ARTICLES
https://doi.org/10.1038/s41477-017-0067-8

nature plants

OPEN

The *Aegilops tauschii* genome reveals multiple impacts of transposons

Guangyao Zhao[1], Cheng Zou[1], Kui Li[2], Kai Wang[2], Tianbao Li[3], Lifeng Gao[1], Xiaoxia Zhang[4], Hongjin Wang[5], Zujun Yang[5], Xu Liu[1], Wenkai Jiang[2]*, Long Mao[1]*, Xiuying Kong[1]*, Yuannian Jiao[4]* and Jizeng Jia[1]*

Wheat is an important global crop with an extremely large and complex genome that contains more transposable elements (TEs) than any other known crop species. Here, we generated a chromosome-scale, high-quality reference genome of *Aegilops tauschii*, the donor of the wheat D genome, in which 92.5% sequences have been anchored to chromosomes. Using this assembly, we accurately characterized genic loci, gene expression, pseudogenes, methylation, recombination ratios, microRNAs and especially TEs on chromosomes. In addition to the discovery of a wave of very recent gene duplications, we detected that TEs occurred in about half of the genes, and found that such genes are expressed at lower levels than those without TEs, presumably because of their elevated methylation levels. We mapped all wheat molecular markers and constructed a high-resolution integrated genetic map corresponding to genome sequences, thereby placing previously detected agronomically important genes/quantitative trait loci (QTLs) on the *Ae. tauschii* genome for the first time.

Wheat is one of the most important food crops in the world. It is also a species with an extremely large and complex genome that contains more TEs than any other known species[1]. This has for some time caused wheat research to lag behind that of crops with smaller genomes like rice, sorghum, etc. TEs were discovered by Barbara McClintock in 1951[2], but for about three decades dating from her discovery, TEs were largely thought to be 'Junk DNA', parasite genes in a host genome, or selfish DNA[3]. TEs make up a large fraction of many plant genomes; indeed, they are the major determinant of genome size. For example, more than 80% of the genomes of maize[4] and wheat[1] are composed of TEs. More recently, TEs have been recognized as important functional components of genomes. Four distinct functional contributions of TEs are now recognized, including their roles in determining genome size and rearrangements, in generating mutations, in altering chromosome architecture and in the regulation of gene expression. TEs are now recognized as an abundant and unexplored natural source of regulatory sequences for host genes, and TE biology has become an active research area in recent years[5–10].

Sequencing technologies such as Illumina HiSeq X Ten and PacBio RS II sequencing, and new library construction methods such as 10x Genomics, as well as new assembly techniques such as DeNovoMAGIC2[11,12], are now significantly accelerating the progress of wheat genome sequencing efforts. Common wheat is a hexaploid species with A, B and D subgenomes. Whole genome shotgun sequencing of the Chinese Spring cultivar[13,14] and its diploid ancestors *Triticum urartu*[15] and *Ae. tauschii*[16] have been reported. Additionally, chromosome sorting was used to sequence chromosome 3B of the hexaploid wheat cultivar Chinese Spring, which advanced wheat research significantly[1]. Although previous efforts generated draft genomes for diploid ancestors and for bread wheat, the majority of contigs could not be mapped to chromosomes in these assemblies. We previously constructed the draft genome of *Ae. tauschii*, and found that it was particularly rich in genes related to adaptation[16]. Here, we present a reference D genome in which more than 92.5% of the genome sequences are anchored to chromosomes. We mapped the global distribution of TEs and examined multiple functional impacts of TEs on D genome evolution, gene structure and gene expression.

Results and discussion

Genome assembly and feature annotation. Over 778 Gb of short read sequences were generated using two Illumina sequencing platforms (HiSeq 2000 and HiSeq 2500) (Supplementary Table 1). Short reads were first assembled by DeNovoMAGIC2[11,12] using 450 bp, 2 kb, 5 kb and 8 kb libraries to generate the V1.0 assembly, which comprises 271,060 contigs (N50 = 50.3 kb) and 117,344 scaffolds (N50 = 6.8 Mb). The assembly was further elongated using SSPACE with 20 kb and 40 kb libraries to generate the V1.1 assembly (N50 = 13.1 Mb). Finally, we used long reads generated with the PacBio RS II sequencing platform to further improve the contig assembly (Supplementary Table 2). The final assembly (V1.2) comprises 188,412 contigs (N50 = 112.6 kb) and 112,517 scaffolds (N50 = 12.1 Mb) (Supplementary Table 3). Our assembly represents a greater than 210-fold improvement in contiguity compared with the previously published *Ae. tauschii* assembly reported by our research group in 2013[16] (see statistics for V0.1 in Supplementary Table 2). The total scaffold length of the V1.2 assembly (4.31 Gb)

[1]Key Laboratory of Crop Gene Resources and Germplasm Enhancement, Ministry of Agriculture, The National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, 100081 Beijing, China. [2]Novogene Bioinformatics Institute, 100083 Beijing, China. [3]Agronomy College, Collaborative Innovation Center of Henan Grain Crops, Henan Agricultural University, 450002 Zhengzhou, China. [4]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, the Chinese Academy of Sciences, 100093 Beijing, China. [5]Center for Information in Biology, School of Life Science and Technology, University of Electronic Science and Technology of China, 610054 Chengdu, China. Guangyao Zhao, Cheng Zou and Kui Li contributed equally to this work. *e-mail: jiangwenkai@novogene.com; maolong@caas.cn; kongxiuying@caas.cn; jiaoyn@ibcas.ac.cn; jiajizeng@caas.cn

spans 95.8% of the estimated genome size (4.5 Gb)[16], and the largest 434 scaffolds cover 90% of the genome (Supplementary Table 3).

A high-density genetic map containing 164,872 single-nucleotide polymorphism (SNP) loci spanning 1,153.6 cM in seven linkage groups was used to anchor the scaffolds to chromosomes. In total, 658 scaffolds were aligned and anchored to the genetic map, with a total length of 4.0 Gb, spanning 92.5% of the assembled genome; 97.9% of the aforementioned SNPs were placed on the scaffolds (Supplementary Table 4). We confirmed that the sizes of the chromosomes in our assembly were consistent with the results of a cytogenetics analysis that we performed for this study: chromosomes 2, 3, 5 and 7 are relatively longer, but chromosomes 1, 4 and 6 are shorter (Table 1 and Supplementary Fig. 2). Attesting to the quality of the assembly, Illumina paired end reads were mapped to our assembly and 99.96% of them could be mapped, which suggests that our assembly contains almost all of the information in the raw reads (Supplementary Table 6). Completeness of gene regions was assessed using CEGMA (conserved core eukaryotic gene mapping approach) and BUSCO (Benchmarking Universal Single Copy Orthologs). Two hundred and forty-three of the 248 (97.9%) conserved core eukaryotic genes from CEGMA were captured in our assembly), and 240 (98.8%) of these were complete. BUSCO analysis showed that 97% of the plant single-copy orthologues were complete. We also mapped 10,748 EST sequences generated from our *Ae. tauschii* full-length cDNA libraries[16] to the assembly; 10,471 (97.4%) of these could be mapped to the scaffolds with greater than 90% coverage, which indicated that gene regions were almost complete in our assembly. To assess large-scale accuracy, we compared our assembly with the sequences of 17 BACs obtained either from the NCBI GenBank or from our in-house library. All of these BACs could be aligned to our assembly, in the correct order, with high sequence identity (15 of 17 alignments had identity > 99%), and had greater than 97% coverage. Only two of the gaps located in genic regions, all other gaps on the BACs were intergenic sequence, repeats or N strings (Supplementary Fig. 1 and Supplementary Table 10). The accuracy of the assembly was further confirmed by contig sequences which were randomly selected from a genome version generated from PacBio data[17]. All of the 120 selected PacBio Contigs could be unambiguously aligned to our assembly with sequence identity greater than 99.3% (Supplementary Table 11). These dramatic improvements in quality result from the use of emerging techniques like DeNovoMAGIC2, large insert mate-pair sequencing and PacBio, especially for TE-rich genomes. We annotated protein-coding genes (PCGs), RNA, pseudogenes and catalogued functional annotations for PCGs. Importantly, we also analysed the distribution of TEs on the seven chromosomes (Fig. 1). The *Ae. tauschii* genome sequence contains 42,828 PCGs, 98% of which were anchored to chromosomes in our assembly (Table 1); this is more than the 39,425 PCGs reported for Chinese Spring[14]. The average lengths, the number of exons and the GC content of the coding regions of the PCGs are similar to those of the genes in the genomes of other grass species (Supplementary Table 19). However, both the intron lengths and the intron GC content of the *Ae. tauschii* PCGs are much larger than those in any other sequenced grass species (Supplementary Table 19). These differences are perhaps due to the dramatically higher number of TE insertions in the introns of the *Ae. tauschii* PCGs. Of the PCGs, 92.6% were functionally annotated based on information from the NR (non-redundant database in NCBI), SwissProt, InterPro, Pfam and KEGG databases (Supplementary Table 20). We identified 25,893 likely pseudogenes with premature stop codons or frameshift mutations, a much higher number than that in rice (1,439–5,608)[18,19] or *Arabidopsis* (801–4,108)[18,20]. When including gene fragments without disabling mutations in a broader pseudogene definition, the total number of pseudogenes reached 267,546 in the *Ae. tauschii* genome, which is two times larger than the number in the maize genome (B73 RefGen_v3, in the 5b + annotation build)[21]. Therefore, *Ae. tauschii* appears to be the plant species with the highest reported pseudogene content. In addition, 3,630 transfer RNA genes, 238 miRNA genes, 1,271 small nuclear RNA genes and 2,856 ribosomal RNA genes were predicted in the genome (Supplementary Table 22).

