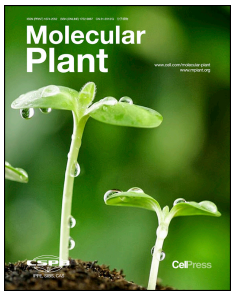
Journal Pre-proof



Gene duplication drove the loss of awn in sorghum

Leina Zhou, Can Zhu, Xiaojian Fang, Hangqin Liu, Shuyang Zhong, Yan Li, Jiacheng Liu, Yang Song, Xing Jian, Zhongwei Lin

PII: S1674-2052(21)00271-9
DOI: https://doi.org/10.1016/j.molp.2021.07.005
Reference: MOLP 1204

To appear in: *MOLECULAR PLANT*
Accepted Date: 8 July 2021

Please cite this article as: **Zhou L., Zhu C., Fang X., Liu H., Zhong S., Li Y., Liu J., Song Y., Jian X., and Lin Z.** (2021). Gene duplication drove the loss of awn in sorghum. Mol. Plant. doi: https://doi.org/10.1016/j.molp.2021.07.005.

Gene duplication drove the loss of awn in sorghum

Leina Zhou[1, †], Can Zhu[1, †], Xiaojian Fang[1], Hangqin Liu[1], Shuyang Zhong[1], Yan Li[1], Jiacheng Liu[1], Yang Song[1], Xing Jian[1], Zhongwei Lin[1,*]

[1]National Maize Improvement Center; Center for Crop Functional Genomics and Molecular Breeding; Joint Laboratory for International Cooperation in Crop Molecular Breeding, Ministry of Education; Beijing Key Laboratory of Crop Genetic Improvement, Laboratory of Crop Heterosis and Utilization, China Agricultural University, Beijing 100193, China

† These two authors contributed equally to this work.

*Corresponding author:

Zhongwei Lin

E-mail: zlin@cau.edu.cn.

Short Summary: Loss of the awn facilitates seed harvest and storage in sorghum. The *awn1* gene due to a duplication from chromosome 10 to 3 became active after recruiting a new promoter from the neighbouring region, repressed the outgrow of the awn and thus drove the loss of awn in sorghum.

Key words: awn, sorghum domestication and improvement, gene duplication, DAP-seq

Abstract

Loss of the awn in some cereals including sorghum is a key transition during cereal domestication or improvement that has facilitated grain harvest and storage. The genetic basis for the loss of awn in sorghum during domestication or improvement remains unknown. Here, we identified a transcription factor gene *awn1* encoding an ALOG domain, which is responsible for awn loss during sorghum domestication or improvement. *awn1* arose from a gene duplication from chromosome 10 that translocated to chromosome 3, recruiting a new promoter from the neighbouring intergenic region filled with "noncoding DNA", and recreating the first exon and intron. The *awn1* acquires high expr`ession after duplication and represses the elongation of awns in domesticated sorghum. Comparative mapping revealed a high collinearity at *awn1* paralog locus on chromosome 10 across cereals and awn growth and development was successfully reactivated on the rice spikelet by inactivating rice *awn1* orthologue. Further RNA-seq and DAP-seq revealed that as a transcription repressor, AWN1 directly bound to the motif in the regulatory regions from three *MADS* genes related to flower development and two genes *DL* and *LKS2* for the development of awn, downregulated the expressions of these genes, and then repressed the elongation of awn. The preexistence of regulatory elements in the neighbouring intergenic region of *awn1* before domestication signified that noncoding DNA may serve as a treasure trove for evolution during adaptation to a changing world. Our results supported that gene duplication can promptly drive the evolution of

gene regulatory network.

Introduction

The awn is a bristle-like extension on the lemma of cereal grains. Awns gradually grow out from the lemma after pollination and turn into a stiff structure at maturity. Although the awn is simple, it also holds several evolutionary benefits for seed propagation in the wild. For example, the awn acts like a self-defence weapon that protects grains from small animals and especially birds that would feed on cereal grains (Jagathesan et al., 1961). Awn facilitates grain dispersal by the wind and by allowing grains to stick to animal fur when released from the mother plant. When seeds fall from the plant at maturity, awns can adjust the trajectory and angle of the falling seed and may contribute to seed germination. In wild wheat, two awns on the grain can periodically bend with changing humidity and push the grain into the soil (Elbaum et al., 2007). Such self-planting mechanism guarantees timely germination. Especially, long awns in wheat can significantly improve yield (Jagathesan *et al.*, 1961; Rebetzke et al., 2016). These characteristics confer a selection advantage for seed propagation in the environment. The selection of awnless crops was a key factor during rice and sorghum domestication or improvement, as grains without awns make for a much easier harvest and storage.

Awn is generally controlled by quantitative trait loci (QTL) during domestication or improvement, and several genes responsible for the development of awn were

identified in different cereals. A deletion in the coding region of rice *Long and Barbed Awn1* (*LABA1*) decreased the synthesis of cytokinin in awn meristem to repress the elongation of awn (Hua et al., 2015). Multiple alleles in rice *An-1* which encoded a basic helix-loop-helix protein downregulated cell division in awn and then removed the awn (Luo et al., 2013). Rice *GRAIN NUMBER, GRAIN LENGTH AND AWN DEVELOPMENT1* (*GAD1*), which encoded a small secretary signal peptide with a EPIDERMAL PATTERNING FACTOR-LIKE domain, harbored a frame-shift insertion and regulated cell division in awn (Jin et al., 2016). Two YABBY genes *Dropping leaf* (*DL*) as a promoter and *TONGARI BOUSHI1* (*TOB1*) as a repressor, and *OsETTIN* with an auxin response factor are involved in rice awn development (Huang et al., 2020; Tanaka et al., 2012; Toriba and Hirano, 2014). In addition, rice awn length is negatively controlled by the gene for *yield, grain length and awn 1* (*gla1*) encoding a mitogen-activated protein kinase (MAPK) phosphatase (Wang et al., 2019). The *LKS2* gene with a SHI domain-containing domain regulated cell division to determine cell number during awn development in barley (Yuo et al., 2012). As a transcriptional repressor in awn development, wheat *B1* encoded a C2H2 zinc finger protein with multiple EAR motifs (Huang *et al.*, 2020). However, the genes responsible for the loss of awn during sorghum domestication or improvement remain unknown.

Here, we combined QTL fine-mapping and association mapping to identify a major QTL *awn1* on chromosome 3 responsible for awn loss in sorghum. We found that *awn1* encoded an ALOG protein and was originated from a gene duplication from

chromosome 10. After duplication, *awn1* amazingly obtained a new promoter in the intergenic region, acquired high expression and then suppressed the elongation of awn. We also conducted RNA-seq and DAP-seq to identify several downstream genes of *awn1*. Our results suggested that regulatory elements in intergenic regions may serve as a treasure trove for evolution and gene duplication can promptly reshape gene regulatory network.

Results

The major QTL *awn1* for awn loss in sorghum

Sorghum is a staple cereal that was domesticated in Africa approximately 6,000 years ago (Kimber, 2000; Lin et al., 2012). A key step during sorghum domestication or improvement was the transition from awned wild sorghum progenitor to awnless domesticated sorghum. In order to discover the genetic basis underlying this key transition during sorghum domestication or improvement, we conducted a QTL mapping analysis of awn formation in a recombination inbred line (RIL) population with 502 lines, derived from the cross between the awned wild sorghum progenitor *Sorghum virgatum* (SV) and the awnless domesticated sorghum cultivar Tx623 (Figure 1 and Supplementary Figure 1). QTL mapping revealed that awn development was controlled by a large-effect QTL at the distal end of chromosome 3 (Figure 2A) that accounted for 62% of the total phenotypic variation. The position overlapped with a site from the previous genome-wild association mapping study (Girma et al.,

2019). This large-effect QTL for awn loss was then designated as to *awn1*. Sorghum grains are composed of two glumes on the outermost whorls and two transparent and thin lemma and palea in the inner whorls (Figure 1A-E). The near-isogenic inbred line (NIL) carrying the SV allele (71~73 Mb on chromosome 3) at *awn1* showed a long awn on the lemma, while the respective NIL bearing the Tx623 *awn1* allele had a fairly short awn on the lemma (Methods and Supplementary Figure 2) that did not outgrow the glume and remained invisible from the outside (Figure 1B-E). The result suggested that *awn1* controls the elongation of the awn in sorghum.

The *awn1* gene encodes an ALOG domain is responsible for awn loss in sorghum

Initial marker screening identified 39 recombinant plants at the *awn1* locus in the RIL population of 502 individuals, and placed *awn1* between the P2 and P3 markers on chromosome 3 (Methods and Figure 2A). The markers P2 and P3 were used to screen a larger population of 3,358 plants (Supplementary Figure 3B), which were derived from selfing of 23 residual heterozygous lines (RHLs) heterozygous at the *awn1* locus and homozygous at other regions. We thus identified another 36 recombinant plants, which allowed us to narrow down the *awn1* mapping interval to within a 9.5-kb fragment according to the Tx623 reference genome (Figure 2B and Supplementary Figure 3B and C), flanked by the markers SNP2 and P9 on chromosome 3.

Sequence comparison between SV and Tx623 revealed several single nucleotide polymorphisms (SNPs) and insertions/deletions between the two parental lines within this 9.5-kb interval. Importantly, of all variants, the domesticated sorghum cultivar

Tx623 contained a large 5.4-kb insertion that harboured the only gene (Sobic.003G421300) in this region (Figure 2B and Supplementary Figure 3D). To determine which variant was responsible for the loss of awn, we conducted sequence analysis for the 9.5-kb fine-mapping fragment from 4 awned wild, 30 awned and 43 awnless domesticated sorghum accessions from a worldwide diversity panel (Supplementary Table 1). We identified 45 variants, of which 23 showed strongly significant signals ($P < 2.2\times10^{-4}$) by association testing (Figure 2C and Supplementary Table 1). The strongest signal was associated with the 5.4-kb insertion ($P = 5.4\times10^{-32}$), while other variants with significant signals were in high linkage disequilibrium with the insertion (Supplementary Figure 4). These results suggested that the 5.4-kb insertion is responsible for the loss of awn during sorghum domestication or improvement. The only gene within the 5.4-kb insertion present in domesticated sorghum Tx623, Sobic.003G421300, contained two exons and one intron and encoded a protein of 282 amino acids (aa) (Figure 2D). The protein harboured an ALOG domain from 18 aa to 151 aa based on InterPro (https://www.ebi.ac.uk/interpro/) (Figure 2D).

Gene duplication gives birth to *awn1* in sorghum

Although SV is the wild progenitor of domesticated sorghum, it lacks the 5.4-kb insertion that carries Sobic.003G421300, as do other wild awned sorghum accessions in this study (Supplementary Table 1). We therefore suspected that this 5.4-kb insertion may have arisen during sorghum domestication or improvement.

Accordingly, we compared the local synteny around the gene Sobic.003G421300 responsible for the awnless phenotype and its four neighbouring genes (Sobic.003G421100, Sobic.003G421201, Sobic.003G421400 and Sobic.003G421500) among sorghum, rice and maize. The four flanking genes showed conservation of sequence and order in all three species, with the exception of Sobic.003G421300, which was exclusively seen in sorghum (Figure 3A). The analysis indicated that the 5.4-kb insertion may have originated from another region of the sorghum genome.

