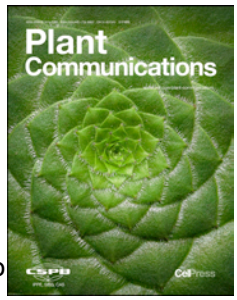
Journal Pre-proof



DeepCBA: a deep learning framework for gene expression prediction in maize based on DNA sequence and chromatin interaction

Zhenye Wang, Yong Peng, Jie Li, Jiying Li, Hao Yuan, Shangpo Yang, Xinru Ding, Ao Xie, Jiangling Zhang, Shouzhe Wang, Keqin Li, Jiaqi Shi, Guangjie Xing, Weihan Shi, Jianbing Yan, Jianxiao Liu

PII: S2590-3462(24)00293-1

DOI: https://doi.org/10.1016/j.xplc.2024.100985

Reference: XPLC 100985

To appear in: *PLANT COMMUNICATIONS*

Received Date: 28 March 2024

Revised Date: 25 May 2024

Accepted Date: 5 June 2024

Please cite this article as: Wang, Z., Peng, Y., Li, J., Li, J., Yuan, H., Yang, S., Ding, X., Xie, A., Zhang, J., Wang, S., Li, K., Shi, J., Xing, G., Shi, W., Yan, J., Liu, J., DeepCBA: a deep learning framework for gene expression prediction in maize based on DNA sequence and chromatin interaction, *PLANT COMMUNICATIONS* (2024), doi: https://doi.org/10.1016/j.xplc.2024.100985.

DeepCBA: a deep learning framework for gene expression prediction in maize based on DNA sequence and chromatin interaction

Zhenye Wang[1,2,3,#], Yong Peng[1,4,#], Jie Li[1,2,3,#], Jiying Li[5], Hao Yuan[1,2,3], Shangpo Yang[1,2,3], Xinru Ding[1,2,3], Ao Xie[1,2,3], Jiangling Zhang[3], Shouzhe Wang[1,4], Keqin Li[1,2,3], Jiaqi Shi[3], Guangjie Xing[3], Weihan Shi[3], Jianbing Yan[1,4], Jianxiao Liu[1,2,3,4,*]

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

[2]Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan 430070, China

[3]College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

[4]Hubei Hongshan Laboratory, Wuhan 430070, China

[5]Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

#These authors contributed equally to this article

[*]Correspondence: Jianxiao Liu (liujianxiao@mail.hzau.edu.cn)

Short Summary

This study develops a highly accurate deep learning-based gene expression prediction model (DeepCBA) based on maize chromatin interaction data. DeepCBA exhibits higher accuracy in expression classification and expression value prediction, and identifies some important motifs involving maize gene promoter proximal interaction (PPI) and proximal-distal interaction (PDI). Moreover, the promoter editing and verification of two reported genes (*ZmCLE7*, *ZmVTE4*) demonstrated new insights of DeepCBA in precise designing of gene expression and even future intelligent breeding.

Abstract

Chromatin interactions create spatial proximity between distal regulatory elements and target genes in the genome, which has an important impact on gene expression, transcriptional regulation, and phenotypic traits. To date, several methods have been developed for predicting gene expression. However, existing methods do not take into consideration the impact of chromatin interactions on target gene expression, thus potentially reduces the accuracy of gene expression prediction and mining of important regulatory elements. In this study, a highly accurate deep learning-based gene expression prediction model (DeepCBA) based on maize

chromatin interaction data was developed. Compared with existing models, DeepCBA exhibits higher accuracy in expression classification and expression value prediction. The average Pearson correlation coefficients (PCC) for predicting gene expression using gene promoter proximal interactions, proximal-distal interactions, and proximal and distal interactions were 0.818, 0.625, and 0.929, respectively, representing an increase of 0.357, 0.16, and 0.469 over the PCC of traditional methods that only use gene proximal sequences. Some important motifs were identified through DeepCBA and were found to be enriched in open chromatin regions and expression quantitative trait loci (eQTL) and have the molecular characteristic of tissue specificity. Importantly, the experimental results of maize flowering-related gene *ZmRap2.7* and tillering-related gene *ZmTb1* demonstrate the feasibility of DeepCBA in exploring regulatory elements that affect gene expression. Moreover, the promoter editing and verification of two reported genes (*ZmCLE7*, *ZmVTE4*) demonstrated new insights of DeepCBA in precise designing of gene expression and even future intelligent breeding. DeepCBA is available at http://www.deepcba.com/ or http://124.220.197.196/.

Keywords: Maize, Gene expression prediction, Chromatin interactions, Deep learning, Promoter editing, Regulatory elements and motifs.

Introduction

Gene expression plays an important regulatory role in the development, growth and reproduction of organisms, and a specific amount of gene product is produced in a particular spatiotemporal manner (Zrimec *et al*., 2020). Predicting gene expression helps to better understand the mechanism and the impact of sequence variation on transcriptional regulation, and is complementary to population-based association analysis methods (Avsec, Ž *et al*., 2021). Therefore, developing accurate gene expression prediction models and mining important variation sites would help reveal the genetic basis of complex traits.

Gene expression is regulated by key genomic regulatory elements in DNA sequences, including promoters, enhancers, silencers, insulators. *etc*. Similarly, epigenetic features such as histone modifications, DNA methylation, transcription factors (Liu *et al*., 2021), and DNase I hypersensitive sites play significant roles in the expression of target genes. Early studies mainly used traditional machine learning methods to predict gene expression based on various types of data (Beer M. A., 2004; Cheng *et al*., 2011; Dong *et al*., 2012; Tasaki *et al*., 2020), with room of improvement on the accuracy. Deep learning methods have good fitting capability in processing

complex nonlinear data and can effectively extract complex features. To date, deep learning methods have been widely used in many fields, such as in image processing and natural language processing. In recent years, researchers have also used deep learning for genome sequence analysis tasks, including the prediction of sequence functions (Zhou *et al*., 2015), transcription factor-binding sites (Zhao *et al*., 2021), chromatin interactions, methylation status, *etc* (Karlić *et al*., 2010; Schmidt *et al*., 2017). Similarly, researchers have developed deep learning models to predict gene expression through analyzing genome sequences, such as Basenji (Kelley *et al*., 2018), ExPecto (Zhou *et al*., 2018), Enformer (Avsec, Ž. *et al*., 2021), and Chromoformer (Lee *et al*., 2022), *etc*. Specifically, ExPecto uses convolutional neural network (CNN) to integrate 40 kb sequences upstream and downstream of gene promoters and uses spatial feature transformation and a linear regression model to predict human gene expression. Enformer uses the Transformer model to analyze the impact of distal elements as distant as 100 kb on gene expression. Studies show that chromatin interactions create spatial proximity between distal regulatory elements and target genes on the genome, which has an important impact on gene expression, transcriptional regulation and phenotypic traits (Schoenfelder *et al*., 2019; Peng *et al*., 2019). However, existing deep learning methods do not consider the impact of chromatin interactions (including promoter proximal regions and distal elements) on target gene expression, resulting in the capture of incomplete sequence information and thus affecting prediction accuracy. Although Enformer can predict the impact of distal elements within 100 kb on target gene expression, this method cannot capture the impact of regulatory elements at the genome-wide level. In addition, the above methods have mainly been used in humans and mice, and there have been few studies in plants.

Maize is a crop with one of the largest cultivation areas in the world. It is not only the most important food crop, but has also come to be used in industry and agriculture. Based on the B73 reference genome sequence, chromatin interaction and gene expression data from multiple tissues, our study has three major contributions:

(1) We developed an accurate maize gene expression prediction model named DeepCBA (Deep neural networks of CNN module, BiLSTM and Attention mechanism) based on chromatin interactions data. DeepCBA has higher AUC and PCC values in gene expression classification and regression prediction tasks, and it improves the PCC of predicting gene expression values by 46.3% compared to the traditional method. Experimental results show the impact of gene promoter proximal interactions, proximal-distal interactions, and proximal and distal interactions on gene expression are 0.801, 0.621, and 0.923, respectively.

(2) The DeepCBA model identified 400-800 motifs that affect gene expression through chromatin interactions in maize shoot and ear tissues. These motifs have obvious tissue specificity and are significantly enriched in expression quantitative trait loci (eQTL) and open chromatin regions. The identified motifs are clustered into 6 groups of core sequences in ear and shoot, and the distribution pattern of the motifs can be divided into 5 and 4 categories in PPI and PDI mode, respectively.

(3) Experimental results of detecting the regulatory elements of the reported genes (*ZmRap2.7*, *ZmTb1*), saturating the promoter regions of *ZmCLE7* and *ZmVTE4*, constructing cross-tissue and cross-genotype transfer learning models reveal the feasibility of DeepCBA in mining distal regulatory elements, precise designing of gene expression and even future intelligent breeding.

Results

Predicting gene expression based on DeepCBA model. The experimental data used in this study are the published maize chromatin interaction and expression data (Li *et al*., 2019) of three tissues (Shoot and Ear). According to the type of elements that interact with genes (Materials and methods), the chromatin interaction data is divided into two categories: the promoter proximal region interaction (PPI) and the promoter-distal region interaction (PDI). The interaction data involves two tissues: Shoot (Li *et al*., 2019) and Ear (Li *et al*., 2019). The average interaction number of PPI in Shoot and Ear is 50198, and the average interaction number of PDI in above two tissues (Shoot and Ear) is 11198. The number of genes involved in the above interaction datasets is 23707. To balance the number of genes in each category, we classify genes into Unexpressed/Expressed/Highly-expressed according to FPKM∈[0-0.1), FPKM∈[0.1-15), FPKM∈[15-max]. The gene expression values of different tissues were between 0 and 500, accounting for 99.6% of all genes in maize genome. We defined the DNA sequence for a specific gene as an inclusion of 1 kb upstream and 0.5 kb downstream of the transcription start site (TSS), and 0.5 kb upstream and 1 kb downstream of the transcription termination site (TTS) (Figure 1A) (Washburn, J. D. *et al*., 2019).

