Conserved noncoding sequences correlate with distant gene contacts in *Arabidopsis* and *Brassica*

Lei Zhang[1], Jian Wu[1], Jianli Liang[1], Runmao Lin[1], Chao Sun[2], Qirui Dai[1], Lupeng Zhang[1], Huiling Guo[1], Ranze Zhao[1], and Xiaowu Wang[1*]

SUPPORTING INFORMATION

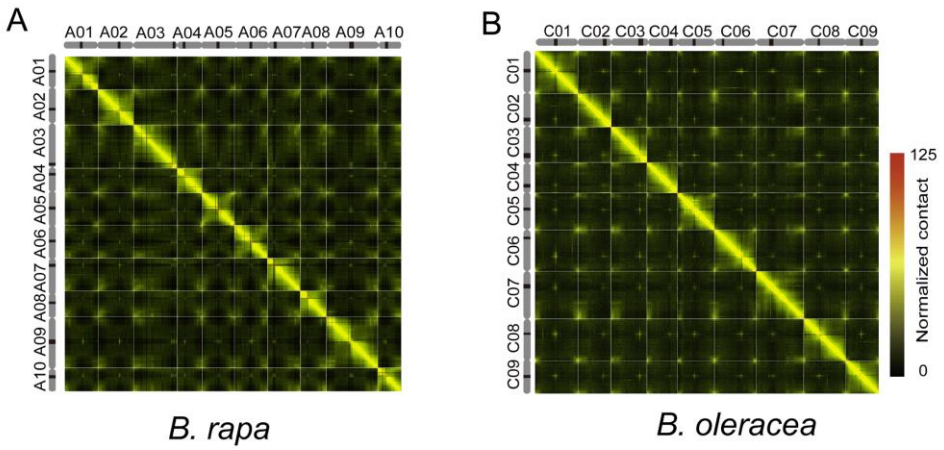

Figure S1. Whole-genome heatmaps of *B. rapa* and *B. oleracea* at 100-kb resolution

The chromosomes are shown from left to right and top to bottom. The chromosomes in the heatmaps are separated by grey bars. The thick black boxes in the top and left insets indicate the approximate centromere locations.

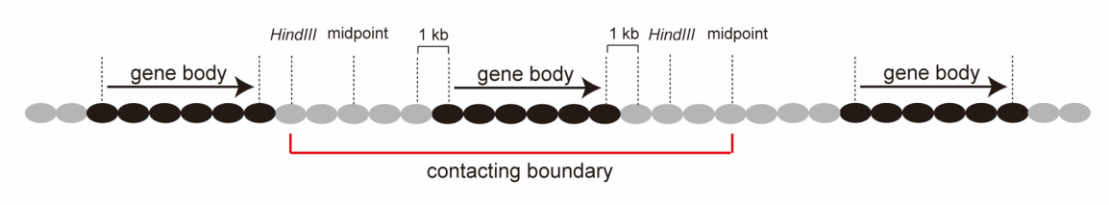

Figure S2. Identification of the contacting boundary of genes

For each gene, the bases at 1 kb upstream and downstream of the gene body were used as focal points. After comparing the nearest restriction site (*HindIII*) of the peripheral region of the focal point and the midpoint of the inter-genic region, the farther one was defined as the contacting boundary of genes. We should note that the contacting boundary of the gene may overlap with that of adjacent genes.