To identify the possible chromosomal origin from which this 5.4-kb insertion jumped to its current location on chromosome 3, we performed a Basic Local Alignment Search Tool (BLAST) with the 5.4-kb insertion as query against the sorghum genome, which a nearly identical copy located at the position of 57 Mb on sorghum chromosome 10 (Figure 3B). The sequence on chromosome 10 was nearly identical to that of the 5.4-kb insertion from chromosome 3 (Supplementary Data 1). The *awn1* paralogous gene (Sobic.010G225100) on chromosome 10, designated as *awn1-10*, only carried one synonymous nucleotide change in the coding region from C to T compared with *awn1* (Supplementary Figure 5). Both *awn1* and *awn1-10* produced an identical protein. The *awn1-10* consisted of two exons and one intron of 485 bp upstream of the start codon (Figure 3B and Supplementary Figure 5). A careful examination of *awn1* and *awn1-10* established that *awn1* resulted from a partial duplication of *awn1-10*, comprising a partial fragment of 359 bp from the *awn1-10* intron, the entire second exon and a 3,595-bp fragment downstream of the 3' untranslated region (3' UTR) on chromosome 10 (Figure 3B and Supplementary

Figure 5). We also detected a repeat sequence of 235 bp and of unknown origin inserted into the 359-bp fragment of the *awn1-10* intron in *awn1* (Figure 3B and Supplementary Figure 5). To test how these various insertions might affect the *awn1* transcript, we performed 5' Rapid Amplification of cDNA Ends (RACE) analysis: we discovered a new first exon upstream of the *awn1* start codon. This new exon of 527 bp consisted of a 276-bp fragment from the inserted *awn1-10* intron, the repeat sequence of 235 bp and a neighbouring segment of 16 bp from the sequences upstream of the insertion site (Figure 3B and Supplementary Figure 5). By contrast, 3' RACE analysis indicated that *awn1* and *awn1-10* transcripts have an identical 3' UTR of 317 bp (Figure 3B and Supplementary Figure 5). Therefore, the *awn1* transcript carries a 5' UTR distinct from that of *awn1-10* transcript, while encoding an identical protein, aside from a silent SNP in the coding region.

We then tested *awn1* and *awn1-10* relative transcript levels across multiple organs: *awn1* was expressed in young growing panicles at the inflorescence development stage 3 and stage 4 (Jiao et al., 2018), leaves and roots (Figure 3C and see Methods). *awn1* was more highly expressed in panicles when compared to leaves or roots, with the strongest expression detected in 3-cm panicles (Figure 3C). Similarly, *awn1-10* was expressed in the same tissues (Figure 3C), with higher transcript levels in panicles relative to leaves or roots, and the highest transcript levels seen in 3-cm panicles (Figure 3C). Notably, *awn1* expression rose to much higher levels than *awn1-10* across all tissues tested here (Figure 3C). These results suggested that the gene duplication in domesticated sorghum was associated with a large increase in

199 *awn1* expression.

200

201 **The function of *awn1* remains conserved in rice**

202 We next conducted a comparative genomic analysis at the *awn1-10* locus across the
203 staple cereals sorghum, rice, maize and wheat wild progenitor (*Aegilops tauschii*),
204 which revealed a syntenic block that covered chromosomal fragments for sorghum
205 chromosome 10, rice chromosome 6, maize chromosomes 6 and 9, and wheat wild
206 progenitor chromosome 7 (Figure 3E). Despite a few reversions and deletions, this
207 syntenic block retained most of the genes and gene order. For the *awn1-10* gene in
208 particular, both sequence and gene structure were highly conserved among these
209 cereal species (Figure 3E and F, and Supplementary Figure 6). It is worth noting here
210 that *awn1-10* was duplicated in maize, with orthologues on chromosome 6 and
211 chromosome 9 (Figure 3E). Cereal ALOG proteins were divided into five clusters and
212 the AWN1 proteins then split into cluster A (Li et al., 2019). To determine whether the
213 function of *awn1* remains conserved in awn development, we turned to genome
214 editing by CRISPR/Cas9 to generate loss of function plants in rice *awn1* gene. We
215 obtained four independent rice events (RE-1, RE-2, RE-3 and RE-4) in the *japonica*
216 rice cultivar Zhonghua11. All gene-edited rice plants carried deletions in the *awn1*
217 coding region that introduced frameshift resulting in early termination of translation
218 (Supplementary Figure 7A). These edited rice plants grew long awns on the lemmas,
219 in sharp contrast to the awnless non-transgenic rice plants (Figure 3G-I and
220 Supplementary Figure 7). These results suggested that the function of *awn1* is

conserved between sorghum and rice.

Sorghum *awn1* recruited a new promoter

The duplication of *awn1-10* from chromosome 10 that generated *awn1* on chromosome 3 was limited to the coding region and 3' UTR, such that the *awn1-10* promoter and 5' UTR are absent from the *awn1* locus, raising an obvious question as to the mechanism driving *awn1* expression (Figure 3B). To answer this question, we first looked at the two genes flanking *awn1*: Sobic.003G421201 and Sobic.003G421400 exhibited different expression patterns and had lower expression levels in the panicle when compared to *awn1* (Figure 3D). This result provided evidence that the *awn1* promoter had not been recruited from the two neighboring genes (Sobic.003G421201 and Sobic.003G421400). We then turned to the chromatin accessibility landscape of the *awn1* promoter, as determined by Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) from a public repository (http://epigenome.genetics.uga.edu/PlantEpigenome/) (Lu et al., 2019). We discovered a region with high chromatin accessibility, located –2,000 bp to –1,500 bp upstream of the *awn1* translation start codon (Figure 4A).

Next, we conducted transient expression assays, with the luciferase reporter gene (*LUC*) driven by a truncated promoter series, targeting the region from –2,200 bp to –1,500 bp upstream of *awn1* with high levels of chromatin accessibility. The 1,500-bp *awn1* promoter fragment did not induce luciferase activity over a promoter-less empty vector (Figure 4B). By contrast, longer promoters, ranging from 1,660 bp to 2,140 bp

in length, strongly upregulated luciferase activity ($P < 1.0 \times 10^{-3}$) (Figure 4B). Based on these results, we defined five promoter segments between –2,200 bp and –1,500 bp to use as probes for electrophoretic mobility shift assays (EMSAs) *in vitro*. We detected obvious signals with the probes Pr1 (from –2,140 bp to –2,000 bp) and Pr4 (from –1,770 bp to –1,600 bp) after incubation with nuclear extracts from young panicles (Figure 4C) at the inflorescence development stage 4. To provide an independent validation of the strong transcriptional activity displayed by the *awn1* promoter, we introduced a *β-GLUCURONIDASE* (*GUS*) reporter construct driven by the *awn1* promoter into rice plants and monitored GUS activity: rice glumes showed strong staining in the transgenic plants (Figure 4D). These results indicated that *cis*-regulatory elements in the neighbouring intergenic region can activate *awn1* transcription in sorghum.

awn1 is a transcriptional repressor that controls awn elongation in sorghum

AWN1 belongs to a family of transcription factors with an ALOG domain. To assess the subcellular localization of AWN1, we introduced a construct encoding an AWN1-GFP fusion protein into onion epidermal cells and maize leaf protoplasts. We observed fluorescent signals for AWN1-GFP in the nucleus of both onion epidermal cells and leaf protoplasts (Figure 5A and Supplementary Figure 8). To test whether AWN1 functions as a transcription factor, we performed transcriptional activity assays by generating chimeric proteins whereby AWN1 was fused to the DNA-binding domain from the yeast GAL4 transcription factor (GAL4-DB) and to the activation

domain from herpes simplex virus protein16 (VP16). The reporter construct for these assays consisted of the luciferase reporter gene driven by a synthetic promoter comprising five copies of the GAL4 upstream activating sequence (UAS) and a TATA box (Figure 5B). While GAL4DB-VP16 strongly activated luciferase expression (as measured by high luciferase activity), the GAL4DB-VP16-AWN1 chimeric protein dramatically repressed luciferase activity by over 15-fold from the same reporter (Figure 5C). The results were independently validated in a yeast two-hybrid assay (Figure 5C and Supplementary Figure 9). The yeast two-hybrid assay showed that the AWN1 and BD (GAL4 DNA binding domain) fusion protein in yeast did not activate the expression of the reporter gene (Supplementary Figure 9), implying that AWN1 has no transcriptional activation. These results indicated that the transcription factor AWN1 acts as a transcriptional repressor.

To better understand the awn development in wild and domesticated sorghum, we made careful observations of developing spikelets from NIL-SV and NIL-Tx623. The awn meristem first appeared on the top of the lemma and was synchronized with the differentiation of the stamen and pistil meristems in both NIL-SV and NIL-Tx623 (Figure 5D and Supplementary Figure 10). The awn meristem developed further at the tip of the lemma and grew to extend beyond the glume in NIL-SV (Figure 5D and Supplementary Figure 10). In sharp contrast, the awn development remained inhibited and became gradually covered by the glume and invisible in NIL-Tx623 (Figure 5D and Supplementary Figure 10).

To explore the role of *awn1* at the molecular level during awn development, we

performed RNA-seq at the end of the inflorescence development stage 4 and DNA affinity purification sequencing (DAP-seq) analyses from spikelets harvested from NIL-SV and NIL-Tx623. From 4,044 differentially expressed (DE) genes (q-value < 0.05), we identified 2,477 downregulated and 1,567 upregulated in NIL-Tx623 relative to NIL-SV (Supplementary Table 2). Gene ontology (GO) analysis (http://systemsbiology.cau.edu.cn/agriGOv2/) identified 32 significantly enriched GO terms (FDR < 0.05) (Supplementary Table 3). The most significantly enriched GO term related to molecular function (F) was transcription factor activity (Supplementary Table 3). Auxin was involved in the development of awn in rice and wheat (Huang *et al.*, 2020; Toriba and Hirano, 2014). The analysis of RNA-seq then identified 25 DE genes related to the auxin pathway. Most of these genes (18 out of 25) for auxin were downregulated after the duplication of *awn1* (Supplementary Table 2 and Supplementary Figure 11A). The 18 DE genes in the pathway of auxin mainly encoded auxin efflux carriers, auxin response factors, auxin-responsive GH3 families and AUX/IAA transcriptional regulator families (Supplementary Figure 11A). In addition, we detected 42,443 AWN1 binding peaks that located within the regulatory regions of 16,066 genes by DAP-seq (Supplementary Figure 12 and Supplementary Table 4). A motif of 11 bp was predominantly enriched in most of these binding sites (Figure 5E and Supplementary Figure 12C). In agreement, we determined that 36% of the DE genes (1,444 out of 4,044) contained the AWN1 binding site in their regulatory regions (<= 3kb upstream of the start codon of a gene), based on DAP-seq results (Supplementary Tables 2 and 4). We noticed 17 genes encoding MADS-box

transcription factors, two YABBY genes and one SHORT INTERNODE (SHI)-like gene in the list of DE genes (Supplementary Tables 2 and 5). Most of these MADS-box genes are involved in flower development (Cui et al., 2010; Li et al., 2011b; Li et al., 2010; Li et al., 2011a; Yamaguchi et al., 2006), while the rice YABBY gene *DL* (Toriba and Hirano, 2014) and the SHI-related gene *LKS2* (Yuo *et al.*, 2012) are related to awn development. Satisfyingly, the intergenic regions (promoters, 5'UTRs, 3'UTRs and introns) of these genes were targeted by the protein of AWN1 in our DAP-seq dataset (Supplementary Table 4).