In this study, a high-precision maize gene expression prediction model, called DeepCBA, was developed to make predictions based on chromatin interactions (Figure 1B). DeepCBA includes three modules, and the convolution neural network (CNN) is used to extract features of the encoded chromatin sequence and reduce the dimensionality. DNA sequences are usually double-stranded, with the two strands connected by hydrogen bonds between bases, known as

reverse complements. The bidirectional long short-term memory network (BiLSTM) can capture bidirectional and spatial information, and it has ability to capture the dependencies between features by accessing long-range context. In this study, we use the BiLSTM to capture distal interactions among chromatin sequence features. The self-attention mechanism is used to capture the contribution of key features for the model.

To evaluate the reliability of DeepCBA, the accuracy of the gene expression classification prediction of the following models was compared: (1) CNN_No_PPI: a CNN model using the 3 kb sequences upstream and downstream of the gene; (2) CNN_PPI: a CNN model using the 6 kb sequence of promoter proximal region interaction (PPI); (3) DeepCBA_PPI: the DeepCBA model using the 6 kb sequence of promoter proximal region interaction (PPI). DeepCBA has excellent model generalization ability in predicting gene expression classification. The results show that the gene expression classification prediction accuracy of DeepCBA using PPI data, or in PPI mode, is significantly better than that of CNN_No_PPI and CNN_PPI (Figure S1). The PCC of predicted gene expression for DeepCBA_PPI in two datasets is 0.954, 0.967, and the PCC of predicted gene expression for CNN_No_PPI in two datasets is 0.309, 0.52 (Figure 1C). To explore the impact of the order of interacting genes on the results of target gene expression, we consider gene order in PPI mode (data augmentation) and the number of chromatin interactions is twice that of the original PPI sequences (no data augmentation). The experimental results show that DeepCBA can achieve better prediction results after the regulatory order between genes in PPI mode is taken into consideration (Figure 1C). The above results show that DeepCBA has higher accuracy than traditional methods in predicting gene expression.

DeepCBA identifies a dynamic range of gene expression values in different interaction modes. When only using PDI sequences to predict gene expression, the PCC of DeepCBA's prediction results of the two tissues are 0.6214 and 0.6278, respectively. While, only using PPI sequences to predict gene expression, the PCC of DeepCBA's prediction results of the two tissues are 0.8060 and 0.8290, respectively. When considering the interaction sequences of PPI and PDI at the same time, the PCC of DeepCBA in Shoot and Ear are 0.9314 and 0.9266, respectively. After considering the order of genes in PPI mode (data augmentation), the PCC of the two tissues are 0.9539 and 0.9672 (Figure 2). The above results show that the PCC of DeepCBA increases by 35.7%, 20%, 50.2% in the case of PDI, PPI, PDI+PPI, respectively. In addition, we conducted expression classification and regression predictions for the other three datasets: Shoot (Peng *et al*., 2019), Ear (Peng *et al*., 2019) and Tassel (Sun *et al*.,

2020) (Materials and methods). Compared to other methods, the regression prediction accuracy of DeepCBA has increased by 23.16%, 20.54%, 19.73% in the three datasets (Figure S2). For the gene expression classification prediction, DeepCBA has better average AUC than the other two methods (Figure S3). The above results reveal the significant role of chromatin interactive regulatory sequences info in predicting gene expression and quantifying the effects of different interactive elements on expression regulation. As the dataset size of chromatin interactions increases, the better predictions of gene expression are.

Interestingly, there are some outliers in the above DeepCBA gene expression prediction results. These outliers all show a tendency of higher true expression but lower predicted values. Taking the expression prediction of PPI sequences as an example, we compare the predicted gene expression (*Pre_exp*) and the real gene expression value (*Real_exp*). When *Pre_exp* < 0.5**Real_exp* or *Pre_exp* > 1.5**Real_exp*, we regard the gene as a candidate gene with a large prediction deviation. Then, the candidate genes are sorted according to the deviation, and it is found that approximately 40% of these genes participate in both PPI and PDI interactions (Table S5). Moreover, the genes with biased predictions have obvious tissue specificity (Figure S4A). The above results indicate that an insufficient amount of chromatin interactions dataset maybe the factor causing the deviation in gene expression prediction, which reveals the complex regulatory network in the process of gene expression.

Genome wide mining of PPI-mediated motifs affecting gene expression. To identify important motifs affecting gene expression, saliency map (Chu *et al*., 2011) was used to calculate the gradient of chromatin sequence. The results show that the sequence region around 750 bp upstream of gene TSS and 250 bp downstream of gene TSS have significant contribution to gene expression prediction (Figure 3A), which is consistent with previous results (Washburn *et al*., 2019). Furthermore, TF-MoDIsco (Avanti *et al*., 2018) was used to mine motifs (Figure S4B, C), and 812 and 897 motifs were identified in Ear and Shoot, respectively (Figure 3B). To validate the reliability of these predicted motifs, we use PlantTFDB database (Tian *et al*., 2020; Jin *et al*., 2017; Jin *et al*., 2015; Jin *et al*., 2014) as the ground truth and query these motifs against the database. The results show that there are 87.6% and 41.14% motifs in Shoot and Ear are matched with the verified conserved domains of 653 higher plants (E-value < 0.5) (Figure S8). Secondly, the motifs identified in Ear and Shoot were compared with the motifs that are bound by 104 maize transcription factors (Tu *et al*., 2020), and results showed that 40.4% and 39.4% of 104 TFs can be matched (Supplementary Data 1 and Supplementary Data 2), respectively. Interestingly, four motifs (ATTTAA, CAGGAA, TAATAT and CACAGA) were validated to be

involved in gene expression regulation (Liu *et al*., 2021; Fu *et al*., 2013; Hufford, M. B *et al*., 2021) (Figure S12A, B). The above results show that the motif sequences identified using DeepCBA have biological significance. The detected motifs are positionally anchored in the 6 kb (3 kb per gene) sequence of the PPI sequence, and the distribution pattern of the motifs can be divided into 5 categories: (1) highly enriched near 250 bp downstream of the TSS; (2) only significantly enriched at specific sites, such as the TSS and TTS; (3) slightly enriched near 250 bp downstream of the TSS; (4) slightly enriched in the TSS and highly enriched in the TTS; and (5) distributed evenly throughout the entire sequence (Figure 3C). By comparing the number of motifs with different patterns in different tissues, it is found that the proportion of motifs in the first category is the highest, which is consistent with the gradient results in sequences (Washburn *et al*., 2019). The number of overlapping motifs in the Ear and Shoot dataset is 95. The distribution of these 95 motifs in the five enrichment patterns is basically consistent in Ear and Shoot (Table S6). To further verify that the motifs identified by DeepCBA affect gene expression, we combine the motifs in pairs to form a motif composition (*Real motif*), and the control is the motif composition formed by random sequence combination (*Random motif*). Then, the two kinds of motif compositions are inserted into 3 kb sequences to perform gene expression prediction (using *N* coding mode in the one-hot coding). The results show that the impact on the expression of the composited motifs identified by DeepCBA is significantly higher than that of the control group (Figure S7B) (p value = 2.06e-13). To further detect the impact of the specific motif sequence changes on expression, we randomly selected two motifs (CCGCCG and CTCTCTC), mutated the above two motifs in the test dataset and predicted the expression of the corresponding genes (Figure S7A). The numbers of genes in Ear and Shoot are 2608 and 4176, respectively. According to the standard that the expression value changes by more than 50%, the results in the Ear dataset show that the proportions of two motif mutations that affect gene expression are as high as 96.24% and 96.43%, respectively. In the Shoot dataset, the proportions of the two motif mutations that affect gene expression are as high as 92.12% and 92.21%, respectively (Figure S7C-F). The above results show that the discovered motifs play important functions in gene expression and regulation.

To further validate whether the motifs identified by DeepCBA have sequence similarities in different categories, MetaLogo (Chen *et al*., 2022) is used to cluster the motifs into core sequences. Six groups of motifs with similar core sequences are identified in Ear and Shoot (Figure 3D, E and Figure S9). Based on the clustering results, we compare the expression of genes containing different numbers of motifs in the gene sequences. Gene expression shows an

upward trend as the number of motif compositions increases (Figure 3F, G), which imply that the expression of genes is the result of the joint regulation of different factors.