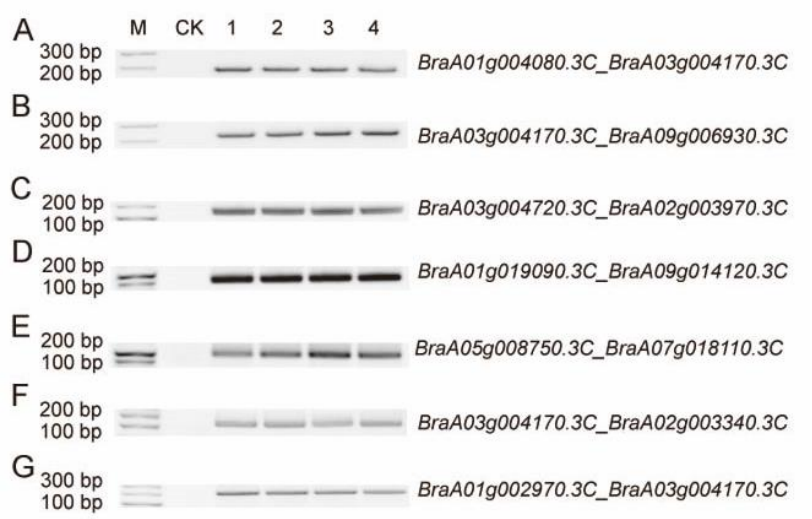

Figure S3. Results from the 3C-PCR experiment in *B. rapa*

The distant contacts of seven gene pairs were assayed using 3C-PCR in *B. rapa*. 'M' indicates the marker, and "CK" indicates the negative control of genome DNA. Numbers 1-4 indicate the four replicates of the 3C-PCR experiment. The PCR reaction used 20 cycles, and the primer sequences were listed in Table S2.

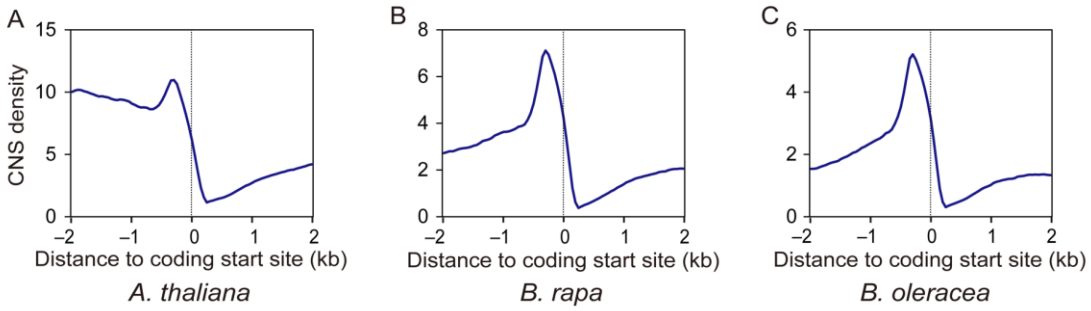

Figure S4. CNS distribution of genes in *A. thaliana*, *B. rapa* and *B. oleracea*

The CNSs are mainly located in the proximal promoter regions of genes. The CNS density was averaged in 500-bp sliding windows with 50-bp steps.

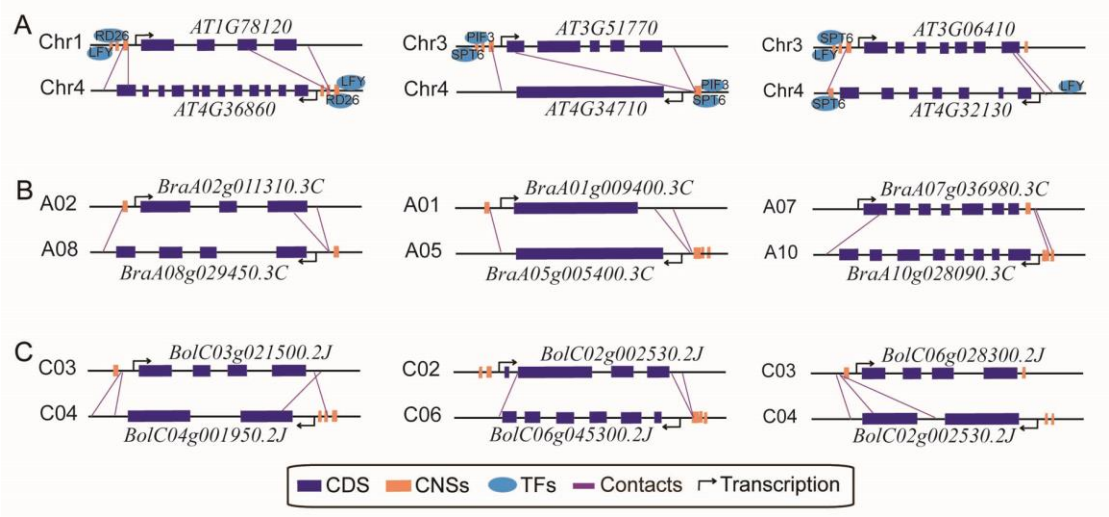

Figure S5. Examples of CNSs correlating with distant gene contacts in *A. thaliana*, *B. rapa* and *B. oleracea*

The three examples are shown in *A. thaliana* (**A**), three examples in *B. rapa* (**B**), and three examples in *B. oleracea* (**C**). The blue bar indicates the CDS of genes, the yellow bar indicates the CNSs, the light blue oval indicates the TFs, the purple lines indicate the distant contacts between different chromosomes, and the black arrow indicates the transcription of genes.