We then selected the three MADS-box genes *MADS3*, *MADS6* and *MADS7*, as well as the sorghum orthologues for *DL* and *LKS2* for further analysis. All five genes were downregulated by the *awn1* in NIL-Tx623 relative to NIL-SV, based on RT-qPCR and RNA-seq data (Supplementary Figure 11B andC). AWN1 showed a strong preference for binding to the promoter regions of the MADS-box and SHI genes (Supplementary Figure 13), while AWN1 had a strong binding peak in a *DL* intron (Supplementary Figure 13). We repeated our transient luciferase activity assay in leaf mesophyll protoplasts with AWN1 as the effector and new constructs whereby the *MADS3*, *MADS6*, *MADS7* and *LKS2* promoters were driving the luciferase reporter gene. In all cases, we observed a strong repression of luciferase activity by AWN1 (Figure 5F and G). Likewise, EMSA revealed the direct binding of Halo-tagged AWN1 to the core 5-bp sequence within the three 11-bp motifs in the *LKS2* promoter, whose gene has a known role in awn elongation (Figure 5H and I and Supplementary Figure 14). Collectively, these results demonstrate that *awn1* represses

awn elongation through the downregulation of genes including *DL* and *LKS2*.

Discussion

Noncoding DNA is not junk but a buried treasure of regulatory elements for evolution

The genomes of all staple cereals have been sequenced across multiple domesticated and wild species. Genome sequence revealed that only a fraction of these genomes (< 20%) encode genes, with the remaining genomic space composed of noncoding sequences (Jiao et al., 2017; Luo et al., 2017; Matsumoto et al., 2005; Paterson *et al.*, 2009). This feature is shared between most organisms with large genomes (Venter et al., 2001). These noncoding sequences appear to have no obvious function and have thus been named junk DNA (Ohno, 1972). However, whether junk DNA has any function for any species is still open for a lively debate (Palazzo and Gregory, 2014; Pennisi, 2012). In this study, *awn1* was duplicated from an ancestral copy mapping to chromosome 10 and inserted into chromosome 3, which resulted in the loss of awn during sorghum domestication or improvement. Amazingly, this new allele recruited a completely new promoter in the neighbouring intergenic region of *awn1* that was originally filled with noncoding DNA (Figure 3B-D). This new promoter was not borrowed from either of the two genes flanking *awn1*, but already existed in the intergenic region before the *awn1* insertion took place. This observation provides a strong support toward the claim that noncoding DNA is not in fact junk, but may instead contribute a large reservoir of regulatory elements as a buried treasure for

353 evolution. These untapped regulatory elements in the noncoding space may thus
354 constitute a great and unpredictable potential for species adaptation to a changing
355 world over the course of evolution.

356

357 **The downstream genes regulated by AWN1**

358 Sorghum awn arises from an awn primordium at the tip of the lemma that then
359 gradually grows out to form the awn. In this study, we determined that *awn1*, the
360 causal gene for a large-effect QTL for awn loss during sorghum domestication or
361 improvement, was a transcription repressor that did not affect the formation of the
362 awn primordium but rather regulated awn outgrowth (Figure 5D). RNA-seq and
363 DAP-seq analysis revealed that *awn1* directly regulated multiple transcription factors,
364 including several MADS-box, YABBY and SHI family members. The *LKS2* gene
365 encodes a SHI domain-containing protein that does not affect cell size and regulates
366 cell number during awn development in barley (Yuo *et al.*, 2012). *MADS1* is involved
367 in awn development in wheat (Huang *et al.*, 2020). Rice *DL* gene can promote the
368 development of awn (Toriba and Hirano, 2014). *MADS3* and *MADS6* can directly
369 intact with *DL* to control the specification of plant floral organ identity and meristem
370 determinacy (Cui *et al.*, 2010; Li *et al.*, 2011b; Li *et al.*, 2010; Li *et al.*, 2011a;
371 Yamaguchi *et al.*, 2006). These facts suggested that these *MADS* gene might be
372 involved in the development of awn. Auxin plays an important role in awn
373 development (Huang *et al.*, 2020; Toriba and Hirano, 2014). *LKS2* affects auxin
374 homeostasis (Yuo et al., 2012). The gene of auxin-response *GH3* family *GH3-8* is

regulated by *MADS1* and *MADS6* in rice (Yadav et al., 2011). In this study, 18 genes in the pathway of auxin were downregulated by *awn1* (Supplementary Figure 11A). Specifically, the transcription of *GH3-8* was also downregulated by the *awn1* gene (Supplementary Figure 11D). The transcriptional repressor AWN1 will directly or indirectly downregulate the expressions of these genes (Figure 6B). Consequently, the growth of the awn will remain dormant and yield the awnless grains during sorghum domestication or improvement.

Gene duplication drove the loss of awn in sorghum

Awn as an important taxonomic trait holds an important place during cereal domestication or improvement. The loss of awn generally facilitates harvest and grain storage. Gene duplication as a key factor plays an important role in evolution, e. g., the duplication and loss of REDUCED COMPLEXITY (RCO) homeodomain protein contribute to leaf shape diversity in the Brassicaceae family (Long et al., 2013; Panchy et al., 2016; Rensing, 2014; Vlad et al., 2014). In this study, a parallel, albeit smaller and recent gene duplication event of *awn1* occurred in sorghum, between chromosomes 3 and 10. The duplication of *awn1* in sorghum caused a drastic change in gene dosage, as *awn1* recruited a completely new promoter that was not associated with the ancestral copy (Figure 3). The newly recruited promoter will quickly incorporate different signaling pathways and give rise to new phenotypes for fitness: gene duplication thus drives the evolution of gene regulatory networks. Our results provide a case that a new gene or regulatory network can be promptly created during

evolution.

This 5.4-kb fragment has no transposable elements, how such fragments translocate in plant genome remains largely unknown. Structural variants based on translocations such as the *awn1* locus may play an important role in plant domestication and new phenotype creation during evolution. To identify whether the 5.4-kb duplication occurs before domestication, we sequenced this 5.4-kb fragment on chromosome 10 from SV and Tx623 (Supplementary Data2). 49 SNPs were present in this 5.4-kb fragment between wild sorghum SV and domesticated sorghum Tx623, while only 11 SNPs were present between *awn1* and *awn1-10* in Tx623 (Supplementary Data1). Much more SNPs present in this fragment with *awn1-10* between wild sorghum and domesticated sorghum indicated that the 5.4-kb duplication might not occur prior to domestication. Clear selection signals occurred at the neighbouring regions of the domesticated *awn1* allele in sorghum (Supplementary Figure 15). We hypothesize that this *awn1* allele spread slowly across sorghum varieties under human selection after the initial gene duplication event, such that most modern-day domesticated sorghums harbour the 5.4-kb insertion of *awn1*.

In summary, gene duplication enhanced the expression of *awn1*, prevented the elongation of the awn, and thus drove the loss of awn on the grain in sorghum.

Methods

Plant materials

A RIL population (F_5 generation) with 502 individuals was derived from a cross between the wild sorghum progenitor *Sorghum virgatum* (SV) and the domesticated sorghum cultivar Tx623. SV is from Egypt and shows typical wild grass characteristics such as shattering, multiple branching, and early heading(Liu et al., 2015). This RIL population and a global sorghum population consisting of 4 awned wild sorghums and 30 awned and 43 awnless domesticated sorghums (Supplementary Table 1) were planted for phenotyping in a randomized block design with three replicates at the China Agricultural University experimental station in Beijing in 2015 (RILs) and 2017 (global sorghum population). The plant materials for QTL fine-mapping were grown in Hainan or Beijing between 2015 and 2017. Each plant was grown at a distance of 25 cm from its neighbours and with a row-to-row distance of 50 cm. The two NILs with SV and Tx623 *awn1* alleles were generated from selfing of a residual heterozygous line (RHL, F_{10} generation), which was heterozygous at *awn1* and homozygous at most other loci (Supplementary Figure 2).

QTL mapping and fine-mapping

We genotyped 288 out of 502 individuals from the RIL population through 198 single sequence repeat (SSR) markers (Liu et al., 2019), which were evenly distributed across the sorghum genome. A genetic map with a total length of 1,528 centimorgan (cM) and an average genetic distance of 7.7 cM between two consecutive markers was generated with the R/qtl (Broman et al., 2003) package from the R program. QTLs were detected in R/qtl using the multiple-QTL mapping method with the

441 phenotype and genotype information from 288 RILs. Simple interval mapping was
442 initially conducted, using the function scanone with the Haley–Knott regression
443 method in R/qtl; a significance threshold (P = 0.05) for the trait was determined with
444 1,000 permutations. The locations of the QTLs with logarithm of the odds (LOD)
445 scores over the threshold were subsequently refined by the function refineqtl. Based
446 on these refined QTL positions, additional QTLs were next detected with the function
447 addqtl. The additional QTLs were added into the model and the locations of all QTLs
448 refined again, when a new significant QTL with a LOD score above the threshold was
449 detected. We repeated the above steps until no more QTLs were detected. The genetic
450 effect and the significance of each QTL were determined using drop-one-QTL
451 analysis in the context of the full model after all the position of all QTLs had been
452 refined.

453 To fine-map *awn1*, we screened the entire set of 502 plants from the RIL
454 population with markers flanking *awn1*. We identified 39 recombinant plants between
455 P1 and P5, which we further genotyped with the P2 and P3 markers (Supplementary
456 Figure 3A). We then used the same P2 and P3 markers to screen a large population of
457 3,358 individuals derived from the selfing of 23 RHLs, carrying heterozygous
458 fragments at the target *awn1* locus while being homozygous at most other loci
459 (Supplementary Figure 1). Rapid marker screening identified 36 recombinant plants
460 between the P2 and P3 markers. With these 36 additional recombinant plants and six
461 newly developed markers, we narrowed down the *awn1* locus to within a region of 9.5
462 kb, flanked by the markers SNP2 and P9. The phenotype of each recombinant plant

463 was confirmed in over 20 progeny plants, which were from the selfing of each
464 corresponding recombinant plant. All primers used in this study are listed in
465 Supplementary Table 6.

466

467 **Association mapping**

468 To test whether the *awn1* gene was responsible for the development of awn in
469 sorghum, we sequenced the 9.5-kb interval defined by fine-mapping from 4 awned
470 wild sorghums, 30 awned and 43 awnless domesticated sorghums (Supplementary
471 Table 1). The awn length was also carefully measured from these sorghums. We
472 identified 45 variants with a frequency of more than 5% across the 77 sequences in
473 the fine-mapping interval. Association mapping testing was performed with a mixed
474 linear model in TASSEL5 (Bradbury et al., 2007). The significance threshold was
475 corrected for multiple testing through Bonferroni correction according to the
476 following equation: $\alpha' \approx \alpha/n = 2.2 \times 10^{-4}$, where α is the nominal significance
477 threshold ($\alpha = 0.01$) and n is the number of variants ($n = 45$).

478

479 **RNA-seq**

480 Total RNA was extracted from young panicles (5 cm, at the inflorescence
481 development stage 4 when awn grew out to extend beyond the glume (Jiao *et al.*,
482 2018).) of the two NILs with the SV or Tx623 *awn1* alleles 50 days after planting
483 (DAP) in biological triplicates, followed by treatment with RNase-Free DNase I
484 (D2215, Takara). The resulting DNA-free RNA sampled was then used as starting

materials for sequencing libraries and sequencing on an Illumina HiSeq-2500 platform. A total of 50 Gb of raw sequencing data was collected. The raw RNA-seq reads were analysed using a common RNA-seq pipeline (Zhang et al., 2019). Briefly, the raw RNA-seq reads were trimmed with Trimmomatic program (Bolger et al., 2014) and further cleaned with fastq_clean (Zhang et al., 2014). After these treatments, the clean reads were aligned to sorghum reference genomes V3.1.1 (McCormick et al., 2018a) on Phytozome (https://phytozome.jgi.doe.gov) using STAR (Dobin et al., 2013). Gene expression based on fragments per kilo-base of exon per million fragments mapped (FPKM) was next obtained through Cufflinks and cuffdiff2 (Trapnell et al., 2014). Differentially expressed (DE) genes between the two NILs were determined based on their corrected P-values (q-value < 0.05).