Three whole genome duplication events have been identified in the evolutionary history of all of the sequenced grass genomes, including the *tau* event at ~150 million years ago (Ma), the *sigma* event at ~127 Ma and the *rho* event at ~70 Ma[4,22]. To infer the evolutionary history of *Ae. tauschii*, we used MCScanx[23] to detect syntenic genomic regions among *Ae. tauschii*, *Oryza sativa*, *Brachypodium distachyon* and *Sorghum bicolor*; all of these intergenomic comparisons showed very strong co-linearity (Supplementary Figs. 12–14), further indicating the high-quality of our *Ae. tauschii* assembly and PCG annotation. An intragenomic comparison of *Ae. tauschii* showed a relatively smaller number of syntenic regions than that in other sequenced grass genomes (Fig. 2a and Supplementary Fig. 15), with the largest syntenic region occurring between chromosomes 1 and 3. Interestingly, this region is also present in the respective syntenic regions of the genomes of five other sequenced grass genomes (Fig. 2b). Genome alignments between *Ae. tauschii* and both *O. sativa* and *S. bicolor* revealed a clear 2-to-2 multiplicity ratio between orthologous regions, which supports the idea that the pan-cereal *rho* event[24] was the most recent whole genome duplication event in the evolutionary history of *Ae. tauschii* (Fig. 2b).

Transposable element analysis. TEs account for fully 85.9% of the assembled sequence of the *Ae. tauschii* genome, similar to what

Table 1 | The gene and TE distribution on the seven chromosomes of the *Ae. tauschii* genome

Chr	Length (Mb)	Genetic length (cM)	Centromere region length (Mb)	Gene counts	Gene density (No. gene/Mb)	Pseudogene counts	TE length (Mb)	LTR length (Mb)	Gypsy length (Mb)	Copia length (Mb)	TE-DNA length (Mb)
Chr1	508.57	144.98	76	5,356	10.5	3,097	441.37	300.21	198.19	85.87	97.02
Chr2	632.94	182.96	136	6,939	11.0	3,856	548.60	371.10	246.78	103.40	122.73
Chr3	627.29	169.00	82	6,129	9.8	3,867	545.32	373.41	250.44	103.68	119.30
Chr4	530.31	120.14	106	4,807	9.1	3,140	471.32	331.69	231.23	83.70	96.29
Chr5	570.43	214.49	82	6,645	11.6	3,375	491.64	341.21	219.62	104.54	100.35
Chr6	497.16	135.95	125	5,179	10.4	3,092	433.32	297.43	198.33	83.45	93.48
Chr7	639.09	186.06	171	6,900	10.8	4,108	548.69	364.43	241.72	103.09	128.35
Chr0	328.57	NA	NA	873	2.7	1,358	217.28	113.47	62.37	43.78	83.86
Total	4,334.36	1,153.58	778	42,828	9.9	25,893	3,697.54	2,492.95	1,648.68	711.51	841.38