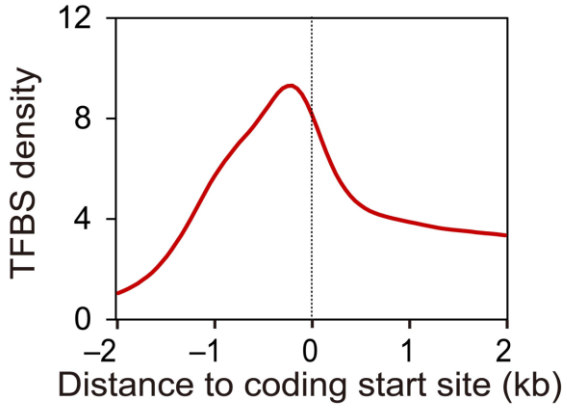

Figure S6. TFBS distribution in genes from *A. thaliana*

The TFBSs are mainly located around the proximal promoter of genes. The TFBS density was averaged in 500-bp sliding windows with 50-bp steps.

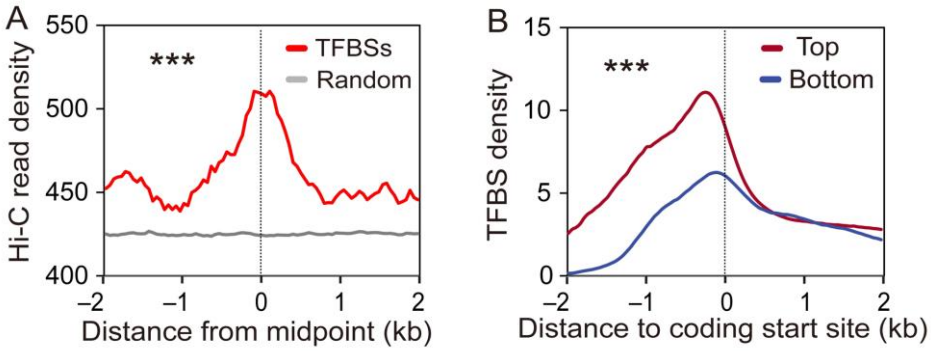

Figure S7. TFBS profiles in *A. thaliana*

(**A**) The normalized density of Hi-C read around TFBSs was higher than that of random TFBS-size regions. The midpoint of TFBSs was used as the focal point. The read density was averaged in 500-bp sliding windows with 50-bp steps. (**B**) Genes with the top 10% normalized number of distant interacting genes (DIGs) had a higher level of TFBS density than genes with the bottom 10% normalized number of DIGs. The TFBS density was averaged in 500-bp sliding windows with 50-bp steps. The *P* values were calculated using a two-sided *t*-test. *** $P < 0.001$.