DNA Affinity Purification (DAP)-seq

A genomic DNA library was constructed following a reported protocol, with some modifications (Bartlett et al., 2017; O'Malley et al., 2016). Briefly, genomic DNA was extracted from the young panicles (3–5 cm) at the inflorescence development stage 3 and 4 of NIL-Tx623, fragmented to 200 bp and ligated with a truncated Illumina TruSeq adaptor to generate the DNA library. The *awn1* coding sequence was fused with Halo-Tag and expressed in Wheat Germ Extract System (Promega) in two independent experiments. Halo-tagged AWN1 was immobilized onto Magne HaloTag beads (Promega), incubated with the genomic DNA library (300 ng) for 1 h, and then washed. The washed beads with bound genomic DNA fragments were tagged with

dual-indexed multiplexing barcodes through PCR amplification for 18 cycles. The resulting libraries from the two independent replicates were pooled and sequenced on an Illumina NavoSeq 6000 sequencer. Input DNA libraries were prepared as described above to control for background noise.

The raw reads from each replicate and input DNA libraries were processed by trimming the adapter sequences and low-quality bases with fastp (Chen et al., 2018). The resulting clean reads were mapped to the sorghum reference genome V3 (http://plants.ensembl.org/) using bowtie2 (Langmead and Salzberg, 2012) v2.35. The mapped reads were then filtered using SAMtools (Li et al., 2009) 1.9 to restrict the reads that aligned to multiple locations with the parameters: -h -q 30 -F 4 -F 256. Peak calling was performed using MACS2 (Zhang et al., 2008) v2.2.7.1 with a cut off q value of 10^{-7}, using the input DNA library as control. Significant overlapping peaks from the two replicates comprised the final list of candidate peaks. We converted the bam files to bigwig files and visualized them in the Integrative Genome Browser (IGV). The most enriched motif for these overlapping peaks was determined using MEME (Bailey et al., 2009) suite v5.5.1.

Scanning electron microscopy

The young panicles from 40 to 45 DAP were fixed in 2.5% glutaraldehyde solution overnight at 4°C, dehydrated through an ethanol series (30% to 100% [v/v]). The fixed samples were then critical-point dried in liquid CO2, sputter-coated with gold, and observed under a Hitachi S-3400N SEM at 10 kV.

529

530 **Subcellular localization**

531 The full-length *awn1* coding sequence from Tx623 was cloned into the
532 pCAMBIA1300-GFP vector. The resulting *35Spro:AWN1-GFP* construct was
533 coprecipitated with gold particles and then introduced into onion epidermal cells by
534 particle bombardment using the helium biolistic device (Bio Rad PDS-1000). The
535 plasmid with *35Spro:AWN1-GFP* was also transformed into the mesophyll protoplasts
536 isolated from the leaves of 10-day-old etiolated B73 seedlings using the polyethylene
537 glycol (PEG) mediated transformation method. The onion epidermal cells and the
538 protoplasts were incubation at 25°C in the dark for 15 h. The subcellular localization of
539 AWN1-GFP was detected at 488-nm laser line on a Nikon C1 confocal laser
540 microscope.

541

542 **Electrophoretic Mobility Shift Assays (EMSAs)**

543 Biotin-labelled oligonucleotide probes were synthesized by Beijing Hippo
544 Biotechnology Company. We mixed 4 μg of nuclear extracts from young panicles (1–
545 3 cm) at the inflorescence development stage 3 or 1 μL of unpurified AWN1 protein
546 fused with HaloTag and expressed with the SP6 High-Yield Wheat Germ Protein
547 Expression system with 8 ng of biotin-labelled annealed oligonucleotides that were
548 purified from a 8% (w/v) polyacrylamide gel, 2 μL of 10× binding buffer, 1 μL of 50%
549 (v/v) glycerol, 1 μL of 100 mM $MgCl_2$, 1 μL of 1 mg/mL poly(dI-dC), 1 μL of 1%
550 (v/v) Nonidet P-40, and double-distilled water for a final reaction volume of 20 μL.

After incubation at 25°C for 20 min, the reactions were separated on 6% (w/v) polyacrylamide gels, and then transferred to N+ nylon membranes (GE). Biotin-labelled DNA was detected with a LightShift Chemiluminescent EMSA kit (Thermo Scientific) according to the manufacturer's instructions.

5' and 3' Rapid Amplification of cDNA Ends (RACE)

Total RNA was extracted from the young panicles (~1 cm) of Tx623 plants collected at 38 DAP. The RNA was then treated with RNase-free DNase I (Takara) and purified using the RNAclean kit (Tiangen). 5' and 3' RACE were then conducted with the SMART RACE cDNA Amplification Kit (Clontech) based on the manufacturer's instructions.

Plant transformation

The rice *awn1* orthologue, mapping to chromosome 6, was edited with two gRNAs via CRISPR/Cas9 genome editing in the *japonica* rice cultivar Zhonghua11. The two gRNAs of 20 bp, each driven by the rice promoters *OsU3* and *OsU6*, respectively, targeted two distinct sites in the coding region of rice *awn1* and were assembled into the CRISPR/Cas9 vector pYLCRISPR/Cas9-MH (Ma et al., 2015). Four rice transformation events (T_0) were allowed to self-pollinate and were phenotyped in the T_1 generation in the greenhouse. A reporter construct placing the *β-GLUCURONIDASE* (*GUS*) gene under the control of a 3.0-kb fragment of the sorghum *awn1* promoter from Tx623 was transformed into Zhonghua11 (ZH11) to

573 determine the tissue specificity of *awn1* expression. All homozygous edited rice plants
574 were confirmed by sequencing across the edited sites.
575
576 **Luciferase transient expression assay**
577 To examine which region in the promoter activates the transcription of the *awn1* gene,
578 a promoter deletion series ranging from 2,100 bp to 1,500 bp from Tx623 was cloned
579 into the LUC vector pGreenII 0800-LUC, which contained the luciferase reporter
580 gene from *Renilla reniformis* (REN) as an internal control under the control of the
581 35S cauliflower mosaic virus (CaMV) promoter, and the firefly luciferase reporter
582 gene (*LUC*) under the control of the *awn1* promoter fragments. These constructs were
583 transformed into maize B73 leaf protoplasts at the four-leaf stage. Freshly isolated
584 protoplasts were mixed with 20 μg DNA of the reporter construct in PEG transfer
585 solution for 18 min at room temperature before being returned to WI medium. After
586 an incubation of 14 h at 25°C, the transformed protoplasts were harvested by
587 centrifugation, lysed in Passive Lysis Buffer (PLB, Promega) and assayed following
588 the Dual-Luciferase Reporter Assay System (Promega). Three biological replicates of
589 each construct were conducted and all assays were repeated three times.
590 We also cloned promoter fragments from *LKS2* (3,516 bp), *MADS3* (2,272 bp),
591 *MADS6* (2,273 bp) and *MADS7* (3,961 bp) upstream from the start codon into the
592 pGreenII-0800-LUC vector to generate another set of reporters. The full-length
593 coding sequence for *awn1* was cloned into the pGreenII 62-SK vector and placed
594 under the control of 35S CaMV promoter as effector construct. The appropriate

595 combinations of reporters and effector constructs were co-transformed into maize leaf
596 protoplasts, using the respective reporter with the empty effector pGreenII 62-SK as
597 control. Luciferase activity was determined as described above.

598

599 **Transcriptional activity assay**

600 To determine the transcriptional activity of the AWN1 protein, we first performed a
601 transcriptional activity assay with the Matchmarker GAL4 Two-Hybrid System 3
602 (Clontech). The full-length and two truncated version of the *awn1* coding sequence
603 were cloned into the pGBKT7 vector to fuse AWN1 with the DNA-binding domain of
604 GAL4 (GAL4-BD). The transcription factor ZmCCT was fused with GAL4-BD as a
605 positive control. All constructs were transformed into yeast strain AH109 according to
606 the manufacturer's instructions. The colonies were diluted and spotted onto yeast
607 synthetic drop-out medium lacking Trp alone or lacking Trp, Ade and His. We next
608 conducted a dual-luciferase transient expression assay in maize leaf protoplasts. The
609 *awn1* coding sequence was cloned into the vector to fuse AWN1 with GAL4-DB and
610 VP16 to generate the effector construct *GAL4DB-VP16-AWN1*. For the reporter
611 construct, a promoter with 5× GAL4 UAS sequence and a TATA box was introduced
612 into pGreenII 0800-LUC. The reporter and effector constructs were co-transformed
613 into maize leaf protoplasts, using the empty effector construct as a control.

614

615 **RT-qPCR**

616 Total RNA was extracted from various plant tissues (leaves, leaf sheathes, roots,

lemma, and young panicles [1~5 cm]) at inflorescence development stage 3 and stage 4 collected from three plants per sorghum NIL (NIL-SV and NIL-Tx623) using an RNA Extraction Kit (TianGen Biotech). First-strand cDNA was synthesized from 2 μg total RNA as starting material, using the TransScript-Uni cDNA Synthesis SuperMix (TransGen Biotech). We performed quantitative PCR (qPCR) in three technical replicates and three biological replicates on a Bio-Rad CFX Maestro system, with the housekeeping sorghum *Ubiquitin* gene as internal control. The final relative transcript levels were determined via the $\Delta\Delta$CT (DDCT) relative quantification method(Livak and Schmittgen, 2001).

Comparative Mapping

Pairwise genomic sequence comparison was conducted with SYNMAP in the comparative genomics database CoGe (http://genomevolution.org/CoGe/). The syntenic map was plotted according to the BLAST results of pairwise genomes derived from these datasets in CoGe, including maize (B73, id333), rice (Nipponbare, id3), wild wheat (*Aegilops tauschii*, id40404) and sorghum (Tx623, id331). The gene identifiers for the *awn1* orthologues are Sobic.003G421300 in sorghum, LOC_Os06g46030 in rice, Zm00001d036617 and Zm00001d046998 in maize, AET7Gv21141300 in wild wheat.

DNA Diversity Analysis

The two 1.5-kb neighbouring fragments of the 5.4-kb insertion in *awn1* were

PCR-amplified from a global sorghum population with 12 wild sorghums, 103 awnless and 57 awned domesticated sorghums (Supplementary Table 7). The resulting PCR products were sequenced on an ABI 3730 sequencer after purification using the QIAquick PCR Purification Kit (Qiagen). These sequences were analysed in ClustalW to construct a nucleotide alignment matrix, which was then imported into DnaSPV5.1 to calculate nucleotide diversity (π) with a sliding window of 100 bp and a step size of 25 bp, excluding the sites with gap (Librado and Rozas, 2009). Tajima's *D* tests were also calculated by DnaSPV5.1 (Librado and Rozas, 2009).

Phylogenetic Analysis

The protein alignment of sorghum AWN1 and related proteins was imported into MEGA7 to generate a phylogenetic tree using the maximum likelihood method (Kumar et al., 2016).

Author Contributions

Z. W. L. designed the study. L. Z., C. Z., X. F., H.L., S. Z., Y. L., J. L., Y. S. and X. J. performed the research. L. Z., C. Z. and Z. W. L. analyzed the data. L. Z., C. Z. and Z.W. L. wrote the manuscript.

Acknowledgments

We thank Jianming Yu from Iowa State University for critical comments on the manuscript. This work was supported by National Natural Science Foundation of

China (92035302 and 31871632 to Z.L.), the National Key Research and Development Program of China (2016YFD0100303 and 2016YFD0101803 to Z.L.) and Chinese Universities Scientific Fund (2021TC065 to Z.L.).

Competing interests

The authors declare no conflicts of interest.

Data availability

RNA-seq data are deposited at the National Center for Biotechnology Information (NCBI) under the SRA accession number PRJNA679987. DAP-seq data are deposited at NCBI under the SRA accession number PRJNA679988.

