44]. Unlike what is observed in lower plants, *CAF1* becomes a polygenic family in higher plants, such as *Arabidopsis thaliana* and rice (**Fig. S2a,b**). Therefore, after the emergence of angiosperm, it is unknown what fate and evolutionary events of *CAF1* experienced to cause the large-scale expansion of the *CAF1* gene. Most likely, *CAF1* may have undergone large-scale replications during the frequent partial or whole genome duplication (WGD) events in flowering plants [45]. Moreover, unlike animals, plants are sedentary and need to evolve a range of mechanisms to respond to environmental changes, which may involve in the expansion of the *CAF1* family. The roles of CAF1 in stress response have been reported in many plants [22–25]. But studies in mammals have shown that the function of CAF1 is quite important in reproductive development [16–19]. The ortholog-function conjecture infers that the orthologous genes in different species perform similar functions [46]. However, there has been no report on the possible role of CAF1 in plant reproductive development.

The structure variation and evolution of the TaCAF1Ia subfamily

Based on the observation and comparing *TaCAF1Ia* members on chromosome group 3 with their synteny members from ancestral species and other Triticeae species (**Fig. S9a,b and S11a**), we speculated that the z site members may be the initial type of *TaCAF1Ia* on the homologous chromosome group 3. By contrast, all the genes on the homologous chromosome groups 4 and 6 which share a similar structure to the Z type genes of the homologous group 3 were considered to be the Z-like type members (**Fig. S9d,e and Fig. S11b,c**). Therefore, we propose that the Z or Z-like types are probably the oldest types in *TaCAF1Ia* subfamily (**Fig. S20**). The Z or Z-like type of *TaCAF1Ia* likely originated from segmental duplication events in the earliest Triticeae ancestor (**Fig. S20**), which may be associated with WGD or reverse retrotransposition [47]. The Y-type *CAF1* was then differentiated from the Z-type gene during an ancient tandem replication event which may happened in a common diploid ancestor for some of the Triticeae species. Along

with the independent evolutionary process in AA and BB progenitors, the Y-like and M types *CAF1s* were also derived from tandem duplication events, respectively (**Fig. S20**). Meanwhile, new tandem replication events occurred continuously within each site (**Fig. S8**). These large amounts of gene duplications may be related to ectopic recombination, replication slippage or retrotransposition during biological evolution in organisms [48]. Moreover, recent reports allege the origin of the DD genome through homoploid hybridization between AA and BB lineages [49]. Here, the 4Dy is not duplicated from any one member of the chromosome group 4, but highly homologous with 3Ay (**Fig. S10**). It is possible that 4Dy derived from the segmental duplication of the 3Ay site during the cross between AA and BB ancestors. In addition, we also provide an important molecular evidence that the S genome is the closest relative of the B genome [50]. Because except in the wheat subB genome, M-type *CAF1* (3Bm/3Sm) was present only in the S genome (**Figs S9, S11**). In short, as a novel gene resource, *TaCAF1Ia* subfamily will be very useful for the phylogenetic origin and evolution analysis in Triticeae.

Wheat and its progenitors have evolved multiple sophisticated and nuanced mechanisms to deal with the continuous duplications of the CAF1Ia subfamily

How would the functional redundancy be dealt with when so many duplicated genes were continuously generated? Here, we found that several strategies appear to have impacted the gene structures since the beginning of the diploid ancestors, including nucleotide base mutation, segment insertion or deletion and others (**Fig. S9**). Furthermore, plants also control gene structure by multiple complicated and nuanced manners. Firstly, according to the results obtained by a large-scale verification of the promoter activities, most of the *TaCAF1Ia* members was silenced in hexaploidy (**Figs. S12-S15**). The cause of these promoter inactivation might be related to DNA methylation, a main approach for wheat to deal with redundancy among homologous chromosomes caused by hexaploidy [51]. Of course, it is also possible that the promoters lack the necessary regulatory sequences during the formation of new duplicated genes, some of which might be pseudogenes [52]. Secondly, through protein interaction analysis of this novel TaCAF1Ia subfamily, it was found that only several CAF1s could form the interaction model of the CCR4-NOT complex by the new interaction sites (Fig. 4**, Fig. S16-S17**). Therefore, the organism prompts many *TaCAF1Ia* members to lose their abilities and activities to handle the functional redundancy in several ways, such as promoter mutation and gene structure mutation. That is, the function performance of homologous genes from different chromosome groups may be subjected to the special selection mechanisms in polyploid species. Additionally, *TaCAF1Ia1*, the target gene of taemiR2275 who was co-expressed with *TaCAF1Ia1*, could be specifically expressed in early stages of anther development (Fig. 3**a,b**). Therefore, it is assumed that a special miR2275-CAF1 pathway was evolved for regulating reproductive development in wheat.