The motifs identified by DeepCBA reveal the regulation of gene expression. To analyze the apparent characteristics and biological functions of the motifs identified by DeepCBA in PPI mode, the identified motifs in Ear are positionally anchored in the original gene sequences. First, we determine the physical location of each motif in the chromosome (Supplementary Data 5) and matched motifs with published eQTLs (Tian *et al*., 2023). The following two processing modes are used as controls: (1) removing motif sequences from PPI sequences and selecting sequences of equal length from the remaining PPI sequences; (2) removing PPI sequences from the whole DNA genome and selecting sequences from the remaining genome sequences. Compared with the control, the motifs identified by DeepCBA are significantly enriched at eQTLs (Figure 4A, B) (****$P < 0.0001$, *t* test). We conducted 100 repeated experiments in the above two controls to reduce the error introduced by randomized experiments. Moreover, we match these motifs with the open chromatin regions identified in 26 lines of the NAM population (Woodhouse *et al*., 2021). The results show that the motifs identified by DeepCBA are significantly enriched in the open chromatin regions identified in the NAM population (Figure 4C, D) (****$P < 0.0001$, *t* test). Similarly, the physical locations of important motifs identified in Shoot are significantly enriched with eQTLs and open chromatin regions (Figure S10) (****$P < 0.00001$, *t* test). Through matching with 104 transcription factors (Tu *et al*., 2020) and eQTLs, there is an identified motif of CATGCA in the sequence of gene *Zm00001d042609*. The motif and the downstream gene *Zm00001d042600* can be bound by the transcription factor nactf109 simultaneously (Supplementary Data 9). The variation in CATGCA in the maize association mapping panel (AMP) leads to expression changes of *Zm00001d042600*, thus affecting the drought resistance phenotype at the seedling stage (Figure 4E). Meanwhile, CATGCA (RY-motif) is a highly conserved motif that exists in the promoters of many seed-specific genes and plays an important role in seed development (Mönke *et al*., 2004). The transcription factors of *ABI3* and *FUS3* play important regulatory roles in the development and maturation process of *Arabidopsis* seeds, and they can combine with RY-motif to regulate the ABA-mediated endosperm maturation process (Guerriero *et al*., 2009). The homologous gene of *ABI* in maize, *Vp1* is also a key factor in regulating seed maturation. In addition, *Vp1* is also expressed in the phloem cells of vegetative tissues under drought stress (Cao *et al*., 2007). *ZmABI19*, a TF containing the B3 domain, can also bind to the RY-motif upstream of the grain filling-specific TF gene *Opaque2* (*O2*) promoter to perform

transactivation to regulate endosperm development in maize. In addition, the deletion of RY-motif will greatly reduce promoter activity in the regulatory regions of *legumin* gene of *V. faba* and *napin* in *Brassica napus* (Reidt *et al.,* 2000). The above cases verify that the motifs identified by DeepCBA have important biological functions in multiple species.

According to the previously published articles and database reports, 7 important motifs are found to be involved in regulating gene expression and plant growth and development, *etc*. For instance, the RY motif (CATGCA) is reported to be involved in the regulation of seed endosperm development (Yang *et al*., 2021). Y-patch (TC motif) is a core regulatory element which can enhance promoter activity (Jores *et al*., 2021) and some experiments have shown that *EjBZR1* can bind to the BRRE motif in the *EjCYP90A* promoter to regulate expression and fruit cell enlargement (Su *et al*., 2021). The function description of some important motifs detected by DeepCBA in PPI mode is shown in Table S7.

DeepCBA reveals regulation motifs for gene expression in PDI mode. The gradient results of motifs in PDI mode show that the region near 250 bp downstream of the gene TSS is more likely to affect gene expression (Figure S5A and Figure S13A). Figure S5B shows the motifs identified by DeepCBA based on PPI and PDI sequences. The motifs identified in two tissue types have obvious tissue specificity, and the motifs identified by the two modes (PPI and PDI) in the same tissue are also different. These results indicate that there are differences in factors that regulate gene expression through PPI and PDI (Figure S13B). To verify the reliability of motifs identified in PDI mode, we compare these motifs against the PlantTFDB (Tian *et al*., 2020; Jin *et al*., 2017; Jin *et al*., 2015; Jin *et al*., 2014) (*E*-value < 0.5). The results show that 51.52% and 48.64% of the motifs in Shoot and Ear match with those of 653 verified higher plants in PDI mode. In addition, 58% and 56% of the 104 transcription factors (Tu *et al*., 2020) match the motifs identified in Ear and Shoot (Supplementary Data 3 and Supplementary Data 4), respectively. Same to the analysis method in PPI mode, the motifs identified by DeepCBA (Supplementary Data 6) are matched with eQTLs and chromatin open regions. The following two processing modes are used as controls: (1) removing motif sequences from PDI sequences and selecting sequences of equal length from the remaining PDI sequences and (2) removing PDI sequences from the whole DNA genome and selecting sequences from the remaining genome sequences. Compared with the control, the motifs identified by DeepCBA are significantly enriched at eQTLs (Figure S5C, D) (****$P < 0.0001$, *t* test). Moreover, we match these motifs with the open chromatin regions identified in 26 NAM population lines. The results show that the motifs identified by DeepCBA are significantly enriched in the open

chromatin regions identified in the NAM population (Figure S5E, F) (****$P < 0.0001$, *t* test). Similarly, the physical locations of important motifs identified in Shoot are significantly enriched with eQTLs and chromatin open regions (Figure S10 and Figure S11) (****$P < 0.00001$, *t* test).

The physical locations of the motifs identified in different tissues in PDI sequences are clustered, and we divide them into the following four categories: (1) highly enriched near 250 bp downstream of the TSS; (2) highly enriched at specific sites; (3) poorly enriched near 250 bp downstream of the TSS; and (4) distributed evenly throughout the whole sequence (Figure 6a). The forms of sequence importance of the above first and second categories are complementary, suggesting that there may be differences in the way they work. Similar to the results in PPI mode, the distribution of the 44 overlapping motifs in the four enrichment patterns is basically same in Ear and Shoot (Table S6). MetaLogo (Chen *et al*., 2022) is used to cluster the core sequences of motifs, and 6 groups of core sequences are identified in Ear and Shoot (Figure S6B). Two motifs, GGCCCA and AAAAAA (Figure S6C, D and Figure S13C), have also been reported in previous studies (Peng *et al*., 2019; Woodhouse *et al*., 2021). Further analysis shows that the conserved region in which the transcription factor TCP binds to DNA is also the GGCCCA motif in maize. Therefore, transcription factors (such as TCP) are likely to play an important role in regulating gene expression through distal elements. Meanwhile, GGCCCA is also known as the site II motif, has been identified in the promoter region of various highly expressed genes, such as ribosomal and DEAD-box RNA helicase genes. The transcription factors of TCP and ASR5 are the examples of proteins known to bind the GGCCCA motif. Moreover, Xu identified several diurnal-related cis elements in seedlings and leaves (Xu *et al*., 2011), including element II of Arabidopsis PCNA-2 (*GGCCCA* or *AGCCCA*).

Similarly, DeepCBA also detected 12 important motifs that are reported in the PDI mode. The functions of these motifs include: enhancing promoter activity, affecting gene expression, heat resistance, salt stress tolerance, forming immune complexes, *etc*. Among them, GGCCCA has been found in the promoter regions of various highly expressed genes, and it is a binding site for the transcription factors of TCP and ASR5 (Xu *et al*., 2011; Oka *et al*., 2017). Besides, association of Telo box (AAACCTA) with site II (GGCCCA) or TEF cis-acting elements appears to be involved in ribosome biogenesis (Gaspin *et al*., 2010). GCC box (GCCGCC) has been reported that the ERF subfamily can bind to the GCC box in response to biotic stress and can also respond to ethylene by enhancing gene expression (Ishige *et al*., 1999; Wu *et al*., 2020).

The function description of some important motifs detected by DeepCBA in PDI mode is shown in Table S8.

DeepCBA identifies regulatory elements in two genes (*ZmRap2.7*, *ZmTb1*) of maize. Based on a series of motifs identified in the PDI mode, we further conducted an in-depth research on the regulatory site of *Vgt1* of the maize flowering-related gene *ZmRap2.7* (Zhao *et al*., 2018; Ricci *et al*., 2019). We extract the 70 kb sequence upstream of gene *ZmRap2.7*, and then split the sequence into sub-sequences with a length of 1.5 kb and input them into DeepCBA model. Saliency map (Chu *et al*., 2011) is used to calculate the sequence gradient, and the results show that the important motifs identified by DeepCBA are mainly distributed in open chromatin regions (Figure 5A). We further narrowed down the regions to two 500 bp regions (chr8: 135941716-135942216 and chr8: 135945716-135946216). After matching the identified motif with 104 transcription factors, enrichment analysis is performed for the above two regions. There are 18 transcription factor-binding sites in the first region and 19 transcription factor-binding sites in the second region (Figure 5B). Importantly, there are up to 16 common transcription factors in the two regions (Table S9 and Table S10 and Supplementary Data 10). Similarly, we extract a 3 kb sequence upstream of the TSS of gene *ZmRap2.7* and input it into DeepCBA model for training and calculate the sequence gradient (Figure 5C). The detected motifs are matched with the chromatin open regions, and the motifs of seven identified transcription factors are found to overlap with open chromatin regions (Figure 5D). We also conduct sequence analysis of maize tillering-related gene *ZmTb1* and its regulatory elements. The 70 kb sequence upstream of *ZmTb1* is selected for verification using DeepCBA, and the important motifs identified are also mainly distributed in the open chromatin region (Figure S14A-C). For the 10 kb (chr 1: 270482176-270492176) region with the highest gradient value, we trained DeepCBA model based on the PDI sequence and detected 314 and 514 motifs in Ear and Shoot involving 93 transcription factors. Similarly, we detected 140 and 262 motifs in the 3 kb (chr1: 270549176-270552176) region with the second highest gradient value in Ear and Shoot, involving 93 transcription factors. The numbers of overlapping motifs in the above two regions were 59 and 118, involving 84 transcription factors (Figure S14D, E). The above results reveal that there are similar transcription factor binding clusters in distal regulatory elements and their regulated genes, thus enabling a deep analysis of gene expression regulation.