REFERENCES

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J.Y., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res **37**:W202-W208. 10.1093/nar/gkp335.

Bartlett, A., O'Malley, R.C., Huang, S.S.C., Galli, M., Nery, J.R., Gallavotti, A., and Ecker, J.R. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat Protoc **12**:1659-1672. 10.1038/nprot.2017.055.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**:2114-2120. 10.1093/bioinformatics/btu170.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23**:2633-2635. 10.1093/bioinformatics/btm308.

Broman, K.W., Wu, H., Sen, S., and Churchill, G.A. (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics **19**:889-890. 10.1093/bioinformatics/btg112.

Chen, S.F., Zhou, Y.Q., Chen, Y.R., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34**:884-890. 10.1093/bioinformatics/bty560.

Cui, R.F., Han, J.K., Zhao, S.Z., Su, K.M., Wu, F., Du, X.Q., Xu, Q.J., Chong, K., Theissen, G., and Meng,

Z. (2010). Functional conservation and diversification of class E floral homeotic genes in rice (Oryza sativa). Plant J **61**:767-781. 10.1111/j.1365-313X.2009.04101.x.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**:15-21. 10.1093/bioinformatics/bts635.

Elbaum, R., Zaltzman, L., Burgert, I., and Fratzl, P. (2007). The role of wheat awns in the seed dispersal unit. Science **316**:884-886. 10.1126/science.1140097.

Girma, G., Nida, H., Seyoum, A., Mekonen, M., Nega, A., Lule, D., Dessalegn, K., Bekele, A., Gebreyohannes, A., Adeyanju, A., et al. (2019). A Large-Scale Genome-Wide Association Analyses of Ethiopian Sorghum Landrace Collection Reveal Loci Associated With Important Traits. Front Plant Sci **10**:691. 10.3389/fpls.2019.00691.

Hua, L., Wang, D.R., Tan, L.B., Fu, Y.C., Liu, F.X., Xiao, L.T., Zhu, Z.F., Fu, Q., Sun, X.Y., Gu, P., et al. (2015). LABA1, a Domestication Gene Associated with Long, Barbed Awns in Wild Rice. Plant Cell **27**:1875-1888. 10.1105/tpc.15.00260.

Huang, D.Q., Zheng, Q., Melchkart, T., Bekkaoui, Y., Konkin, D.J.F., Kagale, S., Martucci, M., You, F.M., Clarke, M., Adamski, N.M., et al. (2020). Dominant inhibition of awn development by a putative zinc-finger transcriptional repressor expressed at the B1 locus in wheat. New Phytologist **225**:340-355. 10.1111/nph.16154.

Jagathesan, D., Bhatia, C., and Swaminathan, M.S. (1961). Effect of Induced Awn Mutations on Yield in Wheat. Nature **190**:468-+. Doi 10.1038/190468a0.

Jiao, Y.P., Lee, Y.K., Gladman, N., Chopra, R., Christensen, S.A., Regulski, M., Burow, G., Hayes, C., Burke, J., Ware, D., et al. (2018). MSD1 regulates pedicellate spikelet fertility in sorghum through the jasmonic acid pathway. Nat Commun **9**ARTN 822 10.1038/s41467-018-03238-4.

Jiao, Y.P., Peluso, P., Shi, J.H., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X.H., Chin, C.S., et al. (2017). Improved maize reference genome with single-molecule technologies. Nature **546**:524-+. 10.1038/nature22971.

Jin, J., Hua, L., Zhu, Z.F., Tan, L.B., Zhao, X.H., Zhang, W.F., Liu, F.X., Fu, Y.C., Cai, H.W., Sun, X.Y., et al. (2016). GAD1 Encodes a Secreted Peptide That Regulates Grain Number, Grain Length, and Awn Development in Rice Domestication. Plant Cell **28**:2453-2463. 10.1105/tpc.16.00379.

Kimber, C. (2000). Origins of domesticated sorghum and its early diffusion to India and China (John Wiley & Sons).

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol **33**:1870-1874. 10.1093/molbev/msw054.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods **9**:357-U354. 10.1038/Nmeth.1923.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Proc, G.P.D. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**:2078-2079. 10.1093/bioinformatics/btp352.

Li, H.F., Liang, W.Q., Yin, C.S., Zhu, L., and Zhang, D.B. (2011a). Genetic Interaction of OsMADS3, DROOPING LEAF, and OsMADS13 in Specifying Rice Floral Organ Identities and Meristem Determinacy. Plant Physiol **156**:263-274. 10.1104/pp.111.172080.

Li, H.F., Liang, W.Q., Jia, R.D., Yin, C.S., Zong, J., Kong, H.Z., and Zhang, D.B. (2010). The AGL6-like gene OsMADS6 regulates floral organ and meristem identities in rice. Cell Res **20**:299-313.

10.1038/cr.2009.143.

Li, H.F., Liang, W.Q., Hu, Y., Zhu, L., Yin, C.S., Xu, J., Dreni, L., Kater, M.M., and Zhang, D.B. (2011b). Rice MADS6 Interacts with the Floral Homeotic Genes SUPERWOMAN1, MADS3, MADS58, MADS13, and DROOPING LEAF in Specifying Floral Organ Identities and Meristem Fate. Plant Cell **23**:2536-2552. 10.1105/tpc.111.087262.

Li, N., Wang, Y., Lu, J., and Liu, C. (2019). Genome-Wide Identification and Characterization of the ALOG Domain Genes in Rice. Int J Genomics **2019**:2146391. 10.1155/2019/2146391.

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics **25**:1451-1452. 10.1093/bioinformatics/btp187.

Lin, Z.W., Li, X.R., Shannon, L.M., Yeh, C.T., Wang, M.L., Bai, G.H., Peng, Z., Li, J.R., Trick, H.N., Clemente, T.E., et al. (2012). Parallel domestication of the Shattering1 genes in cereals. Nat Genet **44**:720-U154. 10.1038/ng.2281.

Liu, H., Liu, H., Zhou, L., Zhang, Z., Zhang, X., Wang, M., Li, H., and Lin, Z. (2015). Parallel Domestication of the Heading Date 1 Gene in Cereals. Mol Biol Evol **32**:2726-2737. 10.1093/molbev/msv148.

Liu, H.H., Liu, H.Q., Zhou, L.N., and Lin, Z.W. (2019). Genetic Architecture of domestication- and improvement-related traits using a population derived from Sorghum virgatum and Sorghum bicolor. Plant Sci **283**:135-146. 10.1016/j.plantsci.2019.02.013.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method. Methods **25**:402-408. 10.1006/meth.2001.1262.

Long, M.Y., VanKuren, N.W., Chen, S.D., and Vibranovski, M.D. (2013). New Gene Evolution: Little Did We Know. Annu Rev Genet **47**:307-333. 10.1146/annurev-genet-111212-133301.

Lu, Z.F., Marand, A.P., Ricci, W.A., Ethridge, C.L., Zhang, X.Y., and Schmitz, R.J. (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. Nat Plants **5**:1250-1259. 10.1038/s41477-019-0548-z.

Luo, J.H., Liu, H., Zhou, T.Y., Gu, B.G., Huang, X.H., Shangguan, Y.Y., Zhu, J.J., Li, Y., Zhao, Y., Wang, Y.C., et al. (2013). An-1 Encodes a Basic Helix-Loop-Helix Protein That Regulates Awn Development, Grain Size, and Grain Number in Rice. Plant Cell **25**:3360-3376. 10.1105/tpc.113.113589.

Luo, M.C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., Deal, K.R., Huo, N.X., Zhu, T.T., Wang, L., Wang, Y., et al. (2017). Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. Nature **551**:498-+. 10.1038/nature24486.

Ma, X.L., Zhang, Q.Y., Zhu, Q.L., Liu, W., Chen, Y., Qiu, R., Wang, B., Yang, Z.F., Li, H.Y., Lin, Y.R., et al. (2015). A Robust CRISPR/Cas9 System for Convenient, High-Efficiency Multiplex Genome Editing in Monocot and Dicot Plants. Mol Plant **8**:1274-1284. 10.1016/j.molp.2015.04.007.

Matsumoto, T., Wu, J.Z., Kanamori, H., Katayose, Y., Fujisawa, M., Namiki, N., Mizuno, H., Yamamoto, K., Antonio, B.A., Baba, T., et al. (2005). The map-based sequence of the rice genome. Nature **436**:793-800. 10.1038/nature03895.

McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B.D., McKinley, B., et al. (2018a). The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J **93**:338-354. 10.1111/tpj.13781.

McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S.Q., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B.D., McKinley, B., et al. (2018b). The Sorghum bicolor reference genome:

improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J **93**:338-354. 10.1111/tpj.13781.

O'Malley, R.C., Huang, S.S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape (vol 165, pg 1280, 2016). Cell **166**:1598-1598. 10.1016/j.cell.2016.08.063.

Ohno, S. (1972). So Much Junk DNA in Our Genome. Brookhaven Sym Biol:366-+.

Palazzo, A.F., and Gregory, T.R. (2014). The Case for Junk DNA. Plos Genet **10**ARTN e1004351 10.1371/journal.pgen.1004351.

Panchy, N., Lehti-Shiu, M., and Shiu, S.H. (2016). Evolution of Gene Duplication in Plants. Plant Physiol **171**:2294-2316. 10.1104/pp.16.00523.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature **457**:551-556. 10.1038/nature07723.

Pennisi, E. (2012). GENOMICS ENCODE Project Writes Eulogy For Junk DNA. Science **337**:1159-1161. DOI 10.1126/science.337.6099.1159.

Rebetzke, G.J., Bonnett, D.G., and Reynolds, M.P. (2016). Awns reduce grain number to increase grain size and harvestable yield in irrigated and rainfed spring wheat. Journal of experimental botany **67**:2573-2586. 10.1093/jxb/erw081.

Rensing, S.A. (2014). Gene duplication as a driver of plant morphogenetic evolution. Curr Opin Plant Biol **17**:43-48. 10.1016/j.pbi.2013.11.002.

Tanaka, W., Toriba, T., Ohmori, Y., Yoshida, A., Kawai, A., Mayama-Tsuchida, T., Ichikawa, H., Mitsuda, N., Ohme-Takagi, M., and Hirano, H.Y. (2012). The YABBY Gene TONGARI-BOUSHI1 Is Involved in Lateral Organ Development and Maintenance of Meristem Organization in the Rice Spikelet. Plant Cell **24**:80-95. 10.1105/tpc.111.094797.

Toriba, T., and Hirano, H.Y. (2014). The DROOPING LEAF and OsETTIN2 genes promote awn development in rice. Plant J **77**:616-626. 10.1111/tpj.12411.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2014). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks (vol 7, pg 562, 2012). Nat Protoc **9**:2513-2513. 10.1038/nprot1014-2513a.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science **291**:1304-+. DOI 10.1126/science.1058040.

Vlad, D., Kierzkowski, D., Rast, M.I., Vuolo, F., Dello Ioio, R., Galinha, C., Gan, X.C., Hajheidari, M., Hay, A., Smith, R.S., et al. (2014). Leaf Shape Evolution Through Duplication, Regulatory Diversification, and Loss of a Homeobox Gene. Science **343**:780-783. 10.1126/science.1248384.

Wang, T., Zou, T., He, Z.Y., Yuan, G.Q., Luo, T., Zhu, J., Liang, Y.Y., Deng, Q.M., Wang, S.Q., Zheng, A.P., et al. (2019). GRAIN LENGTH AND AWN 1 negatively regulates grain size in rice. J Integr Plant Biol **61**:1036-1042. 10.1111/jipb.12736.