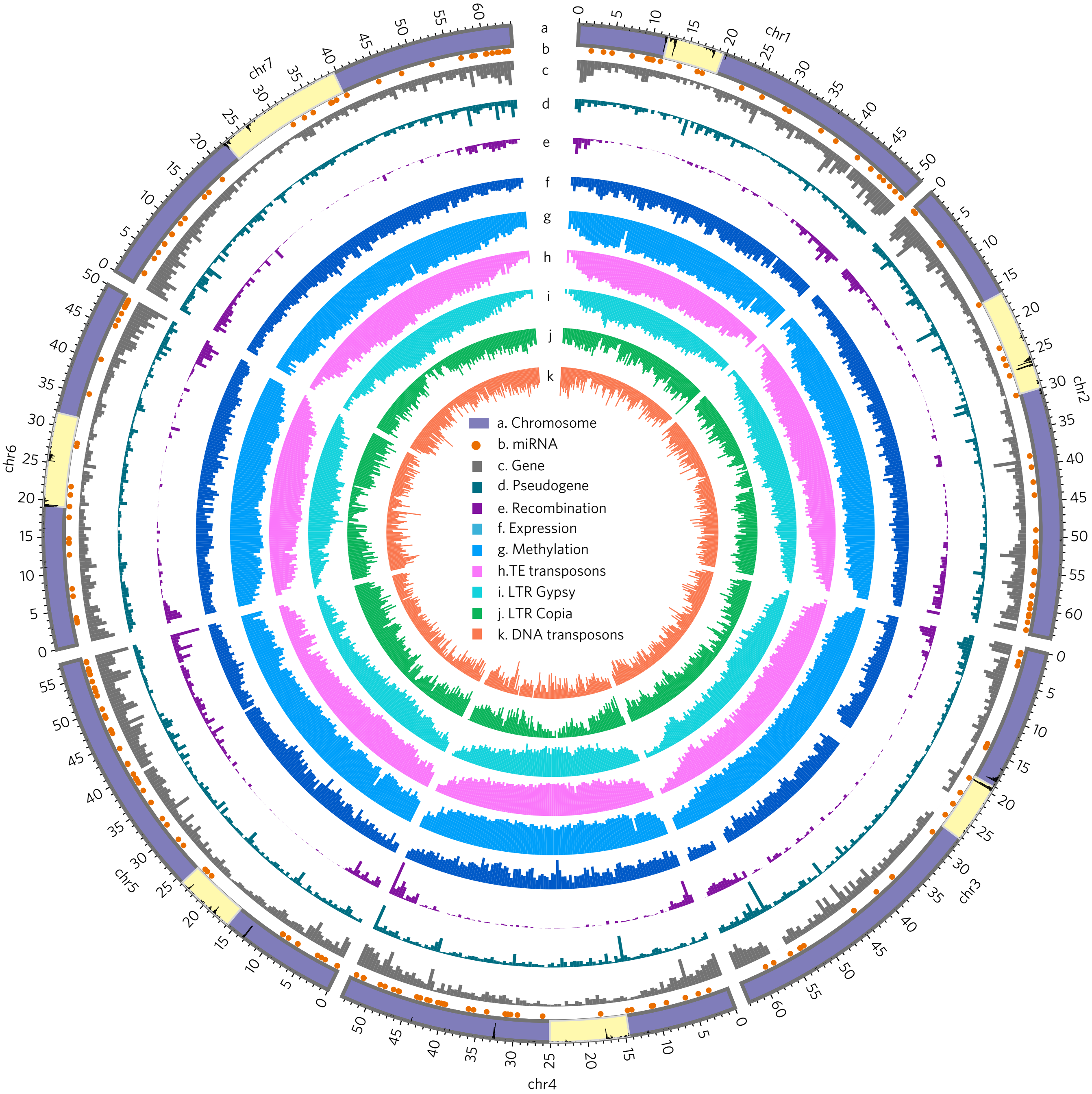

Fig. 1 | Distribution of genomic features in the *Ae. tauschii* genome. Track a, the seven chromosomes. One scale label indicates 10 Mb. The black histogram indicates the distribution of two LTR-RTs (Quinta and Cereba); peaks indicate candidate centromere regions. Track b, miRNA. Track c, Gene density indicated by gene length/5 Mb. Track d, pseudogene density, indicated by gene length/5 Mb. Track e, recombination, measured by cM/Mb with 5 Mb bins. Track f, gene expression, calculated as the average reads per kilobase per million mapped reads (rpkm) in 5 Mb bins. Maximum rpkm from eight tissues and normalization of rpkm by $\log_{10}(\text{rpkm} + 10^{-6})$. Track g, methylation, calculated by methylation base counts/C base counts in Mb bins. Only the most abundant CG methylation contexts are presented. Track h, total TE density across chromosomes, TE length/5 Mb. Track i, Gypsy. The distribution of Gypsy TE density across chromosomes: Gypsy length/5 Mb. Track j, Copia. The distribution of Copia density across chromosomes: Copia length/5 Mb. Track k, DNA TEs. The distribution of DNA transposon density across chromosomes: DNA transposon length/5 Mb.

occurs in bread wheat (chromosome 3B)[1] and maize[4], but much higher than the TE content in rice[25], sorghum[26] or *B. distachyon*[27] (Supplementary Table 13). Note that 85.9% TE content is much higher than what we previously estimated based on our *Ae. tauschii* V0.1 draft genome (62.3%)[16], and also higher than in Chinese Spring (76.6%)[14]. This can be partly attributed to the greater than 210-fold improvement in contiguity of our new assembly. Retrotransposons and DNA transposons cover, respectively, 59.9% and 19.6% of the genome (Supplementary Table 14). Long terminal repeats (LTRs) are the most abundant type of TE, covering 58% of the genome. Three superfamilies (Gypsy, 38.3%; CACTA, 16.8%; Copia, 16.5%) account for 71.7% of the total TE component (Supplementary Table 14). Compared to both rice and maize, the proportion of CACTAs in *Ae. tauschii* is almost tenfold greater[4,25].

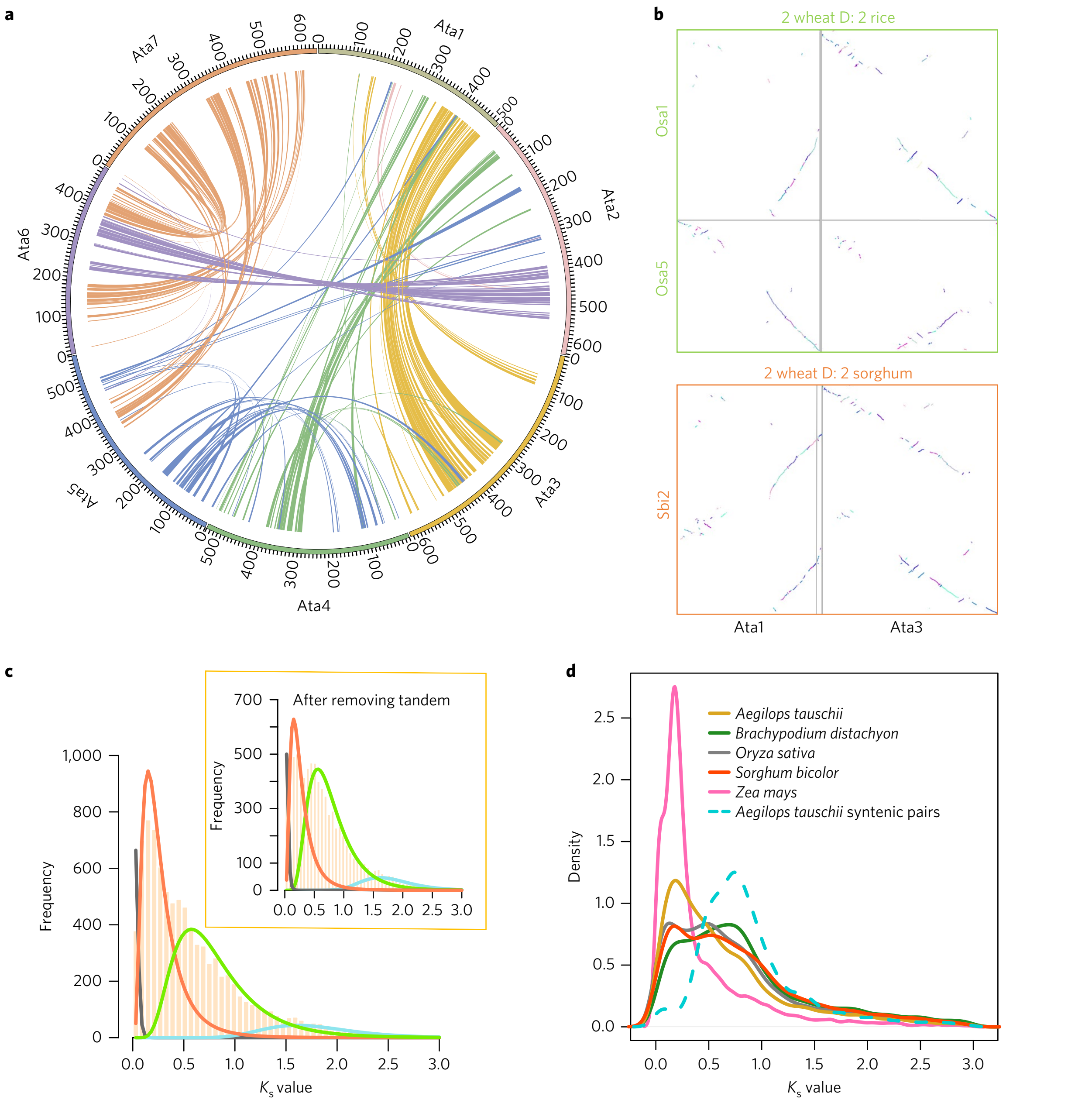

Fig. 2 | Aegilops genome evolution. a, Circos plot showing the syntenic regions in the *Ae. tauschii* genome detected by MCScanx. **b**, Synteny comparisons of *Ae. tauschii* chromosomes 1 and 3 to the genomes of rice and sorghum. Quota ratios of the alignments are shown above the dot plots. **c**, K_s value distributions of 9,569 *Aegilops* paralogues identified from best reciprocal BLASTP searches. Coloured lines superimposed on K_s plots represent significant components identified via a likelihood mixture model implemented in EMMIX. Each component is written as 'colour/mean K_s value/proportion', where colour is the component (curve) colour and proportion is the percentage of duplicated genes assigned to the identified component. There are four components: grey/0.03/0.04, red/0.23/0.44, green/0.72/0.45 and blue/1.71/0.07. The smaller plot represents the same distribution after removal of the tandem duplicates. The four components are grey/0.04/0.04, red/0.23/0.33, green/0.70/0.55 and blue/1.71/0.08. **d**, Density plot of K_s values of best reciprocal BLASTP hits in the genomes of *Ae. tauschii, B. distachyon, O. sativa*, *S. bicolor* and *Z. mays*. Density distribution of K_s values of *Ae. tauschii* syntenic anchor gene pairs.

To further explore the proliferation in each TE superfamily, we here define 'family' according to the peptide sequence similarity of key transposase domains (threshold ≥ 90%; retrotransposase domain (RT) for retrotransposons and the catalytic domains of DD [E/D] (DDE) DNA transposons). In total, we identified 4,669 TE families (100,879 copies with complete transposase domains) among eight superfamilies/orders in our assembly (Supplementary Table 15). Consistent with the DNA-sequence-based TE annotation (above), our domain-based analysis showed that LTR retrotransposons were the most abundant type of TE, and CACTA elements were the most abundant type among the DNA transposons.