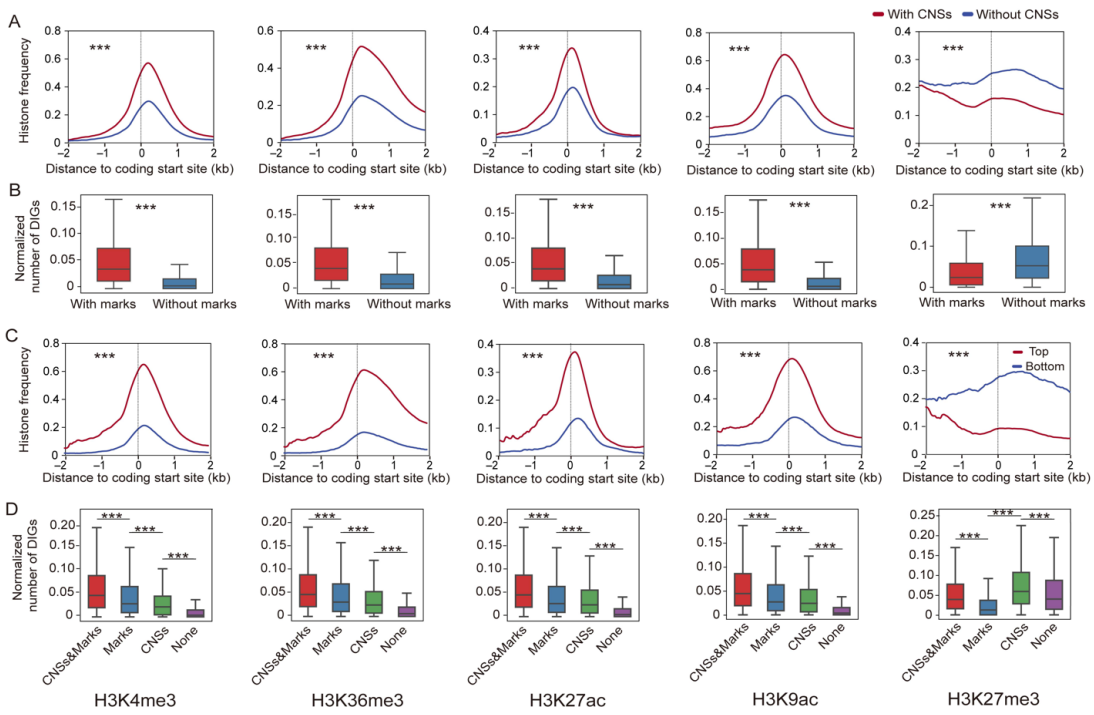

Figure S8. Histone modification profiles in *A. thaliana*

(**A**) Distant contacting genes with CNSs (red) had a significantly higher frequency of active histone marks (H3K4me3, H3K36me3, H3K27ac and H3K9ac) and lower frequency (H3K27me3) than those genes without CNSs (blue). (**B**) Gene with active histone marks (H3K4me3, H3K36me3, H3K27ac and H3K9ac) had a higher normalized number of distant interacting genes (DIGs) than genes with repressive histone marks (H3K27me3). (**C**) Genes with the top 10% normalized number of DIGs had a higher level of active histone frequency (H3K4me3, H3K36me3, H3K27ac and H3K9ac) and a lower level of repressive histone frequency (H3K27me3) than genes with the bottom 10% normalized number of DIGs. Histone frequency was averaged in 500-bp sliding windows with 50-bp steps. (**D**) Genes with CNSs and active histone marks had a higher normalized number of DIGs than those with only CNSs or active histone marks. CNSs&Marks, genes contain both CNSs and histone marks; Marks, genes contain only histone marks; CNSs, genes contain only CNSs; None, no gene contains CNSs or histone marks. The *P* values were calculated using the two-sided *t*-test. *** $P < 0.001$.