Yadav, S.R., Khanday, I., Majhi, B.B., Veluthambi, K., and Vijayraghavan, U. (2011). Auxin-responsive OsMGH3, a common downstream target of OsMADS1 and OsMADS6, controls rice floret fertility. Plant Cell Physiol **52**:2123-2135. 10.1093/pcp/pcr142.

Yamaguchi, T., Lee, D.Y., Miyao, A., Hirochika, H., An, G.H., and Hirano, H.Y. (2006). Functional diversification of the two C-class MADS box genes OSMADS3 and OSMADS58 in Oryza sativa. Plant Cell

18:15-28. 10.1105/tpc.105.037200.

Yuo, T., Yamashita, Y., Kanamori, H., Matsumoto, T., Lundqvist, U., Sato, K., Ichii, M., Jobling, S.A., and Taketa, S. (2012). A SHORT INTERNODES (SHI) family transcription factor gene regulates awn elongation and pistil morphology in barley. Journal of experimental botany **63**:5223-5232. 10.1093/jxb/ers182.

Zhang, M., Zhan, F., Sun, H.H., Gong, X.J., Fei, Z.J., and Gao, S. (2014). Fastq_clean: an optimized pipeline to clean the Illumina sequencing data with quality control. Ieee Int C Bioinform.

Zhang, X., Lin, Z.L., Wang, J., Liu, H.Q., Zhou, L.N., Zhong, S.Y., Li, Y., Zhu, C., Liu, J.C., and Lin, Z.W. (2019). The tin1 gene retains the function of promoting tillering in maize. Nat Commun **10**Artn 5608 10.1038/S41467-019-13425-6.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol **9**Artn R137

10.1186/Gb-2008-9-9-R137.

852 **Figure legends**

853

854 **Figure 1. Phenotypes of *awn1*.**

855 (**A**) The awn is visible on the glumes from NIL-SV, the NIL carrying the wild
856 sorghum (SV) *awn1* allele, but is absent on the glume from NIL-Tx623, the NIL
857 carrying the domesticated sorghum Tx623 (NIL-Tx623) *awn1* allele. (**B,C**) Close-up
858 view of the glumes (**B**) and lemmas (**C**) from NIL-SV. (**D,E**) Close-up view of the
859 glumes (**D**) and lemmas (**E**) from NIL-Tx623 after removing the glumes. The awn
860 remains dormant in NIL-Tx623. gl, glume; le, lemma; pe, pelea.

861 **Figure 2. Fine-mapping of the *awn1* locus.**

862 (**A**) QTL mapping identifies the major QTL *awn1* for awn loss at the bottom of
863 chromosome 3. The red dashed line represents the significance threshold at P = 0.05
864 level. (**B**) Fine-mapping of *awn1* in a large population of 3,358 individuals. The
865 genotypes of two representative recombinant plants between the SNP2 and P9
866 markers are represented graphically. Blue, green and orange bars represent the
867 chromosomal fragments from wild sorghum (SV), domesticated sorghum (Tx623) and
868 heterozygous plants, respectively. Pink flag, the target gene *awn1*; gray bar, the 5.4-kb
869 insertion discovered in Tx623; pink bar, the *awn1* coding region; scale bar, 500 bp. (**C**)
870 Association mapping revealed that the 5.4-kb insertion of *awn1* within the final 9.5-kb
871 fine-mapping interval is responsible for loss of awn in domesticated sorghum. The
872 gene structure is shown on the x axis. Non-significant SNPs are indicated by green
873 circles, while significant SNPs are shown as red circles. A red filled circle

corresponding to the 5.4-kb deletion/insertion was arbitrarily placed in the centre of the 5.4-kb insertion. The red dashed line indicates the significance threshold (3.7) at P =0.01 level after multiple testing correction for 45 variants in the fine-mapping region. (**D**) AWN1 protein sequence, with the ALOG domain marked in the red dashed box.

Figure 3. Gene duplication of *awn1* in cereals.

(**A**) Synteny comparison for *awn1* and its four closest neighbouring genes in sorghum, rice and maize. The sequence and order of these four flanking genes are conserved between these three species. Sb, sorghum; Os, rice; Zm, maize. The bars in the same colour represent orthologues across different species. (**B**) The 5.4-kb insertion of *awn1* on sorghum chromosome 3 was duplicated from *awn1-10* on chromosome 10. The *awn1* and *awn1-10* sequences are nearly identical in the second exon and the additional 3.6-kb fragment downstream of 3' UTR. The new first exon of *awn1* is composed of a 276-bp fragment (green box) from the *awn1-10* intron, a fragment of 235 bp of unknown origin (gray box) and a 16-bp flanking sequence (orange box) at the insertion site on chromosome 3. Blue boxes, 5' and 3' untranslated regions; pink boxes, coding regions; blue and orange arrows, promoters; scale bar, 500 bp. (**C**) Relative expression levels of *awn1* from NIL-Tx623 and *awn1-10* from both NILs in young panicles (YP-1, 3 and 5 cm), leaves and roots. (**D**) Relative transcript levels of *awn1* and its two neighbouring genes (Sobic.003G421201 and Sobic.003G421400) in young panicles (YP), leaves, leaf sheaths (LS) and lemmas from Tx623 plants. (**E**) Comparative genomics reveals a syntenic block corresponding to *awn1-10* on

sorghum (Sb) chromosome 10, rice (Os) chromosome 6, maize (Zm) chromosome 6 and 9 and wild wheat (Aet) chromosome 7. Red point indicates the *awn1-10* gene. (**F**) Phylogenetic tree for sorghum AWN1-10 and related proteins from the four cereal species. (**G-I**) Awns were present on the lemmas of the edited rice line obtained by CRISPR/Cas9 (*RE-1*, top row) compared to control plant (ZH11, bottom row). Close-up views of panicle branch (**H**) and glume (**I**).

Figure 4. Domesticated sorghum *awn1* recruited a completely new promoter.

(**A**) Chromatin accessibility in the promoter region of domesticated sorghum *awn1*, based on ATAC-seq data obtained from the database http://epigenome.genetics.uga.edu/PlantEpigenome/. A region from –2,000 bp to –1,500 bp upstream of the *awn1* start codon shows high chromatin accessibility across two replicates. The A base from the start codon was regarded as position +1. (**B**) Transient expression assays with a firefly luciferase reporter construct (*LUC*) driven by a series of truncated promoter fragments ranged from 1.5 kb to 2.14 kb. All constructs except for the construct containing 1.5-kb promoter fragment have significantly ($P < 0.001$) higher LUC activity than the *LUC* control vector lacking a promoter. **, $P < 0.001$; error bar, SD (n=3). (**C**) EMSA of five probes (Pr1, Pr2, Pr3, Pr4 and Pr5) from the *awn1* promoter incubated with nuclear extracts from panicles. The positions of the five probes (orange lines) are indicated along the *awn1* promoter from –2,200 bp to –1,500 bp. N.E., nuclear extracts from young sorghum panicles (1~3 cm). Arrows indicate the shifted bands induced by nuclear extracts. (**D**) GUS

staining of the glume from rice transgenic plants harbouring the *awn1pro:GUS* reporter. Root without GUS staining was used as control in rice transgenic plant.

Figure 5. AWN1 is a transcriptional repressor.

(**A**) Localization of GFP-AWN1 fusion protein in the nucleus of onion epidermal cells. (**B,C**) AWN1 acts as a transcriptional repressor, as determined by dual-luciferase transient activity assay. Luciferase activity was strongly repressed by the GAL4DB-VP16-AWN1 fusion protein ($P < 1.0 \times 10^{-4}$) relative to GAL4DB-VP16. **, $P < 1.0 \times 10^{-4}$; error bar, SD (n= 3). (**D**) Scanning electron microscope image of spikelets from NIL-SV and NIL-Tx623. White arrow, awn primordium; red star, pistil; orange star, stamen; gl, glume; scale bar, 100 μm. (**E**) DNA logo of the most enriched DNA sequence bound by AWN1, which consists of 11 bp, based on DAP-seq with *in vitro*-translated AWN1 and sorghum genomic DNA. The binding motif contains the 5-bp core sequence AC(A/T)GT. (**F**) Schematic representation of effectors and the four reporter constructs for dual-luciferase transient expression assays. The reporters placed *LUC* under the control of the *MADS3, MADS6, MADS7* or *LKS2* promoters. (**G**) LUC activity is significantly repressed by the overexpression of AWN1 for all four promoters. Student's T test; **, $P < 0.01$. The significant differences mean that AWN1 directly represses the transcriptions of these four genes. (**H**) Two AWN1-binding peaks are detected in the *LKS2* promoter in both replicates. Three 11-bp motifs are found under these two peaks. The numbers in the upper right corner represent the number of reads from input and peak DNA, in RPKM. The red, orange

and green bars under the two peaks indicate the three 11-bp motifs. (**I**) EMSAs of three *LKS2* promoter fragments (LP1, LP2 and LP3), each containing the 11-bp motif. EMSAs were performed in reactions with at least one of the following reagents: Halo tag, Halo-AWN1 protein, biotin-labelled probe, competitor without biotin label, and competitor with the core 5-bp mutated sequence in the 11-bp motif and without biotin label (Supplementary Figure 14). The specificity of binding was tested with competitors. Wild-type competitors dramatically decreased the binding to the probes, whereas mutated competitors did not affect binding. +, present; –, absent.

Figure 6. The downstream genes regulated by AWN1.

(**A**) A translocation of 5.4-kb fragment from chromosome 10 to 3 greatly enhances gene dosage of AWN1 and then represses the development of sorghum awn. Blue and green arrows, promoters; polygons, the *awn1* and *awn1-10* genes. (**B**) AWN1 is a transcriptional repressor that might bind directly to the intron of the awn-related gene *DL*and the promoters of the awn-related gene *LKS2* and the *MADS* genes, and directly or indirectly control the genes related to the auxin pathway, thereby repressing their transcription rates, which will prevent the elongation of the awn on the lemma. Arrows represent promoters, and boxes and thin bar signify exons and intron, respectively. Solid T bars indicate direct repression of genes.