We identified 172 CACTA families that formed four clades in a phylogenetic tree (Fig. 3a). The four clades, Aet-CACTA-1 to Aet-CACTA-4, consisted of 66, 49, 38 and 19 families. To investigate their origins, we identified the CACTA families of five published grass genomes and combined them with the *Ae. tauschii* CACTAs to obtain a six-species CACTA data set containing 855 families (Supplementary Table 17); the genomes included were common

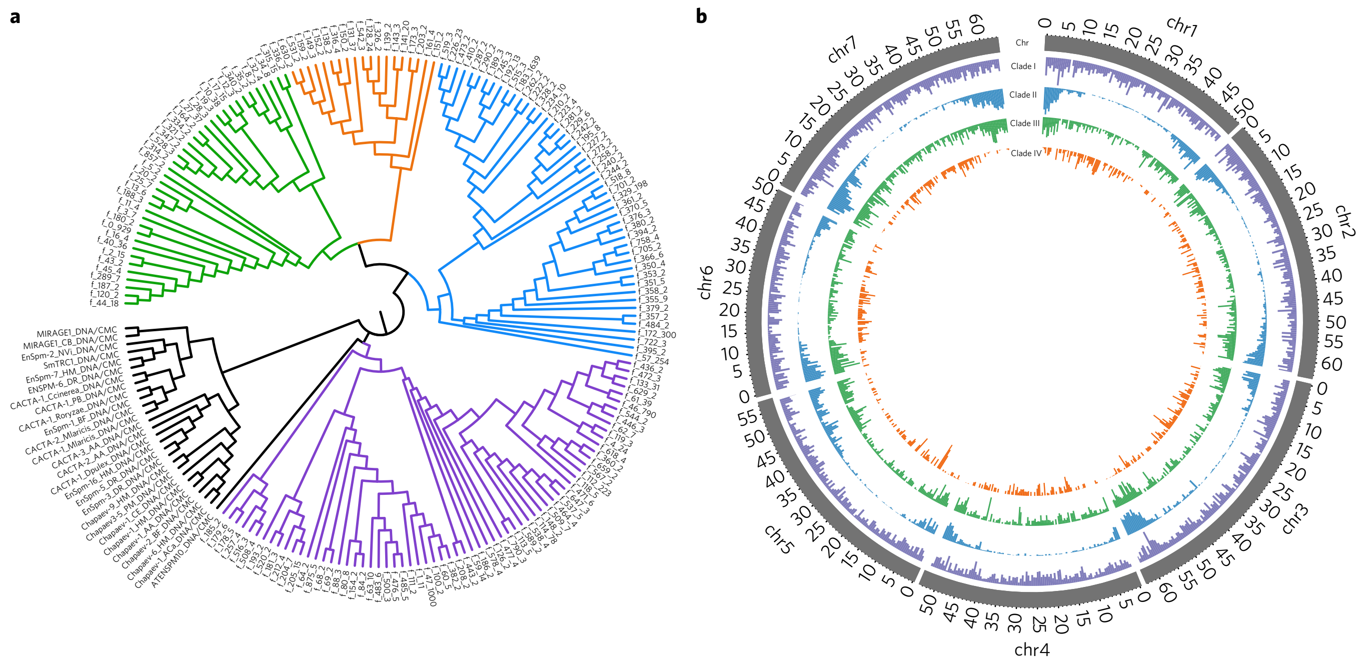

Fig. 3 | Rapid proliferation of transposable elements. **a**, A neighbour-joining phylogeny of CACTA families. Four major clades with high bootstrap support are marked in different colours. Format of sequence ID: 'f number 1–number 2', where number 1 is the family ID and number 2 is the copy number of that family. For example, "f183 1639" indicates that family 183 has 1,639 copies. The black clade is the outgroup. The branch length is not proportional to the genetic distance. **b**, Chromosome-wide distribution of the high-copy-number families (copy-number ≥ 100) among the four major clades of CACTAs. The colour scheme is the same in **a** and **b.**

wheat[13,14], *Triticum urartu*[15], *B. distachyon*[27], rice[25] and sorghum[26]. The topology of the neighbour-joining phylogenetic tree distinguished four clades of CACTAs in *Ae. tauschii*, and clearly suggested that the Aet-CACTA-4 clade originated in the Triticeae lineage, whereas the other three clades originated before the divergence of the major grass groups. A total of 19 families are included in Aet-CACTA-4, relatively fewer than in the three other clades. The centromere-enriched distribution of the Aet-CACTA-4 TEs in *Ae. tauschii* is distinct from the distributions of the other known CACTA subfamilies, which are more abundant in terminal, gene-rich regions (Fig. 3b).

Analysis of the evolutionary relationships of transposase domains showed that the high-copy-number families accounted for only a small proportion of all families, but 59.1% (2,829/4,782) of families had a copy number of 2 or 3 (we set the lowest copy number as 2 in this study) (Supplementary Table 15 and Supplementary Fig. 9). For example, only eight out of 172 CACTA families were high-copy-number families (copy number > 100), but these eight families (5% of all families) had 5,333 copies, representing 84% of all identified CACTA copies. A similar pattern was observed for LTR elements, where 99 (2.5%) of the high-copy-number families accounted for 65,912 (72.6%) of all identified LTR copies. Since CACTA and LTR elements account for over 70% of all *Ae. tauschii* TEs, this pattern suggests that a few dominant families might play key roles in recent genome structural evolution, supporting previous suppositions from a study of bread wheat[1,28].

We next analysed the global distribution of TEs on chromosomes and found that they tended to increase in density along the chromosome from distal regions towards the centromere (Fig. 1 and Supplementary Fig. 19). The distribution of Gypsy TEs is similar to that of total TEs, as indicated by a significant positive correlation ($r = 0.92$, $P < 10^{-13}$). Given that Gypsy TEs represent fully 38.3% (Supplementary Table 14) of total TEs, this high correlation is not surprising. However, the density of both Copia and CACTA TEs decreased from distal regions to centromeric/pericentromeric regions, showing a significant negative correlation with the distribution of total TEs and of Gypsy TEs ($P < 10^{-13}$). To explore the impact of TEs on the distribution of genomic features, we plotted the distributions of both gene and pseudogene density, and also examined gene expression, DNA methylation and recombination on every chromosome (Fig. 1). We found that TEs are associated significantly with all of the other genomic features, highlighting the consequential impact of TEs on this genome. The density of Gypsy elements is significantly positively correlated with average gene expression levels, but is negatively correlated with both gene density and recombination. Similarly, negative correlations were observed for Copia and DNA TEs, although these trends were less obvious. All of these trends are consistent with the findings from an analysis of chromosome 3B of bread wheat[1].

Impact of TEs on gene evolution and regulation. Considering that 85.9% of the *Ae. tauschii* genome is composed of TEs, we explored the impact of TEs on gene duplication, gene structure, methylation, expression and pseudogenization. We found a large number of recently duplicated genes (not resulting from whole genome duplication event in the *Ae. tauschii* genome), some of which in theory may have resulted from TE movement. All-against-all best reciprocal BLASTP searches identified a total of 9,569 paralogous gene pairs. The synonymous nucleotide substitution rate (K_s) for the paralogues was calculated using codeml (PAML)[29]. Interestingly, the K_s distribution showed an older peak at ~0.65 and a more recent peak at ~0.25 (Fig. 2c), suggesting that, in addition to the pan-cereal *rho* event, a larger number of dispersed genes were duplicated more recently in the *Ae. tauschii* genome (3,034 gene pairs with K_s values less than 0.3) than other sequenced grass genomes (Fig. 2d and Supplementary Fig. 16). These genes were classified into gene family categories that are significantly enriched for wounding responses, endopeptidase inhibitor activity, serine-type endopeptidase inhibitor activity and enzyme inhibitor activity (Supplementary Table 23). Careful classification of these recently duplicated gene pairs showed that 1,204 were specific to the wheat D genome (Supplementary Fig. 17). A separate positional analysis revealed that 1,102 of the recently duplicated gene pairs are likely to have arisen via tandem duplication, and the remaining 1,932 non-tandem duplicate pairs were uniformly distributed across all chromosomes (Supplementary Fig. 16). We investigated the sequence similarities between the 1.5 kb flanking sequences of the 1,932 non-tandem duplicates, and found that 23 pairs had similar 5′ and 3′ flanking sequences (defined as 80% identity and 50% of the length), 273 pairs had similar 5′ flanking sequence and 46 pairs had similar 3′ flanking sequences. We further investigated if the flanking regions of these recently duplicated genes overlapped with TEs, and found that the flanking sequences of 3,415 (88%) genes had TEs. Given the tremendous number of TEs in the *Ae. tauschii* genome, and considering that less than about 36% of these resulted from tandem duplication, it seems reasonable to speculate that a burst of TE-associated gene duplication may explain the unexpectedly large number of recently duplicated genes that we observed in our K_s analysis.