The haplotype diversity of TaCAF1a1 was suitable for the traceability analysis on the maternal progenitor of the common wheat

It was previously reported that multiple domestication sites of wheat probably existed in the eastern Mediterranean and the domesticated tetraploid wheat evolved as polymorphic populations rather than single genotypes [53,54]. In this study, we found that not only *CAF1Ia* is suitable for inferring phylogeny of Triticeae species, but also the difference of *TaCAF1Ia1* haplotypes should be suitable for tracing the maternal progenitor origin of the common wheat. We suggest that there are at least four lineages for AABB genome. Among them, lineages I and II were the progenitors for forming AB subgenome in the hexaploid wheat (AABBDD) (Fig. 6**a**). The AABB ancestry of the lineage I originated from the non-Israeli areas (Fig. 5**b**), especially in Turkey. Additionally, although the lineage II contains both Israeli and non-Israeli germplasm resources, we also considered that the non-Israeli origin accessions from the lineage II may be also involved in the formation of common wheat. Because previous studies showed that wild emmer in northern Levant (southeastern Turkey) was most likely the direct progenitor of AB subgenome of wheat [5], and the core area for agriculture is also thought to be emerged in Turkey and Syria [11]. Moreover, the presence of sibling species both in wild tetraploid and wild diploid wheats suggests single rather than multiple events of domestication [55]. However, it was also believed to be a reticulate origin of domesticated emmer based on phylogenetic results and implicit archaeological evidences [56]. The wild emmer of the lineage III have long diverged along the evolutionary path and may have been genetically or geographically different, so they may not have been domesticated (Fig. 6**b**). Here, another origin for tetraploid wheat was identified to be domesticated from wild emmer of the lineage IV, but this lineage was not involved in forming the hexaploid wheat (Fig. 5**b**). According to the above models, we propose the possible original modes of the dual-lineages origin for the common wheat and the potential three-lineages origin for the cultivated tetraploid wheat (Fig. 6). In addition, it is likely that the AABB progenitors of the lineages I and II collected and domesticated by ancient people encountered the crucial opportunity to contact with the DD progenitor at one or more early agricultural areas [57]. Furthermore, it is assumed that such an origin module is difficult to take place because of the well-known hybridizability between different species of wheat for generating mutual-introgression. Unlike in rice keeping the reproductive isolation between subspecies of *indica* and *japonica* [58], the continuous mutual-crosses occurred between the two wheat races from the lineages I and II, add additional difficulties in studying the wheat origin (Fig. 6**a**). However, it is probable that the lineage IV which were involved in the domestication of another series of tetraploid wheat were independently domesticated and may result in geographical or reproductive isolation with DD (Fig. 6**c**). Based on ecological and genetic polymorphism information of *Triticum dicoccoides,* the previous study has already suggested that the center of origin and diversity is northeastern Upper Galilee and the Golan. Elsewhere it occurs just like many "archipelagos" in semi-isolated and isolated populations in the Fertile Crescent [59]. Therefore, it is likely that only some groups were discovered and used by ancient people. Moreover, Hap2, Hap4 and Hap8 could be the dominant haplotypes providing examples of the gene frequency-oriented changes in population caused by the origin of agriculture and artificial intervention [60], which allowed them to be retained and extended.