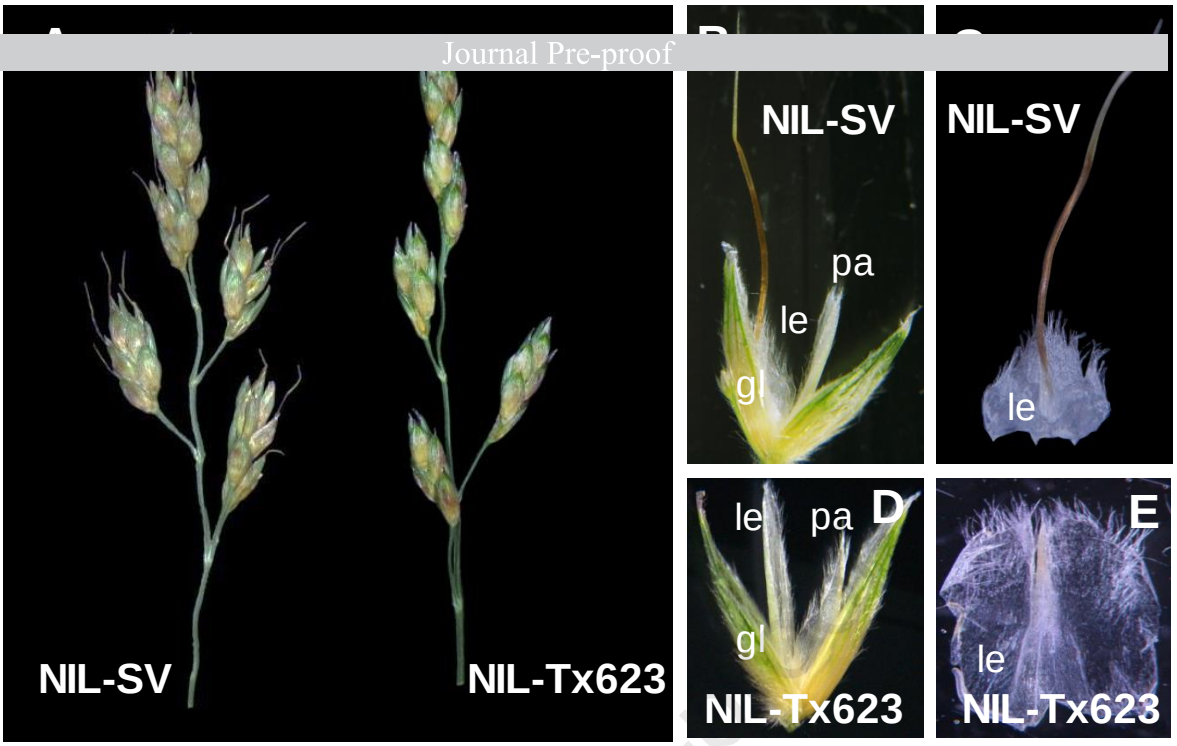


Figure 1

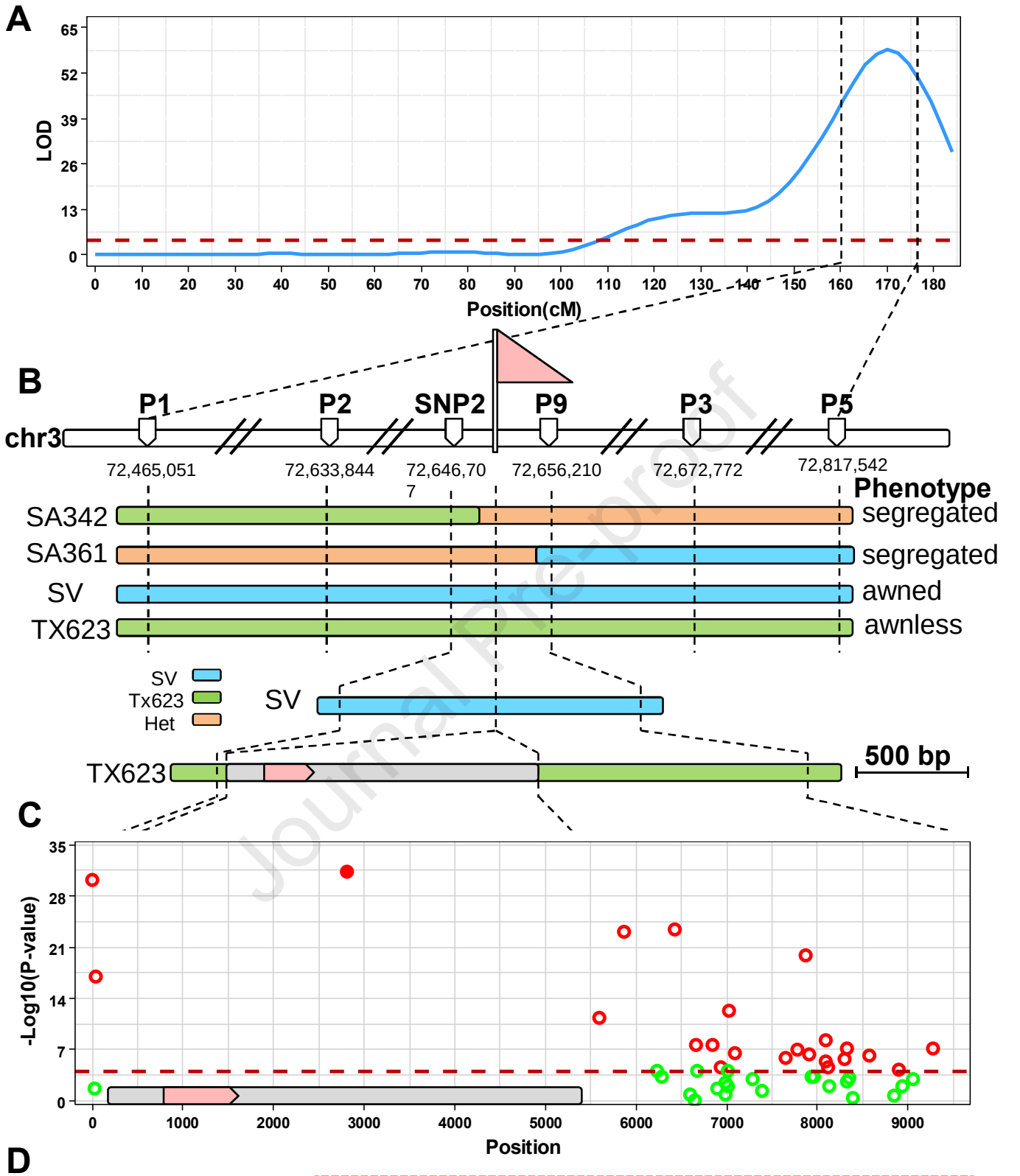


figure2

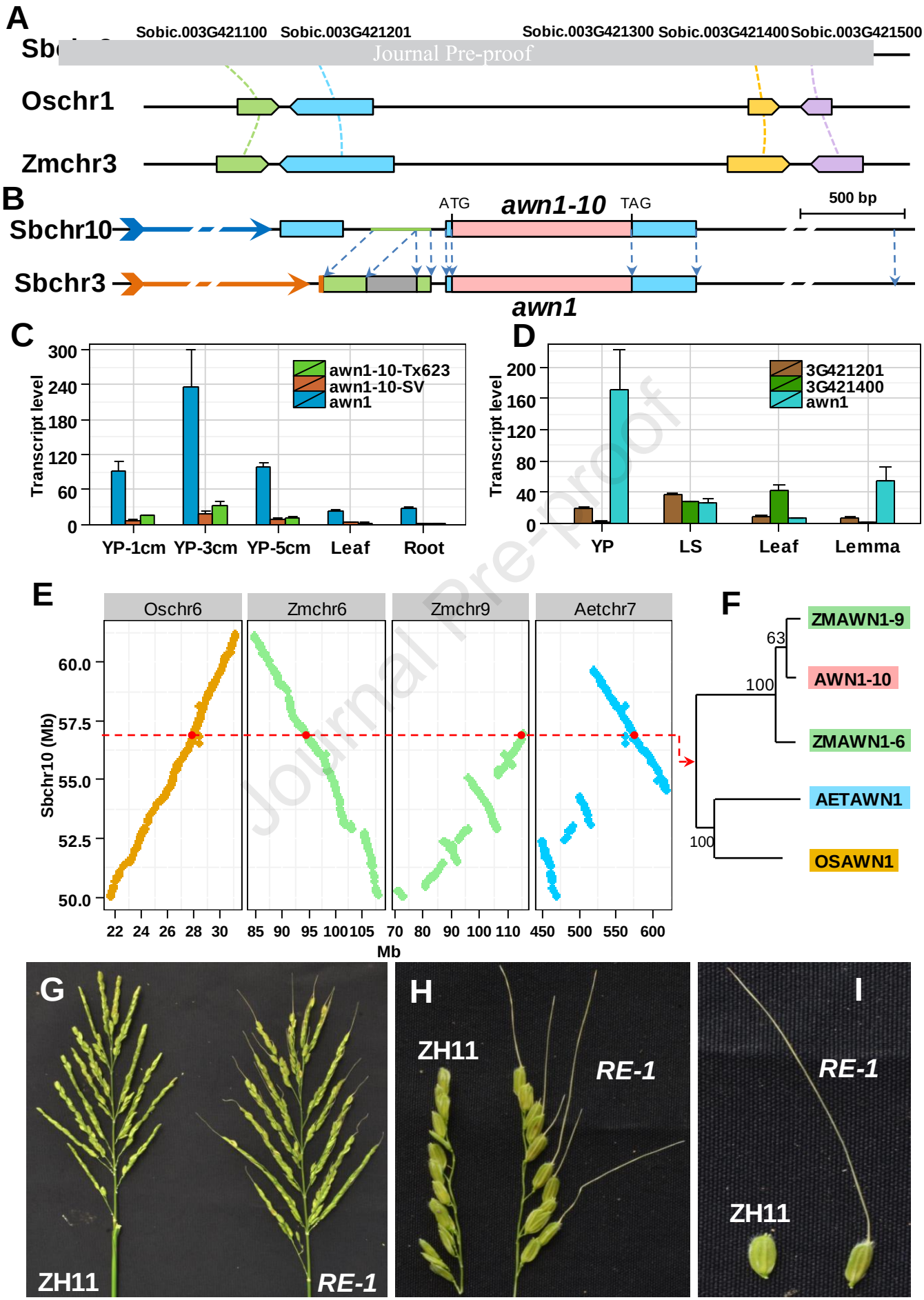
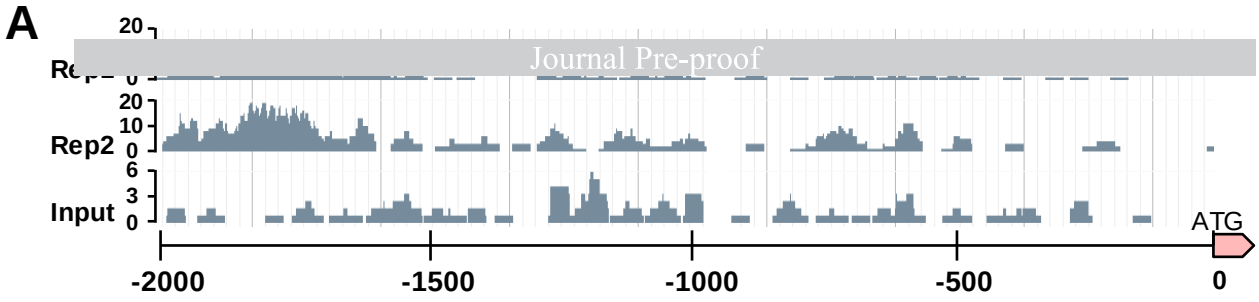
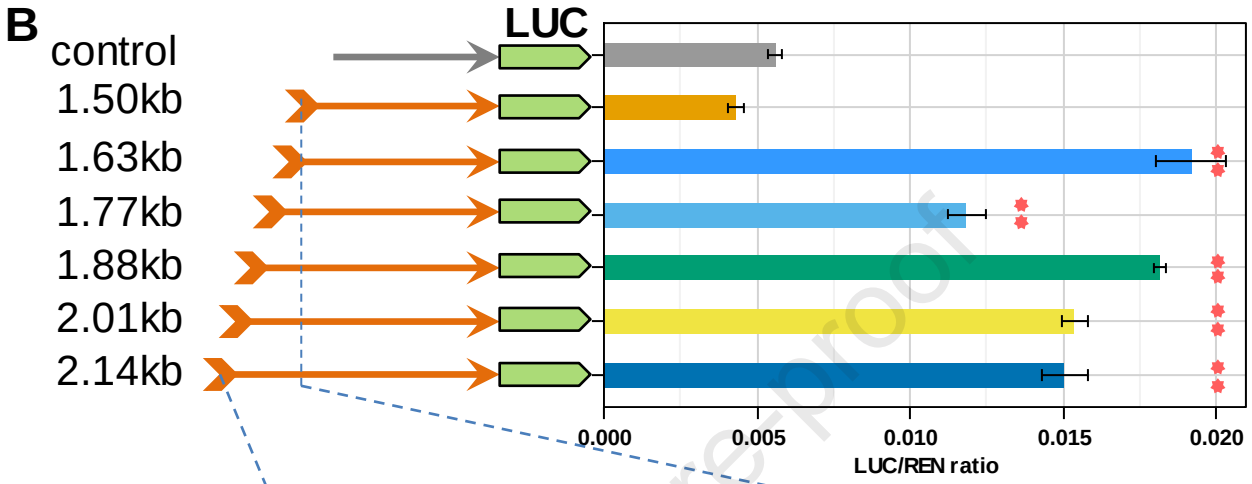
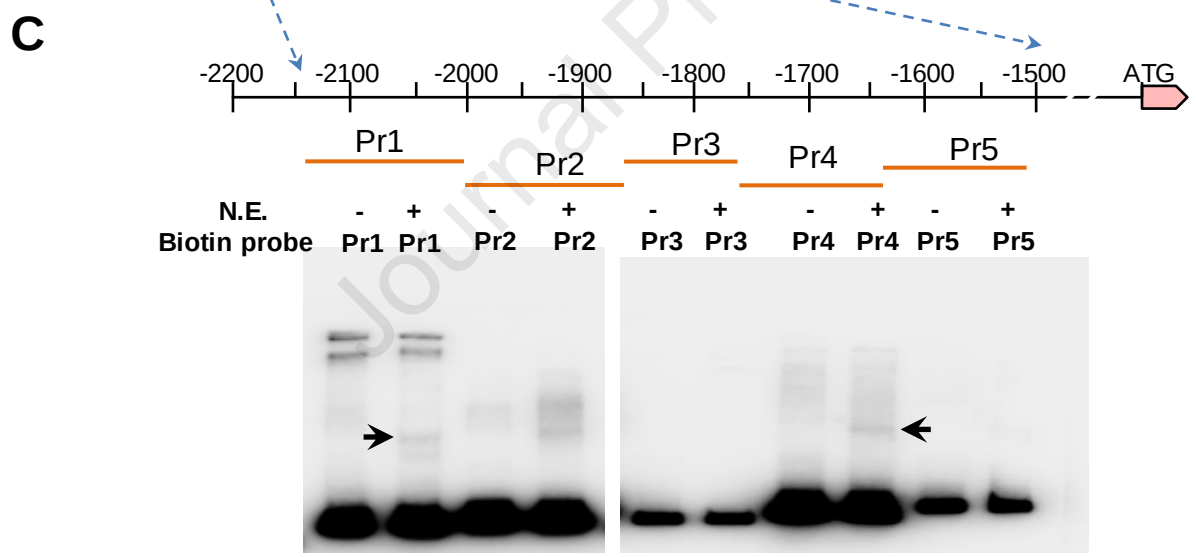