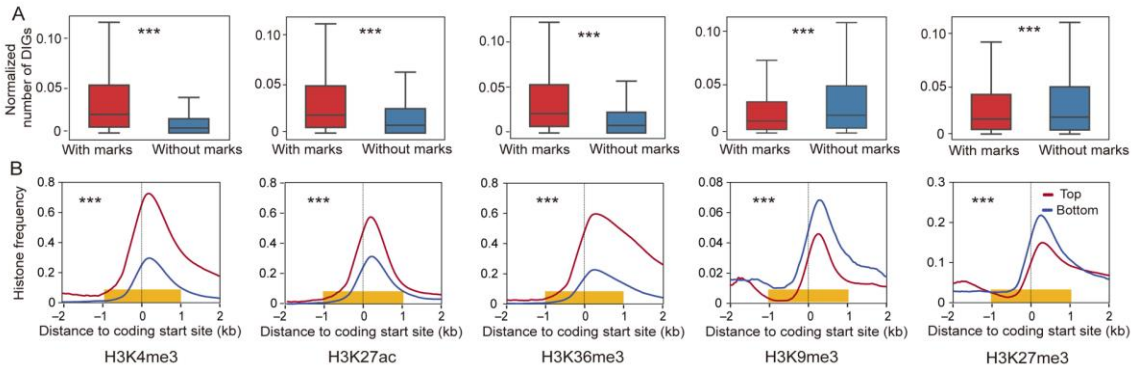

Figure S9. Histone modification profiles in *B. rapa*

(**A**) Genes with active histone marks (H3K4me3, H3K27ac and H3K36me3) had a higher normalized number of distant interacting genes (DIGs) than genes without active marks, while genes with repressive histone marks (H3K9me3 and H3K27me3) had a lower normalized number of DIGs than genes without repressive marks. (**B**) Genes with the top 10% normalized number of DIGs had a higher level of active histone frequency (H3K4me3, H3K27ac and H3K36me3) and a lower level of repressive histone frequency (H3K9me3 and H3K27me3) than genes with the bottom 10% normalized number of DIGs. Histone frequency was averaged in 500-bp sliding windows with 50-bp steps. The *P* values of the regions up- and down-stream 1 kb (yellow block) of the coding start site were calculated using the two-sided *t*-test. *** $P < 0.001$.

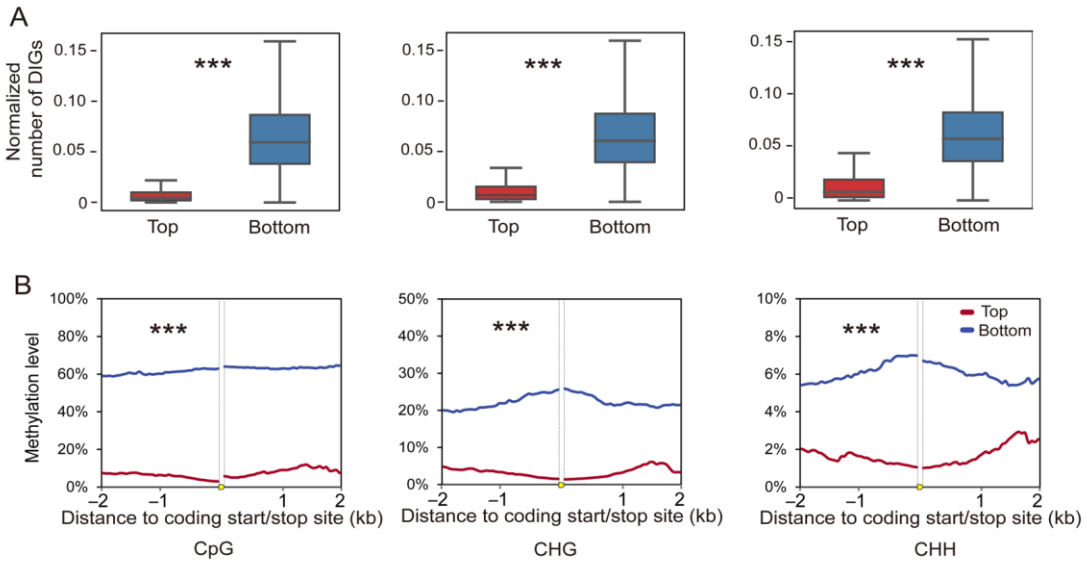

Figure S10. DNA methylation profiles in *A. thaliana*

(**A**) Gene with DNA methylation had a lower normalized number of distant interacting genes (DIGs) than genes without DNA methylation. (**B**) Genes with the top 10% normalized number of DIGs had a lower level of DNA methylation than genes with the bottom 10% normalized number of DIGs. The DNA methylation level was averaged in 500-bp sliding windows with 50-bp steps. The *P* values were calculated using a two-sided *t*-test. *** $P < 0.001$.