We then evaluated the accuracy and validity of the model by analyzing the published resequencing data of 642 hexaploid accessions collected worldwide [3,5,7,30,31]. The results demonstrated that our model is undoubtedly very effective. Similar to the above result of haplotypes identification from the Chinese core germplasm of the hexaploid wheat, Hap2 (40.8%) and Hap8 (58.4%) were the dominant haplotypes in 642 hexaploid accessions. Only one accession contained the rare haplotype, Hap17, which might be a derivative haplotype of Hap13 (**Fig. S21**). Even more remarkably, we found that the haplotype of the remaining 4 hexaploid accessions was identified as Hap4 of the lineage IV. All these four accessions belong to the semi-wild wheat, *Triticum petropavlovskyi* (**Fig. S21**). Interestingly, previous study indicated that *Triticum petropavlovskyi* may be derived from AABB and AABBDD hybridization, but not AABB and DD hybridization [61–63]. These findings have been interpreted as evidence that the lineages IV did not participate in the origin of hexaploid wheat. Thus, the

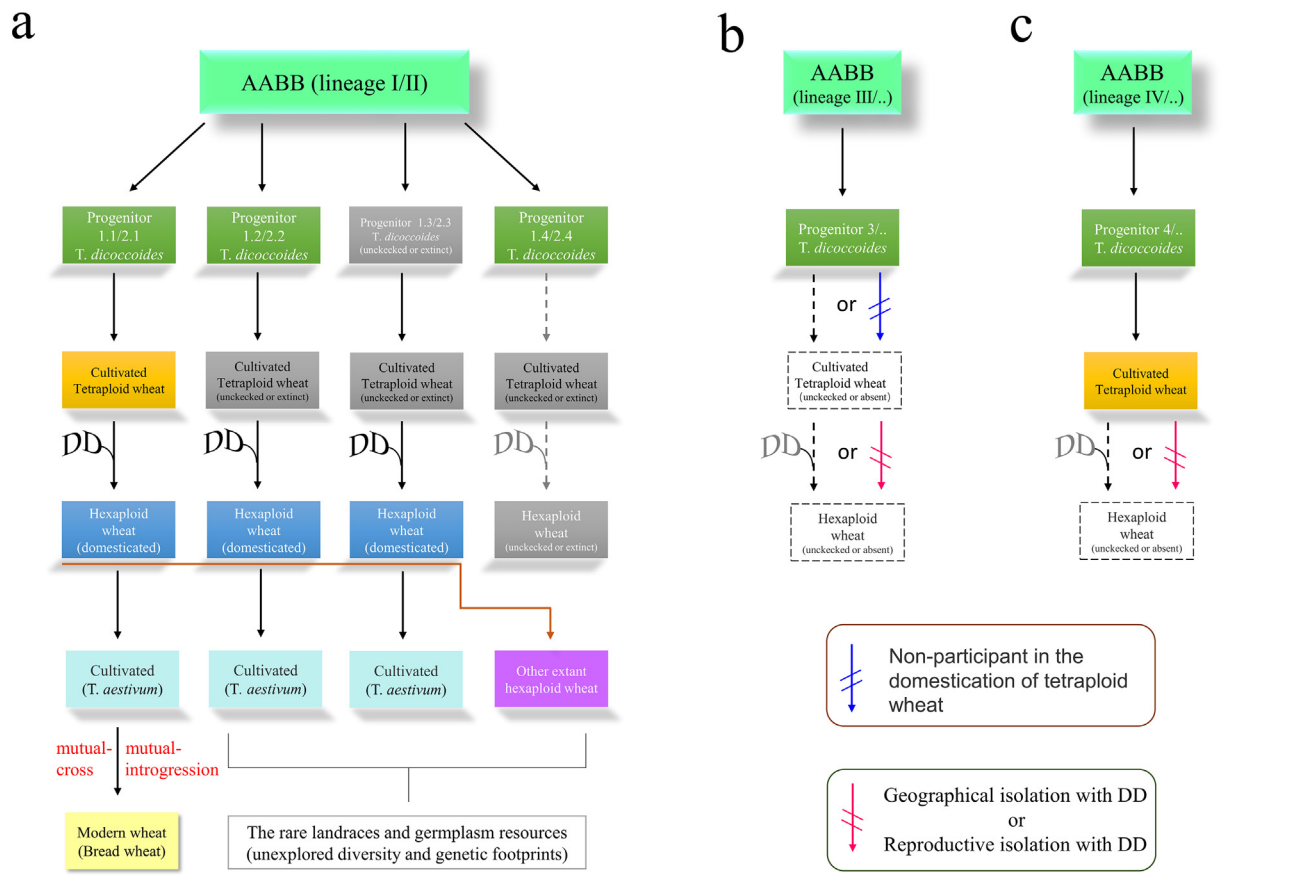


Fig. 6. Hypothetic schematic representation of the evolutionary history of the tetraploid and hexaploidy wheat. The proposed model was based on the results of haplotypes identification of the *TaCAF1Ia1* sequences in the wide range of germplasm resources of wild emmer, durum wheat and common wheat populations. (a) The model of the origin for the AABB genome contributing to form the common wheat. Diagrammatic illustration showed the two divergent lineages (I/II) of the tetraploid ancestries, each of which may contain four types of evolutionary lines. Many tetraploid progenitors were extinct. The rare germplasm resources, such as some landraces, may provide many unexplored diversity and genetic footprints for modern wheat breeding. (b) The AABB ancestry was not involved in the domestication of tetraploid wheat, or they had geographical or reproductive isolation with DD genome. Here, there may exist other undiscovered lineages. (c) The AABB ancestry was involved in the domestication of tetraploid wheat, but never participated in the hybridization with DD genome. Also, there may exist other undiscovered lineages.