DeepCBA accurately predicts gene expression values through promoter saturation mutations. Another method for generating weak alleles through targeting coding regions is using CRISPR-Cas9, which has been widely used in different plants, to edit cis-regulatory

regions (Rodríguez-Leal *et al*., 2017; Liu *et al*., 2021; Song *et al*., 2022). However, screening lines with continuous expression gradient changes in target genes among a large amount of genetic editing material comes with many uncertainties. Therefore, it is crucial to utilize computational tools to predict and screen variations in continuous gradient expression by performing saturation mutations on gene promoters. For validation, we selected the editing results in promoter region of gene *ZmCLE7* in maize (Liu *et al*., 2021). *ZmCLE7* affects yield by changing the ear phenotype, and Ear tissue is used in this study to build the deep learning prediction model. The upstream region with a length of 4 kb (chr4: 8334400-8338400) of gene *ZmCLE7* is selected as the candidate editing region (Figure 6A). Combined with the published results, the CRISPR-Cas9 edited sequences are input into DeepCBA model to predict gene expression (Figure 6B). The results show that the predicted gene expression has a trend consistent with the expression obtained experimentally and the correlation between the predicted gene expression after editing the target segment and the qPCR value is 0.51 (Figure 6C). To explore more precisely how the 4 kb target sequence affects the expression of *ZmCLE7*, we used sliding window methods (window size = 200 bp, step = 200 bp) to process the 4 kb sequence and obtained 12 sequences with a length of 3 kb (Figure 6D). Then, we used the DeepCBA model to predict gene expression for the above 12 edited sequences. The results show that a wider range of expression variation types can be produced than are produced in the biological experiment results (Figure 6E).

To further verify the reliability of the DeepCBA model in gene editing applications, the promoter (chr5: 205820586-205829816) of the gene *ZmVTE4*, which affects the vitamin E content of maize is edited. Combined with the distribution characteristics of different histone modifications in the target region (Figure 6F), we design 7 fragment deletion types for CRISPR-Cas9 editing (Figure 6G). These sequences are input into the DeepCBA model of Ear tissue for expression prediction. Moreover, we extracted RNA from the edited individual plants and detected the expression of target genes and the correlation between the predicted gene expression after editing the target segment and the qPCR value is 0.57 (Figure 6H). The results show the same trend of change of the predicted expression and the real gene expression. Taken together, the above results demonstrate the feasibility of DeepCBA for promoter saturation mutations and provide a powerful tool for the precise design of desired gene expression.

DeepCBA realizes cross-tissue and cross-genotype maize gene expression prediction. With the development of machine learning technologies, many methods have been applied to the study of different tissues, materials and species (Kelley *et al*., 2018). To further explore the

feasibility of predicting expression between different tissues and materials, we constructed a transfer learning model to achieve gene expression prediction across tissues (Ear, Shoot) and materials (B73, SK (Yang *et al*., 2019)) (Figure S16). In addition, we used a new shoot dataset (Peng *et al*., 2019) to further verify the generalization ability of DeepCBA. The new Shoot dataset (Peng *et al*., 2019) contains 43,865 PPIs involving 20,695 genes. The Shoot dataset of SK (Yang *et al*., 2019) contains 7,394 PPIs involving 7,099 genes. The PCC of DeepCBA increases from 0.7601 to 0.8687 after applying transfer learning on the Shoot dataset (Yang *et al*., 2019) (Figure S17A). In addition, we compared the running time of DeepCBA for gene expression prediction with and without transfer learning in different datasets. The results indicate that transfer learning can improve the prediction accuracy while reducing the time of model training and prediction (Figure S17B). For the cross-tissue gene expression prediction: (1) Based on training a prediction model using the PPI and PDI datasets for Shoot (Peng *et al*., 2019), we predict gene expression for the Shoot (Li *et al*., 2019) and Ear (Li *et al*., 2019) (Figure S18A). The PCC for the predicted expression and true expression are 0.8811 and 0.888, respectively. (2) Based on training a prediction model using the PPI and PDI datasets in Ear (Li *et al*., 2019), we predict gene expression for the tissues of Shoot (Peng *et al*., 2019) and Shoot (Li *et al*., 2019) (Figure S18B). The PCC for the predicted expression and true expression are 0.8848 and 0.8737, respectively. (3) Based on training a prediction model using the PPI and PDI datasets in Shoot (Li *et al*., 2019), we predict gene expression for the tissues of Ear (Li *et al*., 2019) and Shoot (Peng *et al.*, 2019). The PCC for the predicted expression and true expression are 0.8931 and 0.8689, respectively (Figure S18C). For the cross-genotype and cross-tissue gene expression prediction, we predict the gene expression of the shoot tissue of SK inbred line based on the three tissue types (Shoot (Peng *et al*., 2019), Ear (Li *et al*., 2019), and Shoot (Li *et al*., 2019)) of B73 material in maize. The PCC for the predicted expression and true expression is 0.8687, 0.8583, and 0.8623, respectively. (Figure S18D, E, F). In all, the PCC for the gene expression prediction among different tissues and materials basically exceed 0.85 through transfer learning. This study provides a reference for predicting gene expression across different tissues and materials and broadens the application scope of the DeepCBA model.

DeepCBA provides a real-time online website. To facilitate open access use of the DeepCBA model, we developed the DeepCBA online website (http://www.deepcba.com/ or http://124.220.197.196/). The website provides the function of high-precision gene expression prediction based on chromatin interactions in maize (and other three crops of rice, cotton and

wheat). Users can select any of the four crops and input any interaction sequences that meet the requirements to predict the expression of the related genes and sequences. Additionally, the website provides a visualization interface to display the gradient importance of the input sequences (Figure 7).

Discussion

The coding regions in maize account for only a small part of the entire genome, and most of the genome is noncoding regions. Many functional loci have been identified in the noncoding regions of maize through association analysis, and several kinds of regulatory elements have been identified through epigenetics analyses at the genome-wide level. However, it is still unclear how the regulatory elements in noncoding regions accurately regulate gene expression. Using deep learning tools to predict the contribution of different regulatory elements to gene expression has important biological significance. As we known, chromatin interactions have important impact on the target gene expression. An important question is therefore to what extent gene expression is determined by the chromatin interactions of DNA sequences.

This study developed a model, called DeepCBA, for high-precision gene expression prediction based on maize chromatin interactions. The CNN is used to extract local features in the DNA sequence, BiLSTM is innovatively used to capture the relationships between distal features, and the self-attention mechanism is used to capture important features. Compared with existing methods, DeepCBA has higher accuracy in gene expression classification and expression value prediction. DeepCBA predicts gene expression accurately by integrating chromatin interaction (PPI, PDI) data in different maize tissues, and the effect size of different regulatory elements on gene expression is quantitatively assessed. It reveals that the average contribution of promoter proximal interaction (PPI), proximal-distal interaction (PDI), proximal and distal interaction (PPI+PDI) for predicting gene expression is 0.817, 0.625 and 0.929, which is improved by 0.357, 0.165 and 0.469 compared to the single sequence method.

Unraveling the black box of deep learning-based applications remains a challenge in biological researches. To interpret the reasoning process of DeepCBA, we used saliency map to calculate the model gradient through the reverse calculation and obtained significance map of DNA sequences. We identified important motifs in the chromatin interaction sequences together with other latent features yet unknown for the prediction of gene expression. Verification against known databases and the published literature shows that the detected motifs have high reliability. In terms of molecular characteristics, the motifs identified in this study

were mainly enriched in eQTLs and chromatin open regions. Moreover, gene expression showed an upward trend as the number of motifs in the motif composition increased (Figure 3F, G). The detected motifs and gradient results of different maize tissues (Ear, Shoot) showed obvious tissue specificity (Figure S5B). The identified motifs are clustered into six groups of core sequences in ear and shoot (Figure S9). In addition, the distribution pattern of the motifs can be divided into 5 and 4 categories in PPI and PDI mode, respectively (Figure 3C, Figure S6A).

Alleles that control important traits (such as crop yield, resistance) often alter the expression level of genes, thereby affecting phenotype. Editing the promoter region helps to intelligently design the expression of target genes, achieving the goal of improving crop yield and resistance (Rodríguez-Leal *et al*., 2017; Liu *et al*., 2021; Song *et al*., 2022). For the regulatory elements in noncoding regions, DeepCBA can detect the gradient effect of regulatory elements at the single base level and evaluate the functional loci of regulatory elements accurately. The feasibility of DeepCBA in exploring regulatory elements that affect gene expression is validated through the reported maize flowering-related gene *ZmRap2.7* and tillering-related gene *ZmTb1* (Figure 5, Figure S14). Through saturation mutations in the promoter and regulatory regions of specific genes (*ZmCLE7*, *ZmVTE4*), de novo gene expression prediction was achieved (Figure 6). This study validated the reliability of DeepCBA through real examples in maize. This model can be widely used for precise design of target gene expression levels in different crops, thereby serving intelligent design and breeding.

To further explore the feasibility of predicting expression in practical applications, a transfer learning model of DeepCBA was constructed to achieve gene expression prediction across tissues (Ear, Shoot) and materials (Figure S18). The gene expression prediction PCC of the cross-tissue and cross-genotype through the transfer learning model exceeds 85%, verifying the wide application scope of DeepCBA. To facilitate the use of DeepCBA model, we have developed a friendly online website (http://www.deepcba.com/ or http://124.220.197.196/) for gene expression prediction and sequence importance visualization about four crops (maize, rice, cotton and wheat).

Interestingly, the expression of some genes predicted using DeepCBA were lower than their actual expression value. This is maybe related to the small number of chromatin interactions in different tissues and materials. That is to say, there are no sufficient data for training DeepCBA and influencing the prediction effect. In addition, the prediction accuracy will be improved through integrating multi-omics data, including chromatin open regions,

transcription factor binding sites, histone modifications, DNA methylation, *etc*.