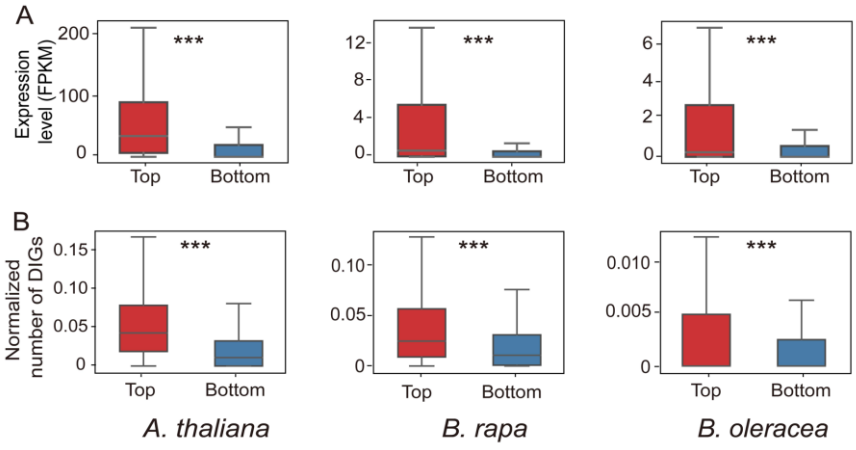

Figure S11. Expression levels of genes with and without distant contacts in *A. thaliana*, *B. rapa* and *B. oleracea*

(**A**) Genes with the top 10% normalized number of distant interacting genes (DIGs) had a higher expression level than genes with the bottom 10%

normalized number of DIGs. (**B**) Genes with the top 10% expression had a higher normalized number of DIGs than those with the bottom 10% expression. The *P* values were calculated using a two-sided *t*-test. *** *P* < 0.001.

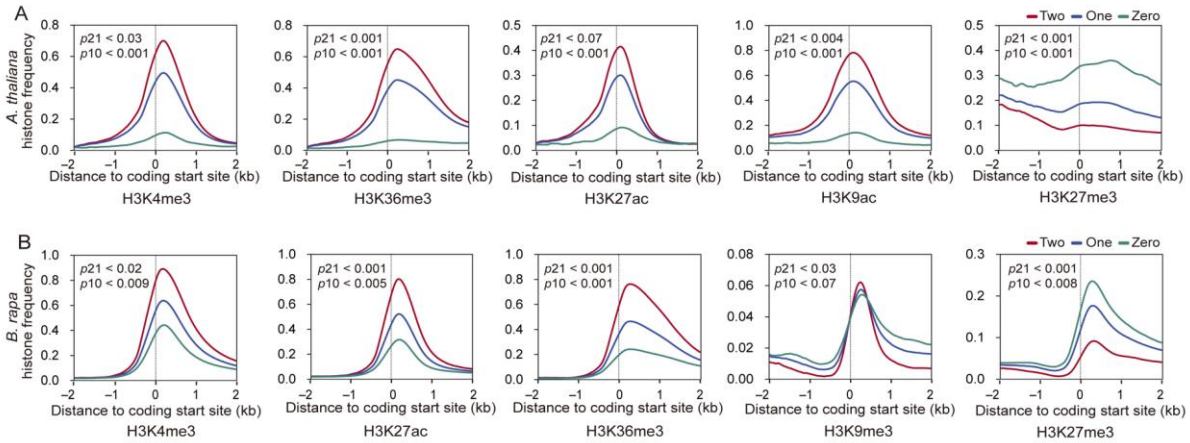

Figure S12. Histone frequency of contacting genes with and without co-expression in *A. thaliana*, *B. rapa* and *B. oleracea*

(A) Distant contacting gene pairs with co-expression had a higher level of active histone marks (H3K4me3, H3K36me3, H3K27ac and H3K9ac) and a lower level of repressive histone frequency (H3K27me3) than those without co-expression in *A. thaliana*. **(B**) Distant contacting gene pairs with co-expression had a higher level of active histone marks (H3K4me3, H3K27ac and H3K36me3) and a lower level of repressive histone marks (H3K9me3 and H3K27me3) than those without co-expression in *B. rapa*. We defined the genes with FPKM > 1 as expressed genes. Two, both genes are expressed; One, only one gene is expressed; Zero, no gene is expressed. "*p*21" indicates the *P* values between histone frequency of gene both expressed and only one expressed, and "*p*10" indicates the *P* values between histone frequency of gene only one expressed and no gene expressed. The *P* values were calculated using a two-sided *t*-test.