dual-lineages original mode for the hexaploid wheat are highly valid. However, the domesticated trajectories of the tetraploid wheat did not completely overlap with the hexaploid wheat. Hap2, Hap4 and Hap8, the three main haplotypes shown to be from three lineages, were identified in other different cultivated tetraploid varieties collected worldwide (Fig. S21). Therefore, the potential three-lineages origin mode for the cultivated tetraploid wheat was also considered.

Conclusions

Overall, here we propose a new dual lineages origin model in common wheat and suggest the possible existence of the three lineages of domestication model in the cultivated tetraploid wheat. It is important to verify the model presented in this study with additional protein families. Moreover, as a conservation subfamily, *CAF1Ia* will be very useful for the phylogenetic analysis of Triticeae. Using the specificity of *TaCAF1Ia1* gene we could easily explore undeveloped footprint in the common wheat, which will be valuable for wheat breeding. As an extensively duplicated subfamily in the polyploid species, the function-performance of homologous genes of *TaCAF1Ia* may be regulated by many special selection and control mechanisms, and as a non-typical *CAF1* subfamily, *TaCAF1Ia* will bring the fresh insight for studying the function of plant *CAF1s*.

Compliance with Ethics Requirements

This article does not contain any studies with human or animal subjects.

CRediT authorship contribution statement

Longqing Sun: Conceptualization, Methodology, Investigation, Validation, Data curation, Writing – original draft. **Ruilian Song:** Methodology, Investigation, Validation. **Yixiang Wang:** Investigation. **Xiaofang Wang:** Validation. **Junhua Peng:** Resources. **Eviatar Nevo:** Resources, Writing – review & editing. **Xifeng Ren:** Supervision, Funding acquisition. **Dongfa Sun:** Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the China Agriculture Research System of MOF and MARA (CARS-5), Accurate Identification of

the Wheat Germplasm Resources in the Mid-lower Reaches of Yangtze River (19211092), National Key Research and Development Program of China (2016YFD0100502), National Natural Science Foundation of China (31601294) and China Postdoctoral Science Foundation (2015M582241). We also thank the Ancell-Teicher Research Foundation for Molecular Evolution for funding *Triticum dicoccoides* studies. The funding agencies had no role in the design of the study, in the collection, analysis, and interpretation of data, and in the writing of the manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2022.04.003.