Compared to other grass species, including maize, rice and *B. distachyon*, the occurrence of both retrotransposons and DNA transposons in both gene bodies and in flanking regions is highest in *Ae. tauschii* (Fig. 4a). Of the predicted genes of *Ae. tauschii*, 45.5% contain at least one TE (Supplementary Table 21), which is two times higher than in maize, another TE-rich genome[4]. Among the 12 TE superfamilies we examined, CACTAs are the most abundant superfamily of TEs inserted in introns; these are present in 8,547 genes (43.9 % of the genes with TE insertions, Supplemental Table 21). We compared the length of inserted TEs in the genic regions and intergenic regions, and found that the TEs inserted in genic regions are significantly shorter than those inserted in intergenic regions (two-sample Kolmogorov–Smirnov (KS) test, $P<0.001$) (Supplementary Fig. 11). As expected, genes containing TEs were on the whole expressed at lower levels than were genes without TEs (Supplementary Fig. 18, KS test, $P<10^{-13}$).

Pseudogenes can be regulators of biological function[30,31]. To test if the high pseudogene content in the *Ae. tauschii* genome detected above is correlated with historic bursts of TE movement in the *Ae. tauschii* genome, we identified processed pseudogenes (those resulting from retrotransposition) and examined the distribution of TEs across pseudogenes. Among the pseudogenes, about 29% of the multi-exon ancestor genes had lost their introns, suggesting that retrotransposition was involved in their pseudogenization. We also found that several superfamilies of retrotransposons and DNA TEs (for example, Gypsy, CACTAs, Helitrons, and so on, Fig. 4b) were enriched in pseudogene bodies and/or in the flanking regions of pseudogenes. More than 80% of the pseudogenes are somehow disrupted by TEs from these superfamilies, suggesting a pivotal role of TE movement in pseudogenization.

Cytosine DNA methylation is important in the epigenetic regulation of gene expression and in silencing transposons and other repetitive sequences[32]. To investigate DNA methylation and gene expression profiles in the *Ae. tauschii* genome, we conducted whole genome bisulfite sequencing and RNA-seq using the same sample tissues. The average percentages of methylation of CG, CHG and CHH contexts in leaf were 89.7%, 59.1% and 2.1%, respectively (Supplementary Table 24). The genome CG and CHG methylation levels are about the same as those in maize (86.4%, 70.9%)[33], and are much higher than those in *B. distachyon* (56.5%, 35.3%)[34] and rice (44%, 24%)[35]. This is consistent with previous observations indicating a positive correlation between genome size, TE content and genomic methylation levels[36]. The methylation level of CG and CHG contexts within genes is much lower than that in the intergenic regions, which might be attributable to the flanking sequences of genes in the *Ae. tauschii* genome being frequently occupied by TEs, and typically heavily methylated (Fig. 4c). Similar to *Arabidopsis*, rice and maize, CG methylation is the most abundant type of methylation in the *Ae. tauschii* genome, and the CG methylation levels are low near both transcriptional start sites (TSSs) and transcrip-

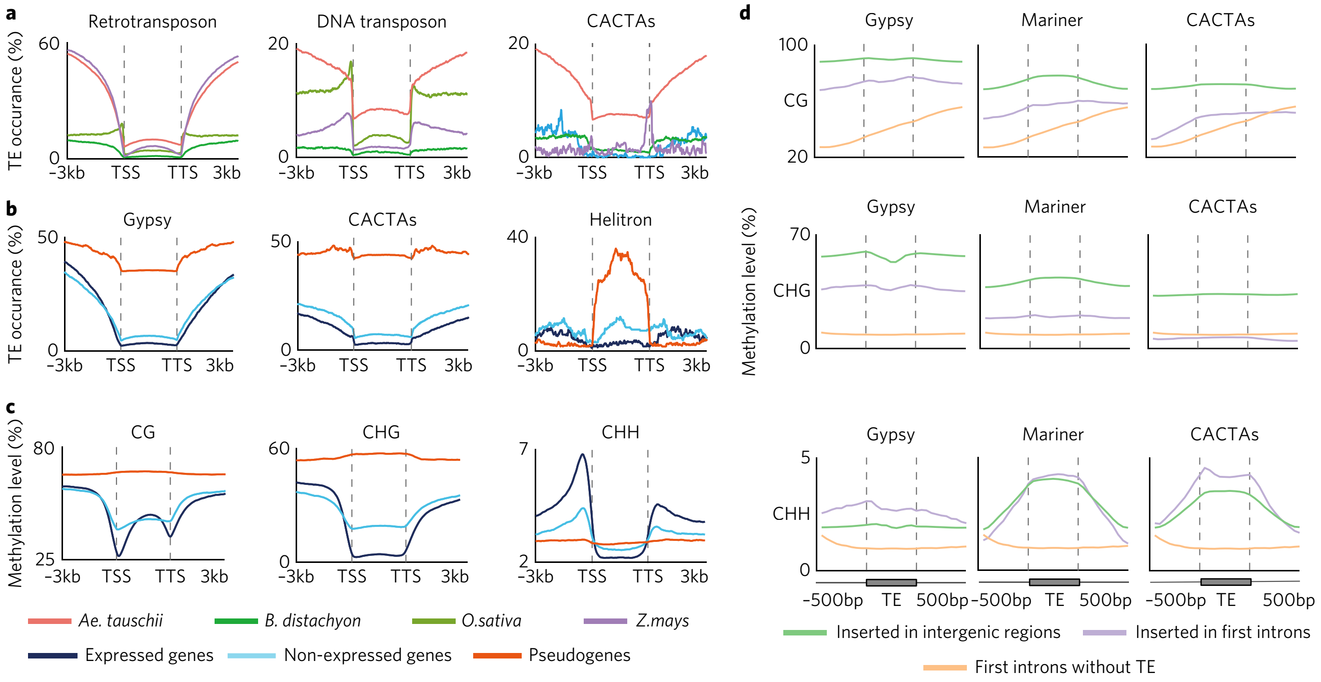

Fig. 4 | TE distribution and methylation profiles across genes in the *Ae. tauschii* genome. a, Distributions of retrotransposons, CACTAs and other DNA transposons across gene bodies and flanking regions (±3 kb). **b**, Distributions of Gypsy elements, CACTAs and Helitrons present in expressed genes, non-expressed genes and pseudogenes. **c**, Methylation levels of expressed genes, non-expressed genes and pseudogenes. **d**, Methylation levels of TEs and their surrounding regions. The analysis in **d** included only isolated TEs (no other TE or gene within 500 bp). The meta-profile of the methylation of the first intron of all genes without TEs represents the control in this plot. TSS, transcription start site; TTS, transcription terminal site.

tional terminal sites (TTSs). Intriguingly, this difference is more obvious in expressed genes than in non-expressed genes, which suggests that the lower level of CG methylation around TSSs and TESs might be important in regulating the activation of gene expression. In contrast, the level of CHG methylation is higher in non-expressed genes than in expressed genes, and this trend is consistent throughout gene bodies (in both exons and introns). Regions with elevated CHH methylation levels located immediately adjacent to TSSs and TESs are referred to as 'mCHH islands'. mCHH islands were first identified in maize, and have been proposed to function as 'insulators' that separate the silencing of TEs from that of nearby genes. However, there is as yet no clear conclusion about the relationship between mCHH islands and gene expression[37–39]. We found that mCHH islands are obvious in *Ae. tauschii* and, further, we found that the CHG methylation level is higher in non-expressed genes than in expressed genes (Fig. 4c), a finding that appears to provide an additional line of evidence to support this gene expression insulator theory. Our combined bisulfite sequencing and RNA-seq analysis revealed that genes with TEs inserted in their introns are expressed at lower levels than are genes without TEs, so we examined the methylation profile around TEs to test if TE insertions affect methylation of the surrounding region (Fig. 4d). Indeed, we found that introns with TE insertions showed higher methylation levels than introns without any TEs, although the level was lower than that for TEs in intergenic regions.