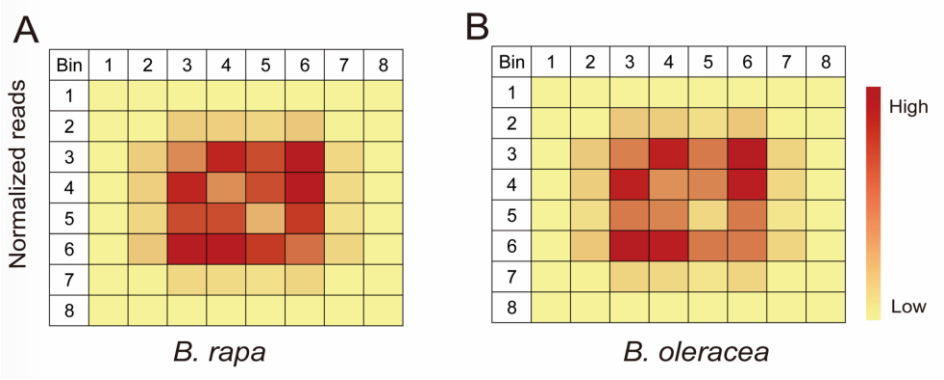

Figure S13. Heatmap of the normalized number of Hi-C reads between two different gene bins in *B. rapa* and *B. oleracea* using data from Xie et al. (2019) The numbers 1-8 on the top and left indicate Bin1-Bin8.

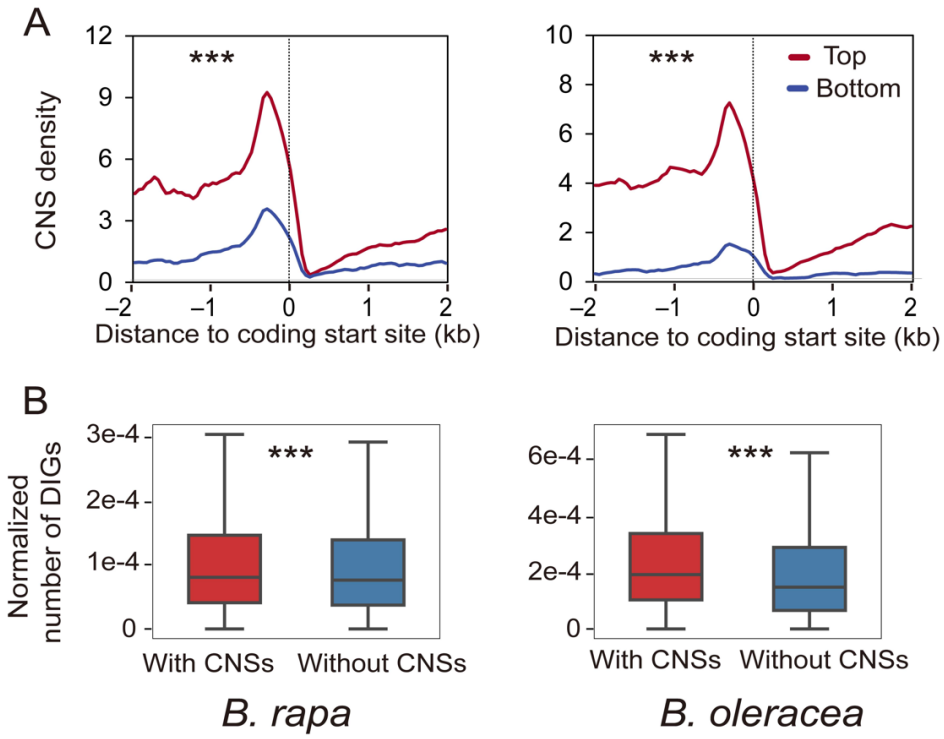

Figure S14. CNS profiles in *B. rapa* and *B. oleracea* using data from Xie et al. (2019)

(**A**) Genes with the top 10% normalized number of DIGs had a higher level of CNS density than genes with the bottom 10% normalized number of DIGs. CNS density was averaged in 500-bp sliding windows with 50-bp steps. The *P* values were calculated using a two-sided *t*-test. (**B**) Genes with CNSs had a

higher normalized number of DIGs than genes without CNSs. The *P* values were calculated using a two-sided *t*-test. *** $P < 0.001$.