References

[1] Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 2012;491(7426):705–10.
[2] Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 2018;361(6403).
[3] Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, et al. Multiple wheat genomes reveal global variation in modern breeding. Nature 2020;588 (7837):277–83.
[4] Pont C, Leroy T, Seidel M, Tondelli A, Duchemin W, Armisen D, et al. Tracing the ancestry of modern bread wheats. Nat Genet 2019;51(5):905–11.
[5] Cheng H, Liu J, Wen J, Nie X, Xu L, Chen N, et al. Frequent intra-and inter-species introgression shapes the landscape of genetic variation in bread wheat. Genome Biol 2019;20(1). doi: https://doi.org/10.1186/s13059-019-1744-x.
[6] He F, Pasam R, Shi F, Kant S, Keeble-Gagnere G, Kay P, et al. Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet 2019;51(5):896–904.
[7] Zhou Y, Zhao X, Li Y, Xu J, Bi A, Kang L, et al. *Triticum* population sequencing provides insights into wheat adaptation. Nat Genet 2020;52(12):1412–22.
[8] Chen H, Jiao C, Wang Y, Wang Y, Tian C, Yu H, et al. Comparative population genomics of bread wheat (*Triticum aestivum*) reveals its cultivation and breeding history in China. BioRxiv 2019;519587.
[9] Juliana P, Poland J, Huerta-Espino J, Shrestha S, Crossa J, Crespo-Herrera L, et al. Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. Nat Genet 2019;51(10):1530–9.
[10] Sansaloni C, Franco J, Santos B, Percival-Alwyn L, Singh S, Petroli C, et al. Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. Nat Commun 2020;11(1). doi: https://doi.org/10.1038/s41467-020-18404-w.
[11] Matsuoka Y. Evolution of polyploid *Triticum* wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification. Plant Cell Physiol 2011;52(5):750–64.
[12] Peng JH, Sun D, Nevo E. Domestication evolution, genetics and genomics in wheat. Mol Breed 2011;28(3):281–301.
[13] Ukleja M, Cuellar J, Siwaszek A, Kasprzak JM, Czarnocki-Cieciura M, Bujnicki JM, et al. The architecture of the *Schizosaccharomyces pombe* CCR4-NOT complex. Nat Commun 2016;7(1). doi: https://doi.org/10.1038/ncomms10433.
[14] Raisch T, Chang C-T, Levdansky Y, Muthukumar S, Raunser S, Valkov E. Reconstitution of recombinant human CCR4-NOT reveals molecular insights into regulated deadenylation. Nat Commun 2019;10(1):1–14.
[15] Bhaskar V, Basquin J, Conti E. Architecture of the ubiquitylation module of the yeast Ccr4-Not complex. Structure 2015;23(5):921–8.
[16] Molin L, Puisieux A. *elegans* homologue of the *Caf1* gene, which encodes a subunit of the CCR4-NOT complex, is essential for embryonic and larval development and for meiotic progression. Gene 2005;358:73–81.
[17] Gou L-T, Dai P, Yang J-H, Xue Y, Hu Y-P, Zhou Yu, et al. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Res 2014;24(6):680–700.
[18] Berthet C, Morera A-M, Asensio M-J, Chauvin M-A, Morel A-P, Dijoud F, et al. CCR4-associated factor CAF1 is an essential factor for spermatogenesis. Mol Cell Biol 2004;24(13):5808–20.
[19] Yamaji M, Jishage M, Meyer C, Suryawanshi H, Der E, Yamaji M, et al. DND1 maintains germline stem cells via recruitment of the CCR4–NOT complex to target mRNAs. Nature 2017;543(7646):568–72.
[20] Chou W-L, Chung Y-L, Fang J-C, Lu C-A. Novel interaction between CCR4 and CAF1 in rice CCR4–NOT deadenylase complex. Plant Mol Biol 2017;93(1-2):79–96.
[21] Chou W-L, Huang L-F, Fang J-C, Yeh C-H, Hong C-Y, Wu S-J, et al. Divergence of the expression and subcellular localization of CCR4-associated factor 1 (CAF1) deadenylase proteins in Oryza sativa. Plant Mol Biol 2014;85(4-5):443–58.
[22] Shimo HM, Terassi C, Lima Silva CC, Zanella JdL, Mercaldi GF, Rocco SA, et al. Role of the *Citrus sinensis* RNA deadenylase CsCAF1 in citrus canker resistance. Mol Plant Pathol 2019;20(8):1105–18.
[23] Wang Pu, Li L, Wei H, Sun W, Zhou P, Zhu S, et al. Genome-wide and comprehensive analysis of the multiple stress-related CAF1 (CCR4-Associated Factor 1) family and its expression in Poplar. Plants 2021;10(5):981. doi: https://doi.org/10.3390/plants10050981.
[24] Fang J-C, Tsai Y-C, Chou W-L, Liu H-Y, Chang C-C, Wu S-J, et al. A CCR4-associated factor 1, OsCAF1B, confers tolerance of low-temperature stress to rice seedlings. Plant Mol Biol 2021;105(1):177–92.
[25] Walley JW, Kelley DR, Nestorova G, Hirschberg DL, Dehesh K. *Arabidopsis* deadenylases AtCAF1a and AtCAF1b play overlapping and distinct roles in mediating environmental stress responses. Plant Physiol 2010;152(2):866–75.