Artificial intelligence is continuing to penetrate various fields, and machine learning has advantages in exploring the impact of different regulatory elements on gene expression. With the development of different deep learning algorithms, the understanding of different regulatory elements will gradually become clearer in the future. We will have a deeper understanding of the effect of variation on gene expression when considering tissue specificity and spatiotemporal specificity. This study will also provide a theoretical basis for accurately designing gene expression and optimizing intelligent breeding in the future.

Materials and methods

Data collection and processing

Data statistical analysis. The experimental data used in this study are the published datasets of maize chromatin interactions and gene expression (Li *et al*., 2019; Peng *et al*., 2019.). The data involves two tissues of shoot and ear. The chromatin interaction includes two types: gene-gene promoter proximal interaction (PPI), gene and proximal-distal interaction (PDI) (Li *et al*., 2019). The Ear (Li *et al*., 2019) dataset contains 35,332 PPIs involving 20,601 genes. The Shoot (Li *et al*., 2019) dataset contains 65,064 PPIs involving 23,707 genes. In addition, we used other three datasets of Shoot (Peng *et al*., 2019), Ear (Sun *et al*., 2020), Tassel (Sun *et al*., 2020) (Table S3) to evaluate the performance of DeepCBA in the prediction of gene expression classification and regression (Figure S2 and Figure S3).

To balance the number of genes in different categories, we classified genes into Unexpressed, Expressed, and Highly-expressed according to the expression range of [0-0.1), [0.1-15), and [15-max], respectively (Table S1). To reduce false-positives in the training process, we divided the training and test datasets based on gene family information. The number of genes with expression in the range of [0-100], [1-100], [0-500], and [0-max] can be seen in Table S4. More than 99% genes had expression in the range of [0-500], so we used these genes for the experimental analysis of expression prediction.

PDI data processing. The PDI datasets (Li *et al*., 2019) of Shoot and Ear include 11,207 and 11,189 PDIs, respectively. The length of the intergenic distal sequences of different tissues is mainly in the range of 1 kb ~ 2 kb, and we set the length of the distal sequences to 1.5 kb. The promoter proximal sequence was also set to 1.5 kb, including 1 kb upstream and 0.5 kb downstream of the gene TSS. When the length of the distal sequence was less than 1.5 kb, we supplemented *N* sequences at both ends of the sequence to 1.5 kb. When the length of the distal

sequence was greater than 1.5 kb, 750 bp sequences were extracted from the middle of the sequence to both sides to form a 1.5 kb sequence.

Data augmentation. For the chromatin interactions of gene A and gene B in PPI mode, it may be that gene A affects the expression of gene B, or it may be that gene B affects the expression of gene A. Therefore, we doubled the number of PPI interactions by adjusting the order of gene sequences, which is called data enrichment.

Quantitative real-time polymerase chain reaction(qRT-PCR). The qRT-PCR data of *ZmCLE7* gene comes from the prior research of (Liu *et al*., 2021). For qRT-PCR analysis, 0.1 g of plant tissue was used to extract total RNA using Quick RNA Isolation Kit (Huayueyang Biotechnology Co. Ltd, Beijing, China). Sources of the analyzed leaves tissue are from *ZmVTE4* Editing materials. EasyScript One-Step gDNA Removal and cDNA Synthesis SuperMix (TransGen Biotech Co. Ltd, Beijing, China) was used to remove the gDNA from the extracted RNA and synthesize first-strand complementary DNA. Real-time fluorescence quantitative polymerase chain reaction with SYBR Green Master Mix (Vazyme Biotech Co. Ltd, Nanjing, China) on a CFX96 Real-Time System (Hercules, CA, USA) was used to quantify the expression level of editing materials. The primers used for quantitative qRT-PCR are listed in Supplementary data (Supplementary Data 13 and 14). The concrete information of the target locations involved in *ZmVTE4* gene editing is shown in Figure S15.

Methods

Model architecture. The DeepCBA model uses one-hot encoding, and it includes three modules: CNN, BiLSTM, and Self-attention. The CNN module includes three parts, and each part contains two convolutional layers. Each convolutional layer connects to a maximum pooling layer to achieve feature dimensionality reduction and feature re-extraction. BiLSTM is used to capture feature information about the proximal and distal chromatin sequences and then mine important motif features that affect gene expression. The self-attention mechanism redistributes the weights of the parameters trained in the model to capture important features. To reduce overfitting, batch normalization and dropout mechanisms are used. In the last layer of DeepCBA, the softmax or linear activation functions are used to perform the prediction task (Figure 1B and Table S13).

The batch size, number of convolutional filters and the size of convolutional kernels are the hyperparameters of DeepCBA. Experimental results show that the batch size has large impact on the prediction results (Figure S19A). The model's prediction accuracy also improves as the number and size of convolutional kernels increase. By comprehensively considering the

running time and memory space, the number and size of convolution kernels are set to 64 and 8, respectively (Figure S19B, C and Table S13).

We supposed that the promoter proximal region sequence P_{seq_G} of target gene *G*, P_{seq_G} consists of two parts: 1 kb sequence upstream and 0.5 kb sequence downstream of gene TSS, 0.5 kb sequence upstream and 1 kb sequence downstream of gene TTS. The expression level of *G* is V_{ag}, P_{seq_A} denotes the promoter proximal region sequence of *gene_A* that has a PPI interaction with *G*, and the following four steps are used to construct the prediction model for gene *G* based on PPI chromatin interactions.

One-hot encoding. One-hot encoding is used to process P_{seq_G} and P_{seq_A} with a length of 3 kb, that is, A=$[1,0,0,0]^T$, C=$[0,1,0,0]^T$, G=$[0,0,1,0]^T$, T=$[0,0,0,1]^T$, N=$[0,0,0,0]^T$. Then, P_{seq_G} is encoded as matrix $M_1 \in R^{4\times3000}$, and P_{seq_A} is encoded as matrix $M_2 \in R^{4\times3000}$. Then, M_1 and M_2 are concatenated vertically and input into the model, P=Concat (M_1, M_2), $P \in R^{4\times6000}$.

CNN Layer. In DeepCBA, the convolution operation in the CNN is firstly used to perform dimensionality reduction and extract important features. For an encoded matrix $M \in R^{(M*N)}$ and a filter $W \in R^{U*V}$, $U << M, V << N$, the convolution operation is shown in Eq. (1).

$$G_{ij} = \sum_{u=1}^{U}\sum_{v=1}^{V} W_{uv} M_{i-u+1,j-v+1} \tag{1}$$

For example, the dimension reduction operation on matrices M_1 and M_2 is shown in Eq. (2) and Eq. (3).

$$G_1[m_1,n_1] = (M_1 * W)[m_1,n_1] = \sum_j \sum_k W[j,k] M_1[m_1 - j, n_1 - k] \tag{2}$$

$$G_2[m_2,n_2] = (M_2 * W)[m_2,n_2] = \sum_j \sum_k W[j,k] M_2[m_2 - j, n_2 - k] \tag{3}$$

W represents the filter, and m_i and n_i represent the number of rows and columns of the matrix after dimension reduction. Then, we obtain the reduced dimension matrices $G_1[m_1,n_1]$ and $G_2[m_2,n_2]$.

The maximum pooling operation is used for secondary dimensionality reduction to solve the problem of overfitting. Based on the feature map $G_i[m_i,n_i] \in R^{M*N*D}$ obtained through the convolution operation, each feature map $G^d \in R^{M*N}$ $(1 \le d \le D)$ can be divided into multiple regions $R^d_{m,n}$ $(1 \le m \le M', 1 \le n \le N')$. Then, the maximum value within $R^d_{m,n}$ is selected to represent the region, as shown in Eq. (4).

$$Y^d_{m,n} = \max_{i \in R^d_{m,n}} G_i \tag{4}$$

BiLSTM Layer. For the promoter proximal sequence P_{seq_G} of gene *G,* and P_{seq_A} denoting the promoter proximal sequence of *gene_A* that has a PPI interaction with *G*, $Y_{m1,m2}$ and $Y_{m1,m2}$ are obtained through the CNN model. m_1=m_2=3, n_1=n_2=128. According to the self-loop update idea of the BiLSTM input gate, forget gate and output gate, the update operation at time *t* (taking Y_{m_1,n_1} as an example) is shown in Eq. (5).

$$f_i^{[m_1,n_1](t)} = \sigma(b_i^f + \sum_j U_{i,j}^f Y_j^{[m_1,n_1](t)} + \sum_j W_{i,j}^f h_j^{[m_1,n_1](t-1)}) \tag{5-1}$$

$$S_i^{[m_1,n_1](t)} = f_i^{[m_1,n_1](t)} S_i^{[m_1,n_1](t-1)} g_i^{[m_1,n_1](t)} \sigma(b_i + \sum_j U_{i,j} Y_j^{[m_1,n_1](t)} + \sum_j W_{i,j} h_j^{[m_1,n_1](t-1)}) \tag{5-2}$$

$$g_i^{[m_1,n_1](t)} = \sigma(b_i^g + \sum_j U_{i,j}^g Y_j^{[m_1,n_1](t)} + \sum_j W_{i,j}^g h_j^{[m_1,n_1](t-1)}) \tag{5-3}$$

$$O_i^{[m_1,n_1](t)} = \sigma(b_i^o + \sum_j U_{i,j}^o Y_j^{[m_1,n_1](t)} + \sum_j W_{i,j}^o h_j^{[m_1,n_1](t-1)}) \tag{5-4}$$

$$h_i^{[m_1,n_1](t)} = tanh(S_i^{[m_1,n_1](t)} O_i^{[m_1,n_1](t)}) \tag{5-5}$$

h^t denotes the current hidden layer vector. i, j denote the *i* and *j-th* neurons, respectively. h^t includes the output of all LSTM cells. b^f, U^f, W^f represent the bias value, input weight and cycle weight of the corresponding threshold units, respectively. $O_i^{[m_1,n_1](t)}$ represents the output of the *i-th* neuron at the current time *t*. BiLSTM fully integrates the temporal and pre/post feature information of Y_{m_1,n_1} and Y_{m_2,n_2}, reduces its dimensionality to 3*64, and obtains $O_i^{[m_1,n_1]}$ and $O_i^{[m2,n2]}$. m_1=m_2=3, n_1=n_2=64.