Integrated genetic map and key agronomic genes/QTLs map. Extensive previous research efforts have genetically mapped a large number of agronomically important genes/QTLs. However, most of these QTLs could not be physically mapped, and individual results are not typically comparable because studies used different mapping populations and different sets of markers. To address this deficiency, we mapped all available markers, including 735 first-generation restriction-fragment length polymorphisms (RFLPs)[40], 3,536 second-generation marker SSRs[41] and millions of third-generation SNPs (90 K[42], 820 K[43] and 660 K used in our laboratory) to the *Ae. tauschii* genome, and generated a high-resolution integrated genetic map corresponding to the genome sequences (Fig. 5). By using this integrated genetic map, we anchored 50 genes/QTLs (with marker sequences previously detected in various populations) to the D genome. We also identified 256 agronomically important genes that have been identified by map-based cloning in cereal crops (203 from rice, 22 from wheat, 16 from barley and 15 from maize). These genes include 53 conferring disease resistance, 19 for abiotic stress tolerance, nine for domestication, 135 for development, 12 for quality and 28 for yield (Supplementary Table 25). Thereby, we generated the first genome-based gene/QTL map for *Ae. tauschii* (Fig. 5 and Supplementary Table 26). A total of 33, 54, 38, 33, 37, 20 and 58 genes/QTLs were anchored to chromosomes 1D to 7D respectively, suggesting that 2D and 7D have made relatively stronger positive contributions to wheat improvement than the other chromosomes. All of these results highlight that, in addition to its utility in genomics research, a chromosome-scale reference genome is a valuable resource for molecular breeding and gene cloning.

Conclusions

We used highly efficient sequencing and assembly techniques to construct a high-quality reference genome for the TE-rich species *Ae. tauschii*. Our assembly has a scaffold N50 value of 13.1 Mb, and 92.5% of the scaffolds were anchored to chromosomes. We also developed a recombination map and anchored 97.9% of genes and 94.7% of pseudogenes and, impressively, 94.1% of TEs. We discovered a new, Triticeae-specific CACTA family, namely Aet-CACTA-4, which contains 19 subfamilies. The physical distribution of Aet-CACTA-4 is distinct from the other three known CACTA subfamilies. The *Ae. tauschii* genome has the largest number of pseudogenes among all examined genomes, which may be attributable to historical bursts of TE activity. About half of the *Ae. tauschii* genes have TE insertions, and bisulfite and transcriptome sequencing revealed that these genes had both elevated methylation levels and reduced transcription. We mapped all of the genetic markers from wheat studies spanning three decades, and constructed the first accurate integrated genetic–physical map of *Ae. tauschii*. We used these tools to map almost all of the previously detected agronomically important genes/QTLs to chromosomes. These resources should contribute immediately to advancing molecular breeding programmes and disease resistance initiatives, and will facilitate the basic functional characterization of many important and long-sought wheat genes.

Methods

Genome sequencing. The genomic DNA of *Ae. tauschii* AL8/78 was used to construct multiple types of libraries, including short insert size (450 bp) libraries, mate-paired (2 kb, 5 kb, 8 kb, 20 kb and 40 kb) libraries and PacBio SMRT Cell libraries. For the 450 bp short inserts, the library was sequenced on an Illumina HisSeq2500 instrument with 250 bp per end. In total, we produced over 778 Gb of short read sequences (Supplementary Table 1). PacBio SMRT Cell libraries were sequenced with a PacBio RS II instrument; over 53 Gb of raw data were obtained. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Genome assembly and evaluation. The genome was assembled using the software package DeNovoMAGIC2 (NRGene, Nes Ziona, Israel)[11,12] software (Supplementary Information 2.2), which is a DeBruijn graph-based assembler designed to efficiently extract the underlying information from raw reads to solve the complexity of the DeBruijn graph because of genome repetitiveness. Sequencing data from the PCR-Free library and the Nextera MP libraries were used for DeNovoMAGIC2 assembly. PCR duplicates, an Illumina adaptor (AGATCGGAAGAGC), and Nextera linkers (for MP libraries) were removed from the raw sequencing data. Overlapping reads from the PE 450 bp 2 × 250 bp libraries were then merged with a minimal required overlap of 10 bp to create the stitched reads. Following these pre-processing steps, merged PE reads were scanned to detect and filter reads with putative sequencing errors (reads containing a sub-sequence that does not reappear several times in other reads). The first step of the DeNovoMAGIC2 assembly algorithm consists of building a De Bruijn graph (kmer = 191 bp) of contigs from the overlapping PE reads. Next, PE reads are used to find reliable paths in the graph between contigs for repeat resolving and contig extension. Later, contigs are linked into scaffolds with PE and MP information, estimating gaps between the contigs according to the distance of PE and MP links. A final fill gap step uses PE and MP links, as well as De Bruijn graph information, to detect a unique path connecting the gap edges.

Mate-paired data (20 kb, 40 kb) were mapped to the basic assembly using bowtie (http://bowtie-bio.sourceforge.net/index.shtml), and only unique mapping reads were retained. Further scaffolding was performed by SSPACE (https://www.baseclear.com/genomics/bioinformatics/basetools/SSPACE, V3.0). PBJelly (http://www.winsite.com/Home-Education/Science/PBJelly/) was used to fill gaps using approximately 10X of SMRT sequencing data. We generated an F_2 mapping population of 490 individuals derived from a cross between the *Ae. tausch*ii accessions Y2280 and AL8/78. The F_2 individuals were grouped using JoinMap4.0 and then ordered using MSTmap. Finally, using the *kosamb*i function, we generated a high-resolution genetic map, which includes 164,872 SNPs developed by restriction site-associated DNA (RAD) tag sequencing technology; the total length is 1,153.58 cM[16]. The high-density genetic map was used to anchor the scaffolds to chromosomes using BLAST[44]. The completeness of gene regions of our assembly was evaluated using both CEGMA (Core Eukaryotic Gene Mapping Approach, http://korflab.ucdavis.edu/datasets/cegma/) and BUSCO (Benchmarking Universal Single-Copy Orthologs, http://busco.ezlab.org/).

Genome annotation. Protein-coding region identification and gene prediction were conducted using a combination of homology-based prediction, *de novo* prediction, and transcriptome-based prediction methods. Protein sequences from nine plant genomes (*B. distachyon, S. bicolor, O. sativa, Zea mays, Hordeum vulgare, Triticum aestivum, T. urartu, Setaria italic* and *Panicum virgatum*) were downloaded from Ensemble (Release 33) and were aligned to the *Ae. tauschii* assembly using TblastN[44] with an E-value cut-off of 1×10^{-5}. The BLAST hits were conjoined using Solar software[45]. GeneWise (https://www.ebi.ac.uk/Tools/psa/genewise) was used to predict the exact gene structure of the corresponding genomic regions for each BLAST hit. A collection of wheat FLcDNAs (16,807 sequences) were directly mapped to the *Ae. tauschii* genome and assembled by PASA (http://pasapipeline.github.io/). Five *ab initio* gene prediction programs, Augustus (http://augustus.gobics.de/, version 2.5.5), Genscan (http://genes.mit.edu/GENSCAN.html, version 1.0), GlimmerHMM (http://ccb.jhu.edu/software/glimmerhmm/, version 3.0.1),