[26] Bernhardt N, Brassac J, Kilian B, Blattner FR. Dated tribe-wide whole chloroplast genome phylogeny indicates recurrent hybridizations within Triticeae. BMC Evol Biol 2017;17(1):1–16.
[27] Su Q, Liu L, Zhao M, Zhang C, Zhang D, Li Y, et al. The complete chloroplast genomes of seventeen *Aegilops tauschii*: Genome comparative analysis and phylogenetic inference. PeerJ 2020;8:e8678.
[28] Fan X, Sha L-N, Dong Z-Z, Zhang H-Q, Kang H-Y, Wang Yi, et al. Phylogenetic relationships and Y genome origin in *Elymus* L. sensu lato (Triticeae; Poaceae) based on single-copy nuclear *Acc1* and *Pgk1* gene sequences. Mol Phylogenet Evol 2013;69(3):919–28.
[29] Petersen G, Seberg O, Salomon B. The origin of the H, St, W, and Y genomes in allotetraploid species of *Elymus* L. and *Stenostachys* Turcz. (Poaceae: Triticeae) Plant Syst Evolut 2011;291(3-4):197–210.
[30] Guo W, Xin M, Wang Z, Yao Y, Hu Z, Song W, et al. Origin and adaptation to high altitude of Tibetan semi-wild wheat. Nat Commun 2020;11(1). doi: https://doi.org/10.1038/s41467-020-18738-5.
[31] Hao C, Jiao C, Hou J, Li T, Liu H, Wang Y, et al. Resequencing of 145 landmark cultivars reveals asymmetric sub-genome selection and strong founder genotype effects on wheat breeding in China. Molecular Plant 2020;13 (12):1733–51.
[32] Chen C, Chen H, Zhang Yi, Thomas HR, Frank MH, He Y, et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Molecular plant 2020;13(8):1194–202.
[33] Sandve SR, Rudi H, Asp T, Rognli OA. Tracking the evolution of a cold stress associated gene family in cold tolerant grasses. BMC Evol Biol 2008;8:245.
[34] Ma S, Wang M, Wu J, Guo W, Chen Y, Li G, et al. WheatOmics: A platform combining multiple omics data to accelerate functional genomics studies in wheat. Molecular Plant 2021;14(12):1965–8.
[35] Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 2016;33(7):1870–4.
[36] Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 2009;25(11):1451–2.
[37] Excoffier L, Lischer H. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 2010;10(3):564–7.
[38] Leigh JW, Bryant D. Popart: full-feature software for haplotype network construction. Methods Ecol Evol 2015;6(9):1110–6.
[39] Hart KJ, Oberstaller J, Walker MP, Minns AM, Kennedy MF, Padykula I, et al. Plasmodium male gametocyte development and transmission are critically regulated by the two putative deadenylases of the CAF1/CCR4/NOT complex. PLoS Pathog 2019;15(1):e1007164.
[40] Koch P, Löhr HB, Driever W, Müller F. A mutation in cnot8, component of the Ccr4-not complex regulating transcript stability, affects expression levels of developmental regulators and reveals a role of *Fgf3* in development of caudal hypothalamic dopaminergic neurons. PLoS ONE 2014;9(12):e113829.
[41] Sun L, Sun G, Shi C, Sun D. Transcriptome analysis reveals new microRNAs-mediated pathway involved in anther development in male sterile wheat. BMC Genomics 2018;19(1):1–17.
[42] Tang TTL, Stowell JAW, Hill CH, Passmore LA. The intrinsic structure of poly (A) RNA determines the specificity of Pan2 and Caf1 deadenylases. Nat Struct Mol Biol 2019;26(6):433–42.
[43] Ma P-F, Liu Y-L, Jin G-H, Liu J-X, Wu H, He J, et al. The *Pharus latifolius* genome bridges the gap of early grass evolution. Plant Cell 2021;33(4):846–64.
[44] Ortiz-Ramírez C, Hernandez-Coronado M, Thamm A, Catarino B, Wang M, Dolan L, et al. A transcriptome atlas of *Physcomitrella patens* provides insights into the evolution and development of land plants. Molecular Plant 2016;9 (2):205–20.
[45] Guo H, Lee T-H, Wang X, Paterson AH. Function relaxation followed by diversifying selection after whole-genome duplication in flowering plants. Plant Physiol 2013;162(2):769–78.
[46] Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science 2015;348(6237):921–5.
[47] Messing J, Bennetzen JL. Grass genome structure and evolution. Plant Genomes 2008;4:41–56.
[48] Schrider DR, Hahn MW. Gene copy-number polymorphism in nature. Proc Roy Soc B: Biol Sci 2010;277(1698):3213–21.
[49] Huynh S, Marcussen T, Felber F, Parisod C. Hybridization preceded radiation in diploid wheats. Mol Phylogenet Evol 2019;139:106554. doi: https://doi.org/10.1016/j.ympev.2019.106554.
[50] Dvorak J, Deal KR, Luo M-C, You FM, von Borstel K, Dehghani H. The origin of spelt and free-threshing hexaploid wheat. J Hered 2012;103(3):426–41.
[51] Chang A- Y-F, Liao B-Y. DNA methylation rebalances gene dosage after mammalian gene duplications. Mol Biol Evol 2012;29(1):133–44.
[52] Wicker T, Mayer KFX, Gundlach H, Martis M, Steuernagel B, Scholz U, et al. Frequent gene movement and pseudogene evolution is common to the large