Self-Attention Layer. O_{m_1,n_1} and O_{m_2,n_2} denote the feature matrices of P_{seq_G} and $P_{seq\_A,}$ respectively. It composites O_{m_1,n_1} and O_{m_2,n_2} vertically and integrates the attention mechanism to realize the redistribution of weights and predict target gene expression. Through compositing O_{m_1,n_1} and O_{m_2,n_2} vertically, O_{s_1,s_2} is obtained. s_1=m_1+m_2=6, s_2=n_1+n_2=64. To improve the accuracy of feature extraction, an attention mechanism is used after the BiLSTM module to realize the redistribution of weights, as shown in Eq. (6) (taking O_{s_1,s_2} as an example).

$$f_i^{s_1,s_2} = tanh(W_w O_{s_1,\ s_2} + b_w), a_i = \frac{exp(b_i^T b_w)}{\sum_i exp(b_i^T b_w)}, V_s = \sum_i a_i f_i^{s_1,s_2} \tag{6}$$

b_i represents the implicit representation of feature $O_{s1,s2}$ in the BiLSTM layer. The importance of feature $O_{s1,s2}$ is measured by the similarity between b_i and the sequence vector b_w. Then, *tanh* is used to normalize the weight of each feature. Each feature is multiplied by its corresponding weight through the attention mechanism and a summation is conducted to obtain the output vector V_s. By setting *dropout* to prevent overfitting, it uses a linear function to obtain

the expression V_{ag} of the target gene *G*.

Transfer learning. Transfer learning is widely used to reduce the training time of the model and achieve better results with a small amount of data. We use transfer learning to fine-tune and train DeepCBA model, thus to predict gene expression across tissues and genotypes (Figure S16 and Figure S17). Specifically, we freeze the first 0 to 19 layers of the pre-trained model and conduct fine-tuning through small batch training (learning rate=1e-4, batch size=64, epochs=200).

DeepCBA mines important DNA sequence regions and motifs. To mine important motifs that affect gene expression in the PPI and PDI modes, saliency map (Chu *et al*., 2011) is used to calculate the significance map (true positive, TP; true negative, TN), and it generates the model gradient through the reverse calculation. The gradient value is denoted as 4*6000**N*, where *N* represents the sum of the number of TPs and TNs. Finally, it uses TF-MoDIsco (Avanti Shrikumar *et al*., 2018) to mine important motifs (Figure S4B, C). The important identified motifs that affect gene expression are compared with the TOMTOM (Gupta *et al*., 2007) platform and the 653 conserved structural regions of higher plants in the PlantTFDB (Table S11and Table S12) to evaluate the reliability of the motifs.

Different motif compositions affect gene expression. (1) Background sequences are generated by filling a sequence with a length of 3 kb based on the *N* coding mode. (2) The important motifs identified by DeepCBA are combined to form a motif composition (*Real motif*) and A/C/G/T are randomly combined to form a motif composition with equal length (*Random motif*). (3) The *real motif* and *random motif* are embedded into the background sequence and the gene expression is predicted through the DeepCBA model (Figure S7B).

Motif enrichment in the open chromatin region. To mine the characteristics of important motifs identified by DeepCBA in Shoot and Ear, the true physical location of each motif in the chromosome are identified. Then, the physical locations of the identified motifs are matched to the chromatin open regions in the NAM population. The operation is repeated 100 times to establish the following two controls: (1) Motif sequences removed from the PPI and PDI sequences and sequences with equal length selected from the remaining PPI and PDI sequences. (2) PPI and PDI sequences removed from the whole genome and sequences with equal length selected from the remaining genome sequences.

Data Availability Statement

Methods, including statements of data availability and any associated accession codes and references, are available at https://github.com/Jie-Lii/DeepCBA.

Funding

This work has been supported by the National Key Research and Development Program of China [2022YFD1201504], the National Key R&D Program of China [2023ZD04076, 2023ZD04061], the Fundamental Research Funds for the Central Universities [2662022YLYJ010, 2021ZKPY018, 2662021JC008, SZYJY2021003], the Major Project of Hubei Hongshan Laboratory [2022HSZD031], the Major Science and Technology Project of Hubei Province [2021AFB002], the Yingzi Tech & Huazhong Agricultural University Intelligent Research Institute of Food Health [IRIFH202209].

Author contributions

Conceptualization, J, Y., and J, L. (Jianxiao Liu); Methodology, Z, W. and J, L. (Jie Li); Writing - Original Draft, Z, W.; Writing - Review & Editing, Z, W., Y, P., and J, L. (Jianxiao Liu); Funding Acquisition J, L. (Jianxiao Liu).

Acknowledgments

Thanks to the anonymous reviewers for the constructive comments, which helped us to substantially improve the manuscript. Thanks to the experimental teaching center of College of Informatics in Huazhong Agricultural University for providing the experimental environment and computation resources.

Conflicts of interest

The authors declare no competing financial interests.

References

Avsec, Ž., Agarwal, V., Visentin, D. et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods.* **18**:1196–1203.

Beer, M. A., Tavazoie, S. (2004). Predicting gene expression from sequence. *Cell*, **117(2)**: 185-198.

Bailey, T. L., Johnson, J., Grant, C. E., et al. (2015). The MEME suite. *Nucleic Acids Research*. **43(W1)**: W39-W49.

Cheng, A., Grant, C. E., Noble, W. S., et al. (2019). MoMo: discovery of statistically significant post-translational modification motifs. *Bioinformatics*. 35(16): 2774-2782.

Cheng, C., Yan, K. K., Yip, K. Y., et al. (2011). A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biology*. **12**: 1-18.

Chu, C. (2011). Saliency mapping of figure and ground of motion in Chinese. *Journal of Chinese Language Teachers Association*. **46(2)**: 49-69.

Chen, Y., He, Z., Men, Y., et al. (2022). MetaLogo: a heterogeneity-aware sequence logo generator and aligner. *Briefings in Bioinformatics*. **23(2)**: bbab591.

Cao, X., Costa, L. M., Biderre-Petit, et al. (2007). Abscisic acid and stress signals induce Viviparous1 expression in seed and vegetative tissues of maize. *Plant Physiology*. **143(2)**: 720-731.

Dong, X., Greven, M. C., Kundaje, A., et al. (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*. **13**: 1-17.

Fu, J., Cheng, Y., Linghu, J., et al. (2013). RNA sequencing reveals the complex regulatory network in the maize kernel. *Nature Communications*. **4(1)**: 2832.

Guerriero, G., Martin, N., Golovko, A. et al. (2009). The RY/Sph element mediates transcriptional repression of maturation genes from late maturation to early seedling growth. *New Phytologist*. **184(3)**: 552-565.

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics.* **27(7)**: 1017-1018.

Gaspin, C., Rami, J. F., & Lescure, B. (2010). Distribution of short interstitial telomere motifs in two plant genomes: putative origin and function. *BMC Plant Biology*. **10**: 1-12.

Gupta, S., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biology*. **8**: 1-9.

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., et al. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*. **373(6555)**: 655-662.

Ishige, F., Takaichi, M., Foster, R., et al. (1999). AG-box motif (GCCACGTGCC) tetramer confers high-level constitutive expression in dicot and monocot plants. *The Plant Journal*. **18(4)**: 443-448.

Jin, J., Tian, F., Yang, D. C., et al. (2016). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*. gkw982.

Jin, J., He, K., Tang, X., et al. (2015). An Arabidopsis transcriptional regulatory map reveals distinct functional and evolutionary features of novel transcription factors. *Molecular Biology and Evolution*. **32(7)**: 1767-1773.

Jin, J., Zhang, H., Kong, L., et al. (2014). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research*. **42(D1)**: D1182-D1187.

Karlić, R., Chung, H. R., Lasserre, J., et al. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*. **107(7)**: 2926-2931.

Kelley, D. R., Reshef, Y. A., Bileschi, M., et al. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*. **28(5)**: 739-750.

Lee, D., Yang, J., & Kim, S. (2022). Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer. *Nature Communications*. **13(1)**: 6678.

Li, E., Liu, H., Huang, L., et al. (2019). Long-range interactions between proximal and distal regulatory regions in maize. *Nature Communications*. **10(1)**: 2633.

Liu, L., Zhang, G., He, S., et al. (2021). TSPTFBS: a docker image for trans-species prediction of transcription factor binding sites in plants. *Bioinformatics*. **37(2)**: 260-262.

Liu, L., Gallagher, J., Arevalo, E. D., et al. (2021). Enhancing grain-yield-related traits by CRISPR–Cas9 promoter editing of maize CLE genes. *Nature Plants*. **7(3)**: 287-294.

Mönke, G., Altschmied, L., Tewes, A., et al. (2004). Seed-specific transcription factors ABI3 and FUS3: molecular interaction with DNA. *Planta*. **219**: 158-166.

Machanick, P., Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*. **27(12)**: 1696-1697.

O'Connor, T., Grant, C. E., Bodén, M., et al. (2020). T-Gene: improved target gene prediction. *Bioinformatics*. **36(12)**: 3902-3904.

Oka, R., Zicola, J., Weber, B., et al. (2017). Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biology*. **18**: 1-24.