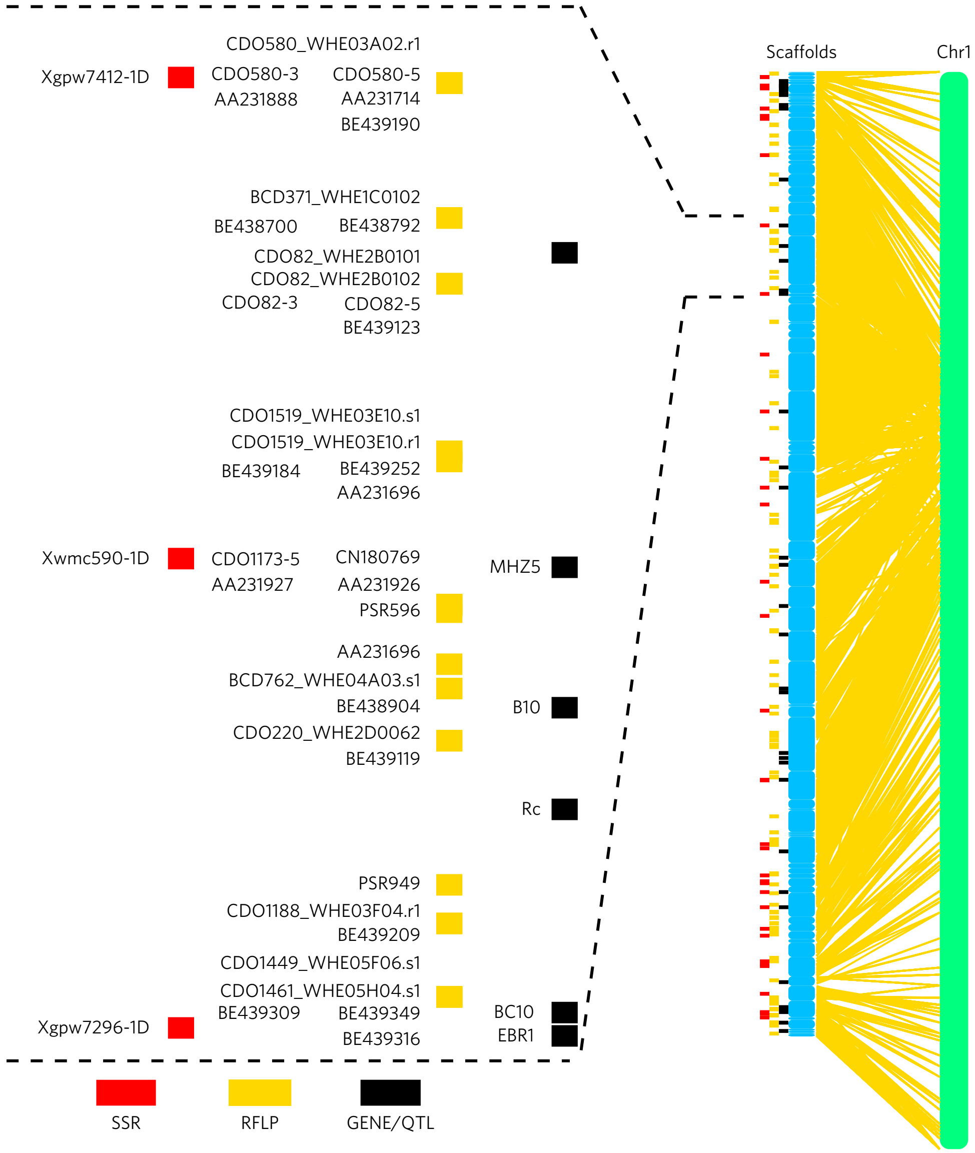

Fig. 5 | A high-resolution integrated genetic map which assists anchoring agronomically important genes/QTLs of *Ae. tauschii.* Different types of molecular markers, including RFLPs, SSRs and SNPs, were anchored to the genome. By utilizing this integrated genetic map, previously detected genes and QTLs were accurately located on chromosome 1D. The complete information for all seven chromosomes is illustrated in Supplementary Tables 26–29.

Geneid (http://genome.crg.es/software/geneid/) and SNAP (http://korflab.ucdavis.edu/software.html), were used to predict coding regions in the repeat-masked genome. RNA-seq data were mapped to the assembly using Tophat (http://ccb.jhu.edu/software/tophat/index.shtml, version 2.0.8). Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/, version 2.1.1) was then used to assemble the transcripts into gene models. Functional annotation of protein-coding genes was achieved using BLASTP[46] (E-value 1×10^{-5}) against two integrated protein sequence databases: SwissProt (http://web.expasy.org/docs/swiss-prot_guideline.html) and NR. Protein domains were annotated by searching against the InterPro (http://www.ebi.ac.uk/interpro/, V32.0) and Pfam databases (http://pfam.xfam.org/, V27.0), using InterProScan (V4.8) and HMMER (http://www.hmmer.org/, V3.1), respectively. The Gene Ontology (GO, http://www.geneontology.org/page/go-database) terms for each gene were obtained from the corresponding InterPro or Pfam entry. The pathways in which the genes might be involved were assigned by BLAST against the KEGG database (http://www.kegg.jp/kegg/kegg1.html, release 53), with an E-value cut-off of 1×10^{-5}.

RNA sequencing and analysis. To aid with gene annotation and to address many biological questions using gene expression level information, we produced a total of 53.21 Gb of RNA-seq data from eight different organs, including pistil, root, young seed, young spikes, young stamen, stem, young leaf and sheath. For this RNA-seq analysis, detailed information about treatment of plant material and experimental process has been described in the literature previously[16]. RNA-seq data were mapped to the genome using Tophat (version 2.0.8). Only the aligned reads located within 600 bp of each other were defined as concordantly mapped pairs; these were used in the downstream quantification analysis. The minimum and maximum intron length was set to 5 bp and 50,000 bp respectively. All other parameters were set to the default values. cufflinks30 (version 2.1.1) (http://cufflinks.cbcb.umd.edu/) was then used to estimate the expression level for each gene based on reads that have been uniquely mapped to the genome. An 'expressed gene' was defined as a gene with RPKM > 1 in leaf or root organs of 3-week-old seedlings. The remaining genes were defined as 'non-expressed genes'.

TE analysis. We studied the evolutionary relationships of two class I and five class II TEs by genome-wide identification of RT/DDE domains; we then constructed phylogenetic trees. To identify RT and DDE domains, we first searched known representative RT/DDE sequences against the assembly using TBLASTN[44] and extracted all regions of E-value less than 1×10^{-10}, and then excluded overlapping hits and retained regions for which the size ≥ 50% of representative sequences. Known representative sequences for DDE domains were from Yuan and Wessler[47], and representative sequences for RT domains were from an in-house plant TE database. A total of 100,879 domain regions were identified.

Giemsa-C banding and fluorescence *in situ* hybridization (FISH). Chromosome preparation from *Ae. tauschii* accession AL8/78 root tips was performed as described by Han et al.[48]. The chromosome Giemsa-C banding procedure followed

Gill et al.[49]. Sequential fluorescence *in situ* hybridization (FISH) with synthesized labelled oligonucleotide probes Oligo-pSc119.2 and Oligo-pTa535 was performed as described by Tang et al.[50]. Images were captured with an Olympus BX-51 microscope equipped with a DP-70 CCD camera. The karyotype of C-banding and FISH patterns of *Ae. tauschii* chromosomes were identified according to homoeology with the D genome of wheat[49].

Whole genome bisulfite sequencing. Bisulfite data from Illumina sequencing was aligned to the *Ae. tauschii* genome using Bismark[51] (version 0.12.5) with bowtie 2, requiring perfect matches. The mapping-quality was set to 10 to filter the mapping results to obtain unique mapping reads. Bismark extractors were used to identify three types of methylation (CHG, CHH, and CG). The visualization of the methylation results was conducted using deeptools[52].

Genome evolutionary analysis. Genome structural syntenic analyses were performed with the MCScanx toolkit[23]. Owing to the large number of recent duplicates found in the wheat D genome, top 20 BLASTP gene pairs were used as inputs for MCScanx to see if their E-value was less than 1×10^{-5}. Paralogous pairs of sequences were identified from the best reciprocal matches in all-by-all BLASTP searches. For each pair of homologous genes, protein sequences were aligned using CLUSTALW2[53], and nucleotide sequences were then forced to fit the amino acid alignments using PAL2NAL[54]. K_s values were calculated using the Nei–Gojobori algorithm[55] implemented in the codeml package of PAML[29]. Orthogroups, or putative gene families, were constructed using the OrthoMCL method[56].