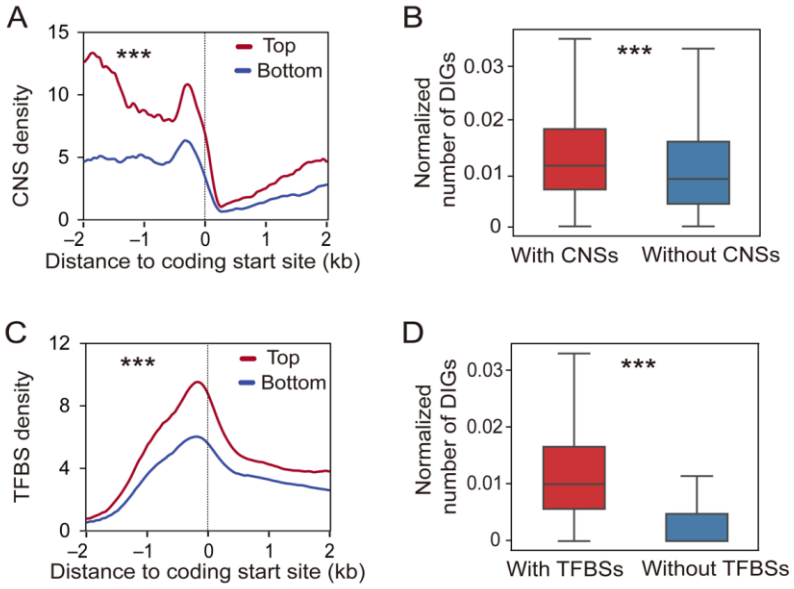

Figure S15. CNS and TFBS profiles of *A. thaliana* Hi-ChIP dataset

(**A**) Genes with the top 10% normalized number of DIGs had a higher level of CNS density than genes with the bottom 10% normalized number of DIGs. (**B**) Genes with CNSs had a higher normalized number of DIGs than genes without CNSs. (**C**) Genes with the top 10% normalized number of DIGs had a higher level of TFBS density than genes with the bottom 10% normalized number of DIGs. (**D**) Genes with TFBSs had a higher normalized number of DIGs than genes without TFBSs. The density was averaged in 500-bp sliding windows with 50-bp steps. The *P* values were calculated using a two-sided *t*-test. "***" indicates $P < 0.001$.

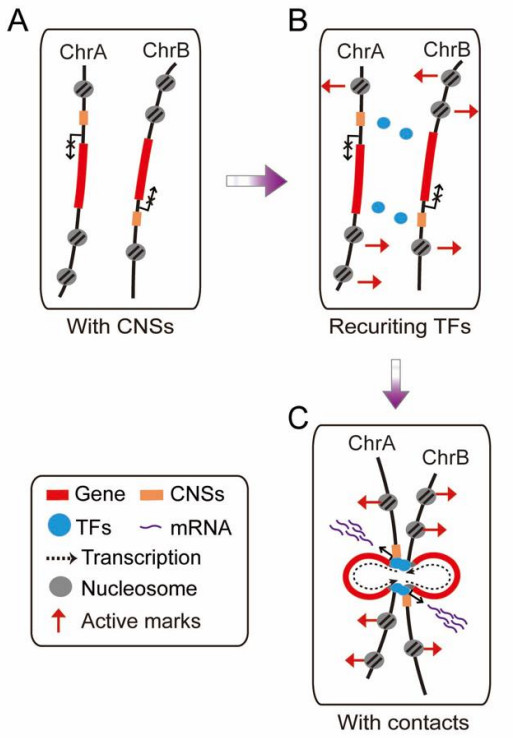

Figure S16. Schematic model of CNSs correlating with distant gene contacts As shown in the example, the potential mechanism of CNSs correlating with distant gene contacts is represented. Genes enriched with CNSs on promoter regions may recruit TFs or other regulatory proteins to involve the distant gene contacts, mainly focusing on the proximal promoter of one gene and the downstream region of the other gene. Furthermore, active marks are placed on the promoter regions, RNA polymerase II may be recruited, and gene transcription begins. We noted that other factors could also be associated with chromatin contacts; thus, other possible mechanisms of distant gene contacts should be envisioned.

Table S1. Summary of Hi-C data in *A. thaliana*, *B. rapa*, and *B. oleracea*

Table S2. Primer sequences for the 3C-PCR experiment in *B. rapa*