Peng, Y., Xiong, D., Zhao, L., et al. (2019). Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. *Nature Communications*. **10(1)**: 2632.

Ricci, W. A., Lu, Z., Ji, L., et al. (2019). Widespread long-range cis-regulatory elements in the maize genome. *Nature Plants*. **5(12)**: 1237-1249.

Rodríguez-Leal, D., Lemmon, Z. H., Man, J., et al. (2017). Engineering quantitative trait variation for crop improvement by genome editing. *Cell*. **171(2)**: 470-480.

Rodgers-Melnick, E., Vera, D. L., Bass, H. W., et al. (2016). Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences*. **113(22)**: E3177-E3184.

Reidt, W., Wohlfarth, T., Ellerström, M. et al. (2000). Gene regulation during late embryogenesis: the RY motif of maturation-specific gene promoters is a direct target of the FUS3 gene product. *The Plant Journal*, **21(5)**, 401-408.

Schmidt, F., Gasparoni, N., Gasparoni, G., et al. (2017). Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Research*. **45(1)**: 54-66.

Schoenfelder, S., Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*. **20(8)**: 437-455.

Song, X., Meng, X., Guo, H., et al. (2022). Targeting a gene regulatory element enhances rice grain yield by decoupling panicle number and size. *Nature Biotechnology*. **40(9)**: 1403-1411.

Shrikumar, A., Tian, K., Avsec, Ž., et al. (2018). Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5. arXiv preprint arXiv:1811.00416.

Tian, F., Yang, D. C., Meng, Y. Q., et al. (2020). PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Research*. **48(D1)**: D1104-D1113.

Tu, X., Mejía-Guerra, M. K., et al. (2020). Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nature Communications*. **11(1)**: 5089.

Tasaki, S., Gaiteri, C., Mostafavi, S., et al. (2020). Deep learning decodes the principles of differential gene expression. *Nature Machine Intelligence*. **2(7)**: 376-386.

Tian, T., Wang, S., Yang, S., et al. (2023). Genome assembly and genetic dissection of a prominent drought-resistant maize germplasm. *Nature Genetics*. **55(3)**: 496-506.

Theune, Marius L., et al. (2019) Phylogenetic analyses and GAGA-motif binding studies of BBR/BPC proteins lend to clues in GAGA-motif recognition and a regulatory role in brassinosteroid signaling. *Frontiers in Plant Science.* **10**: 428233.

Washburn, J. D., Mejia-Guerra, M. K., Ramstein, G., et al. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proceedings of the National Academy of Sciences*. **116(12)**: 5542-5549.

Wu, J., Deng, Y., Hu, J., et al. (2020). Genome-wide analyses of direct target genes of an ERF11 transcription factor involved in plant defense against bacterial pathogens. *Biochemical and Biophysical Research Communications*. **532(1)**: 76-81.

Woodhouse, M. R., Cannon, E. K., Portwood, J. L., et al. (2021). A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biology*. **21**: 1-10.

Zrimec, J., Börlin, C. S., Buric, F., et al. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nature Communications*. **11(1)**: 6141.

Zhou, J., Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*. **12(10)**: 931-934.

Zhao, H., Tu, Z., Liu, Y., et al. (2021). PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. *Nucleic Acids Research*. **49(W1)**: W523-W529.

Zhou, J., Theesfeld, C. L., Yao, K., et al. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*. **50(8)**: 1171-1179.

Zhao, H., Zhang, W., Chen, L., et al. (2018). Proliferation of regulatory DNA elements derived from transposable elements in the maize genome. *Plant Physiology*. **176(4)**: 2789-2803.

Zrimec, J., Börlin, C. S., Buric, F., et al. (2020). Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nature Communications*. **11(1)**: 6141.

Figure Legends

Figure 1 The workflow of DeepCBA. **A** Two kinds of chromatin interactions: PPI and PDI. 1.5 kb gene proximal sequence of TSS and TTS. **B** Five steps in DeepCBA: sequence encoding, feature extraction using CNN, temporal and distal feature extraction using BiLSTM, attention mechanism and gene expression prediction. **C** The PCC comparison results of the three methods for predicting gene expression values. CNN_No_PPI: The CNN model only uses gene upstream and downstream sequences. DeepCBA_No_PPI: The DeepCBA model only uses gene upstream and downstream sequences. DeepCBA_PPI: The DeepCBA model uses interaction sequences. Data augmentation denotes considering gene order in PPI mode during model training.

Figure 2 Performance of DeepCBA for maize gene expression value prediction in different modes. From left to right are the results of predicting gene expression in the case of PDI, PPI, PDI+PPI, respectively. **A-D** The distribution of predicted values and true values of gene expression when DeepCBA model is used to predict gene expression of Shoot in the case of PDI, PPI, PDI+PPI, respectively. **E-H** The distribution of predicted values and true values of gene expression when DeepCBA model is used to predict gene expression of Ear in the case of PDI, PPI, PDI+PPI, respectively.

Figure 3 The motifs influencing gene expression are identified based on the PPI sequence. A The impact on expression prediction of two interacting sequences input into the DeepCBA model. **B** The Venn diagram shows the motifs identified by DeepCBA in Shoot and Ear in PPI mode. **C** Five different

distribution patterns of motifs identified in Shoot and Ear: (1) Highly enriched near 250 bp downstream of the TSS. (2) Highly enriched at specific positions. (3) Poorly enrichment near 250 bp downstream of the TSS. (4) Poorly enrichment near TSSs but high enrichment near TTSs. (5) Even distribution across the whole sequence. **D, E** The core motif sequence obtained using MetaLogo based on the motifs identified in Ear and Shoot in PPI mode. There are 6 core sequences obtained in the two tissues. **F, G** Expression changes in Expressed and Highly expressed genes containing different numbers of motifs in Ear and Shoot.

Figure 4 The epigenetic features and examples of regulated gene expression of motifs identified by DeepCBA (Ear). A The matching number of motifs with different lengths and eQTLs in PPI mode. The motif sequence was removed from the PPI sequences, and sequences with lengths of 6 ~ 10 were randomly selected from the remaining sequences as controls (****$P < 0.0001$, ***$P < 0.01$, **$P < 0.05$, *t* test). **B** The matching number of motifs with different lengths and eQTLs in PPI mode. PPI interaction sequences were removed from the whole genome, and sequences with lengths of 6 to 10 were randomly selected from the remaining sequences as controls (****$P < 0.0001$, ***$P < 0.01$, **$P < 0.05$, *t* test). **C, D** The matching number of motifs identified by DeepCBA in PPI mode and the chromatin open region in the NAM population. The mode of control selection is the same as that of eQTL matching. **E** For the identified CATGCA motif in the *Zm00001d042609* sequence in PPI mode, the motif and the downstream gene *Zm00001d042600* can be bound by the transcription factor nactf109 simultaneously. The reported results verified that the variation in the CATGCA of the maize association mapping panel (AMP) will lead to expression changes in the gene *Zm00001d042600*, thus affecting the maize drought resistance phenotype in the seedling stage.

Figure 5 Identification of regulatory elements in *ZmRap2.7*. A The open chromatin region distribution of 70 kb sequences upstream of gene *ZmRap2.7* in different tissues, and the sequence gradient value calculated through DeepCBA. **B** The identified motifs and transcription factors that can be bound in the two regions (chr8: 135941716-135942216; chr8: 135945716-135946216) with the highest gradient value. **C, D** The motifs and transcription factors that can be bound in the 3 kb region upstream of the TSS of gene *ZmRap2.7*. (In the figure, ①②③④ represent DNase-seq, ATAC-seq, H3K4me3, and H3K9ac respectively.)

Figure 6 DeepCBA edits the maize genes of *ZmCLE7* and *ZmVTE4* to achieve accurate expression prediction. A The distribution results of four histone modifications (H3K27ac, H3K4me3, H3K27me3, H3K9ac) and chromatin open regions within the 4 kb upstream region of the gene *ZmCLE7*. **B** Schematic diagram of 6 pieces of editing information for *ZmCLE7* in the published literature (Liu *et al*., 2021). **C** DeepCBA was used to predict the expression of gene-edited sequences in (b) and compare it with Quantitative Real-time PCR (qPCR) results in the published literature. **D** Using the sliding windows method (window size = 200 bp, step = 200 bp) to process the 4 kb sequence of *ZmCLE7*. **E** The predicted expression of the edited sequences in (d) using DeepCBA. **F** Distribution of three histone modifications (H3K27ac, H3K4me3, H3K9ac) and open chromatin regions within the 4 kb region upstream of the gene *ZmVTE4*. **G** Gene editing events in the 4 kb region upstream of *ZmVTE4*. **H** Expression comparison of *ZmVTE4* predicted by DeepCBA and the results of leaf Quantitative Real-time PCR (qPCR) for the 4 kb upstream region editing of *ZmVTE4*.

Figure 7 The online website of DeepCBA. A The functions of DeepCBA online website. **B** DeepCBA provides the function of high-precision gene expression prediction based on chromatin interaction of four crops: maize, rice, cotton and wheat. Users can freely select relevant models to achieve the prediction tasks. **C** DeepCBA implements a parallel computing algorithm. The prediction results are sent to users via email and users can view the results according to the Job_id. **D** DeepCBA provides a visualization interface to display the gradient importance of the input sequences affecting gene expression.