Integrated genetic map and key agronomic genes/QTLs map. We collected all of the known sequences for the molecular markers of the D genome that have been generated in the past three decades, among which there were 735 RFLP markers, 3,536 SSR markers and nearly one million SNP markers. We next anchored these markers to the genome by matching marker sequences with the D genome sequences. QTLs located on the D subgenome of common wheat were also mapped to the integrated map based on their flanking marker information. We also anchored agronomically important genes on the D genome by using a similar approach.

Life Sciences Reporting Summary. Further information on experimental design and reagents is available in the Life Sciences Reporting Summary.

Data availability. The genome sequence and the annotation are available from the National Centre for Biotechnology Information (NCBI) as BioProject ID PRJNA182898. This Whole Genome Shotgun project is deposited at DDBJ/EMBL/GenBank under accession number AOCO00000000. The version described in this paper is the second version, AOCO02000000. The data that support the findings of this study are available from the corresponding author upon request.

Received: 21 September 2017; Accepted: 30 October 2017;
Published online: 20 November 2017

References

1. Choulet, F. et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721 (2014).
2. McClintock, B. Chromosome organization and genic expression. *Cold Spring Harb. Symp. Quant. Biol.* **16**, 13–47 (1951).
3. Orgel, L. E. & Crick, F. H. C. Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607 (1980).
4. Schnable, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
5. Feschotte, C., Jiang, N. & Wessler, S. R. Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* **3**, 329–341 (2002).
6. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285 (2007).
7. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).
8. Lisch, D. Epigenetic regulation of transposable elements in plants. *Annu. Rev. Plant Biol.* **60**, 43–66 (2009).
9. Bucher, E., Reinders, J. & Mirouze, M. Epigenetic control of transposon transcription and mobility in *Arabidopsis*. *Curr. Opin. Plant Biol.* **15**, 503–510 (2012).
10. Lisch, D. Regulation of the mutator system of transposons in maize. *Methods Mol. Biol.* **1057**, 123–142 (2013).
11. Lu, F. et al. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun.* **6**, 6914 (2015).
12. Avni, R. et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93–97 (2017).
13. Brenchley, R. et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705–710 (2012).
14. The International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat *(Triticum aestivum)* genome. *Science* **345**, 1251788 (2014).
15. Ling, H. Q. et al. Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87–90 (2013).
16. Jia, J. et al. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91–95 (2013).
17. Zimin, A. V. et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).
18. Zou, C. et al. Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.* **151**, 3–15 (2009).
19. Thibaud-Nissen, F., Ouyang, S. & Buell, C. R. Identification and characterization of pseudogenes in the rice gene complement. *BMC Genomics* **10**, 317 (2009).
20. Xiao, J. et al. Pseudogenes and their genome-wide prediction in plants. *Int. J. Mol. Sci.* **17**, 1991 (2016).
21. Law, M. et al. Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. *Plant Physiol.* **167**, 25–39 (2015).
22. Jiao, Y., Li, J., Tang, H. & Paterson, A. H. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *The Plant Cell* **26**, 2792–2802 (2014).
23. Wang, Y. et al. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucl. Acids Res.* **40**, e49 (2012).
24. Tang, H., Bowers, J. E., Wang, X. & Paterson, A. H. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc. Natl Acad. Sci. USA* **107**, 472–477 (2010).
25. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
26. Paterson, A. H. et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
27. The International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
28. Devos, K. M. Grass genome organization and evolution. *Curr. Opin. Plant Biol.* **13**, 139–145 (2010).
29. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
30. Pink, R. C. & Carter, D. R. Pseudogenes as regulators of biological function. *Essays in Biochemistry* **54**, 103 (2013).
31. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038 (2010).
32. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
33. Li, Q. & Eichten, S. R. Genetic perturbation of the maize methylome. *The Plant Cell* **26**, 4602–4616 (2014).
34. Eichten, S. R. & Stuart, T. DNA methylation profiles of diverse *Brachypodium distachyon* align with underlying genetic diversity. *Genome Res.* **26**, 1520–1531 (2016).
35. Li, X. et al. Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics* **13**, 300 (2012).
36. Alonso, C., Perez, R., Bazaga, P. & Herrera, C. M. Global DNA cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms. *Front. Genet.* **6**, 4 (2015).
37. Li, Q. et al. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proc. Natl Acad. Sci. USA* **112**, 14728–14733 (2015).
38. Gent, J. I. et al. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Res.* **23**, 628–637 (2013).
39. Regulski, M. et al. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res.* **23**, 1651–1662 (2013).
40. Sharp, P. J., Chao, S., Desai, S. & Gale, M. D. The isolation, characterization and application in the Triticeae of a set of wheat RFLP probes identifying each homoeologous chromosome arm. *Theor. Appl. Genet.* **78**, 342–348 (1989).
41. Somers, D. J., Isaac, P. & Edwards, K. A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **109**, 1105–1114 (2004).
42. Wang, S. et al. Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnol. J.* **12**, 787–796 (2014).
43. Winfield, M. O. et al. High-density SNP genotyping array for hexaploid wheat and its secondary and tertiary gene pool. *Plant Biotechnol. J.* **14**, 1195–1206 (2016).
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
45. Yu, X. J., Zheng, H. K., Wang, J., Wang, W. & Su, B. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. *Genomics* **88**, 745–751 (2006).
46. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402 (1997).

47. Yuan, Y. W. & Wessler, S. R. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl Acad. Sci. USA* **108**, 7884–7889 (2011).
48. Han, F., Lamb, J. C. & Birchler, J. A. High frequency of centromere inactivation resulting in stable dicentric chromosomes of maize. *Proc. Natl Acad. Sci. USA* **103**, 3238–3243 (2006).
49. Gill, B. S., Friebe, B. & Endo, T. R. Standard karyotype and nomenclature system for description of chromosome bands and structural aberrations in wheat (*Triticum aestivum*). *Genome* **34**, 830–839 (1991).
50. Tang, Z., Yang, Z. & Fu, S. Oligonucleotides replacing the roles of repetitive sequences pAs1, pSc119.2, pTa-535, pTa71, CCS1, and pAWRC.1 for FISH analysis. *J. Appl. Genet.* **55**, 313–318 (2014).
51. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
52. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucl. Acids Res.* **42**, 187–191 (2014).
53. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
54. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucl. Acids Res.* **34**, W609–W612 (2006).
55. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
56. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).

Acknowledgements

We thank J. Dvorak and M.C. Luo for the AL8/78 line; L. Pan, L.L. Zheng, Y.H. Liu, X.Y. Zhang and D.P. Li for material preparation. This research was supported by the National Key R&D Program for Crop Breeding (2016YFD0101004), the National Natural Science Foundation of China (31261140368), and the CAAS-Innovation Team Project. Y.N. Jiao acknowledges start-up funding from the Youth Thousand Talents Program.

Author contributions

J.J., Y.J., X.K., L.M. and W.J. initiated the project and designed the study. G.Z., C.Z., K.L., K.W., T.L., L.G., X.Z., H.W. and Z.Y. performed the research. G.Z., C.Z., K.L., K.W., T.L., L.G., X.Z., H.W. and Z.Y. generated and analysed the data. J.J., Y.J., X.K., W.J., G.Z., C.Z., K.L. and X.L. wrote the paper.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-017-0067-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to W.J. or L.M. or X.K. or Y.J. or J.J.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Jizeng Jia, Yuannian Jiao, Xiuying Kong, Long Mao and Wenkai Jiang

☐ Initial submission ☒ Revised version ☐ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▸ Experimental design

1. Sample size

Describe how sample size was determined.

not applicable

2. Data exclusions

Describe any data exclusions.

not applicable

3. Replication

Describe whether the experimental findings were reliably reproduced.

not applicable

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

not applicable

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

not applicable

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a | Confirmed

☒ ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)

☒ ☐ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☒ ☐ A statement indicating how many times each experiment was replicated

☒ ☐ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)

☒ ☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons

☒ ☐ The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted

☒ ☐ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)

☒ ☐ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

not applicable

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

available

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

not applicable

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

not applicable

b. Describe the method of cell line authentication used.

not applicable

c. Report whether the cell lines were tested for mycoplasma contamination.

not applicable

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

not applicable

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

not applicable

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

not applicable