figure3







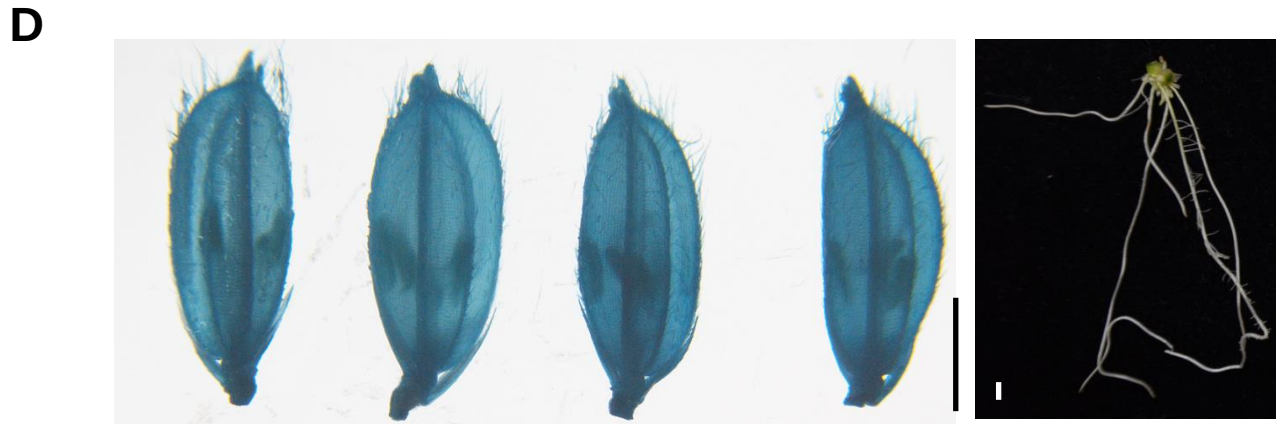


figure4

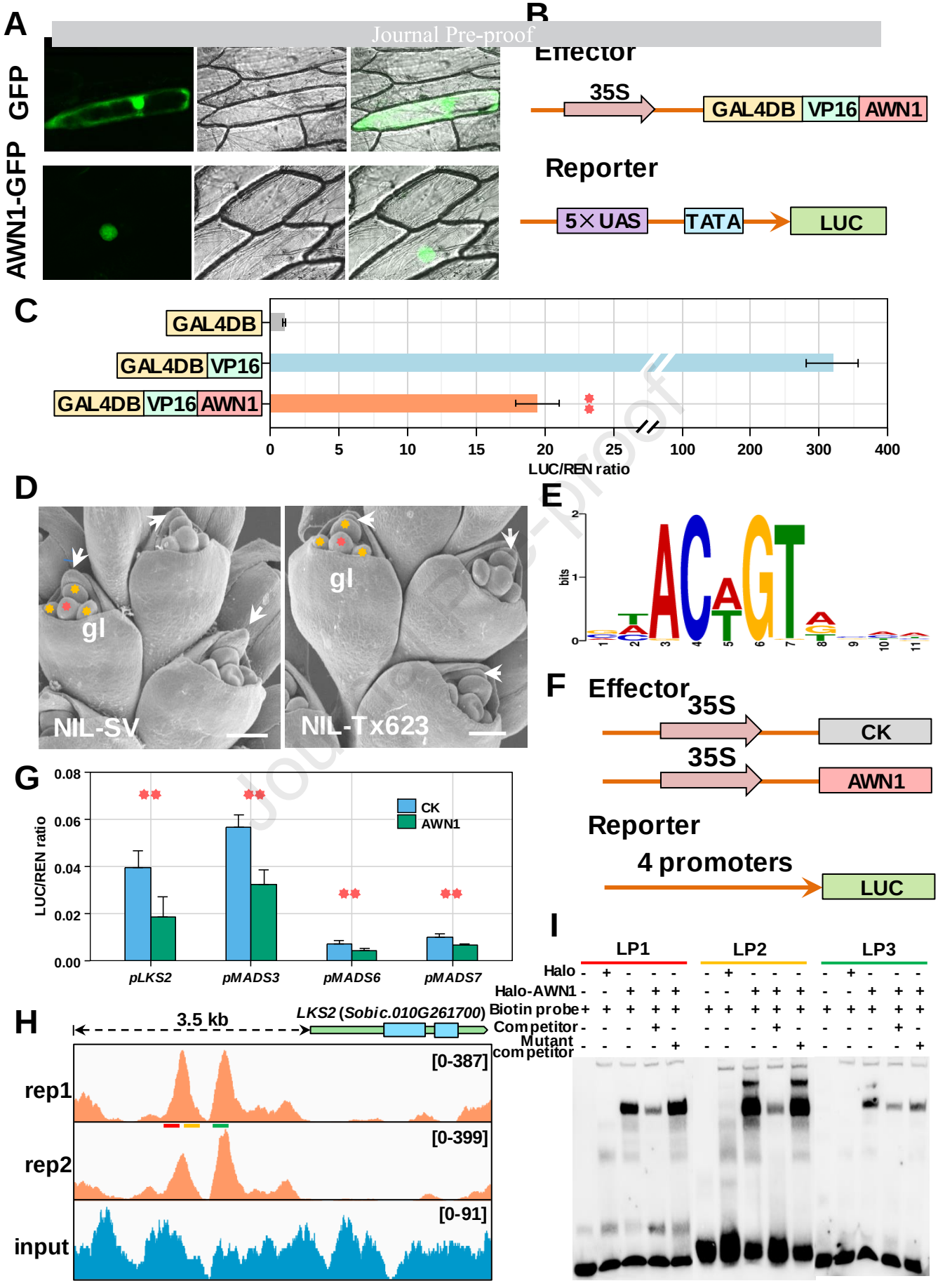


figure5

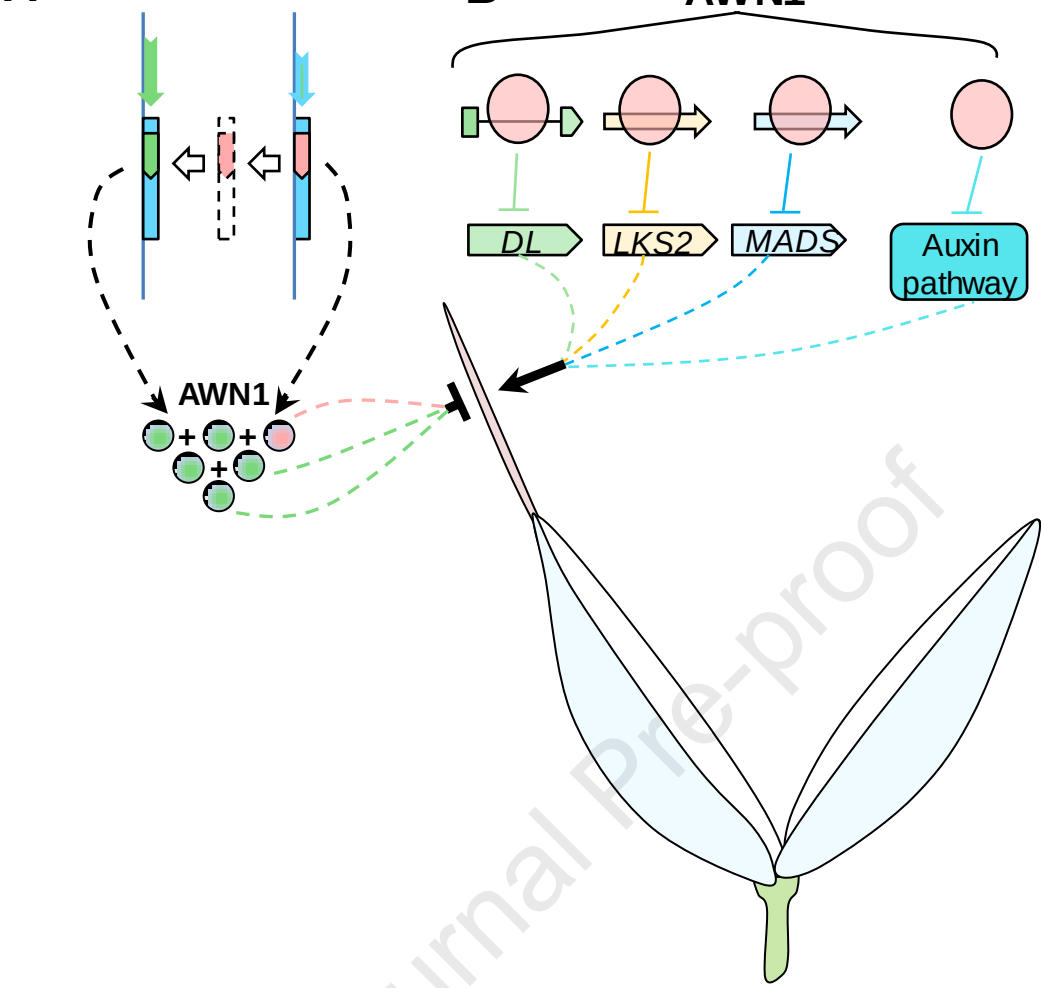


Figure 6