and complex genomes of wheat, barley, and their relatives. Plant Cell 2011;23 (5):1706–18.
[53] Feldman M, Kislev ME. Domestication of emmer wheat and evolution of free-threshing tetraploid wheat. Israel J Plant Sci 2007;55(3):207–21.
[54] Zohary D, Hopf M, Weiss E. Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin: Oxford University Press on Demand; 2012.
[55] Zohary D. Monophyletic vs. polyphyletic origin of the crops on which agriculture was founded in the Near East. Genet Resour Crop Evol 1999;46 (2):133–42.
[56] Civáň P, Ivaničová Z, Brown TA, Niedz R. Reticulated origin of domesticated emmer wheat supports a dynamic model for the emergence of agriculture in the fertile crescent. PLoS ONE 2013;8(11):e81955.
[57] Matsuoka Y, Mori N. Reproductive and genetic roles of the maternal progenitor in the origin of common wheat (Triticum aestivum L.). Ecol Evol 2020;10 (24):13926–37.
[58] Ouyang Y, Li G, Mi J, Xu C, Du H, Zhang C, et al. Origination and establishment of a trigenic reproductive isolation system in rice. Mol Plant 2016;9 (11):1542–5.
[59] Nevo E. Evolution of wild emmer wheat and crop improvement. J Syst Evolut 2014;52(6):673–96.
[60] Feldman M, Fernández-Domínguez E, Reynolds L, Baird D, Pearson J, Hershkovitz I, et al. Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia. Nat Commun 2019;10(1). doi: https://doi.org/10.1038/s41467-019-09209-7.
[61] Kang HY, Fan X, Zhang HQ, Sha LN, Sun GL, Zhou YH. The origin of *Triticum petropavlovskyi* Udacz. et Migusch.: demonstration of the utility of the genes encoding plastid acetyl-CoA carboxylase sequence. Mol Breed 2010;25 (3):381–95.
[62] Chen Q, Kang H-Y, Fan X, Wang Y, Sha L-N, Zhang H-Q, et al. Evolutionary history of *Triticum petropavlovskyi* Udacz. et Migusch. inferred from the sequences of the 3-phosphoglycerate kinase gene. PloS One 2013;8(8):e71139.
[63] Akond AS, Watanabe N, Furuta Y. Comparative genetic diversity of *Triticum aestivum-Triticum polonicum* introgression lines with long glume and *Triticum petropavlovskyi* by AFLP-based assessment. Genet Resour Crop Evol 2008;55 (1):133–41.