Figures

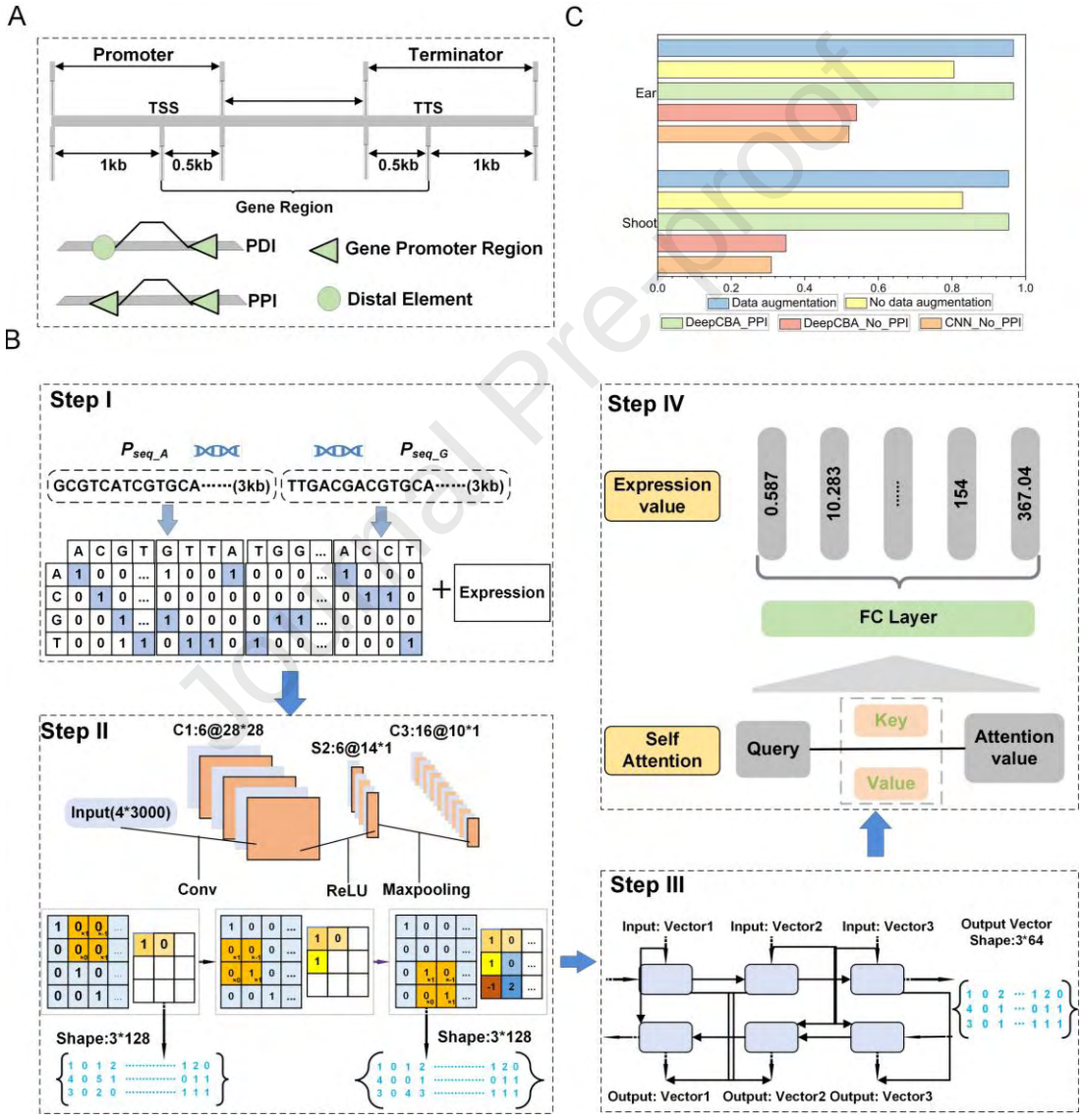


Figure 1

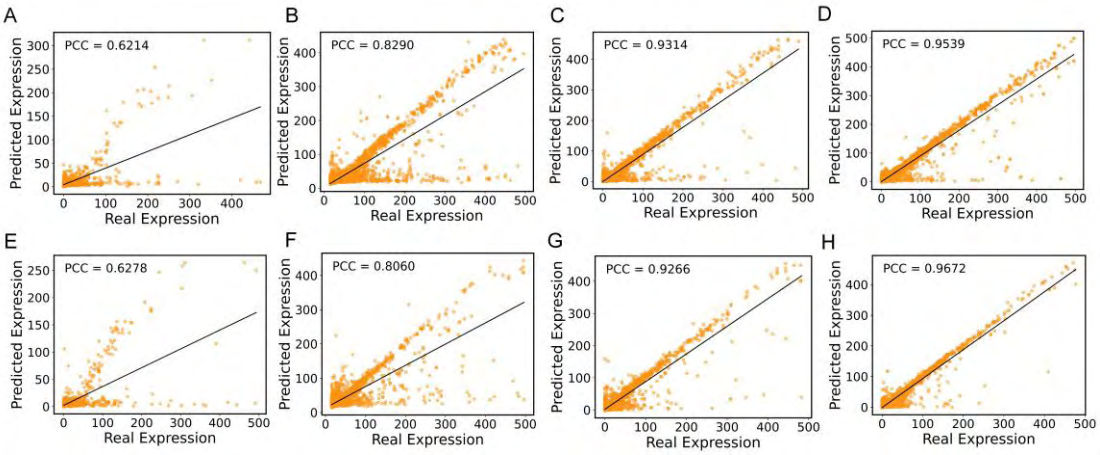


Figure 2

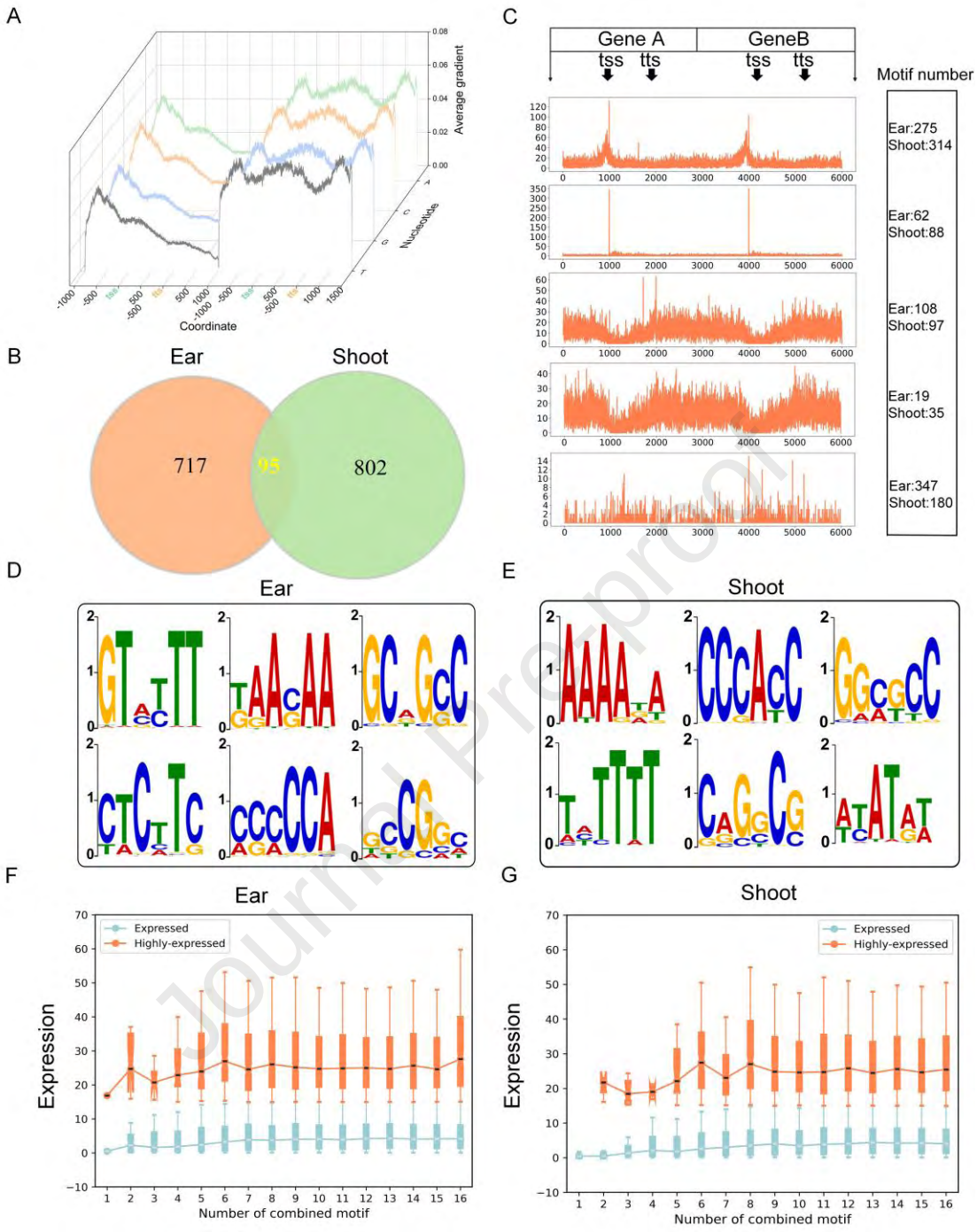

A
Average gradient
Nucleotides
Coordinate
B
Ear
Shoot
717
95
802
C
Gene A
GeneB
tss
tts
tss
tts
Motif number
Ear:275
Shoot:314
Ear:62
Shoot:88
Ear:108
Shoot:97
Ear:19
Shoot:35
Ear:347
Shoot:180
D
Ear
E
Shoot
F
Ear
Expression
Expressed
Highly-expressed
Number of combined motif
G
Shoot
Expression
Expressed
Highly-expressed
Number of combined motif

Figure 3

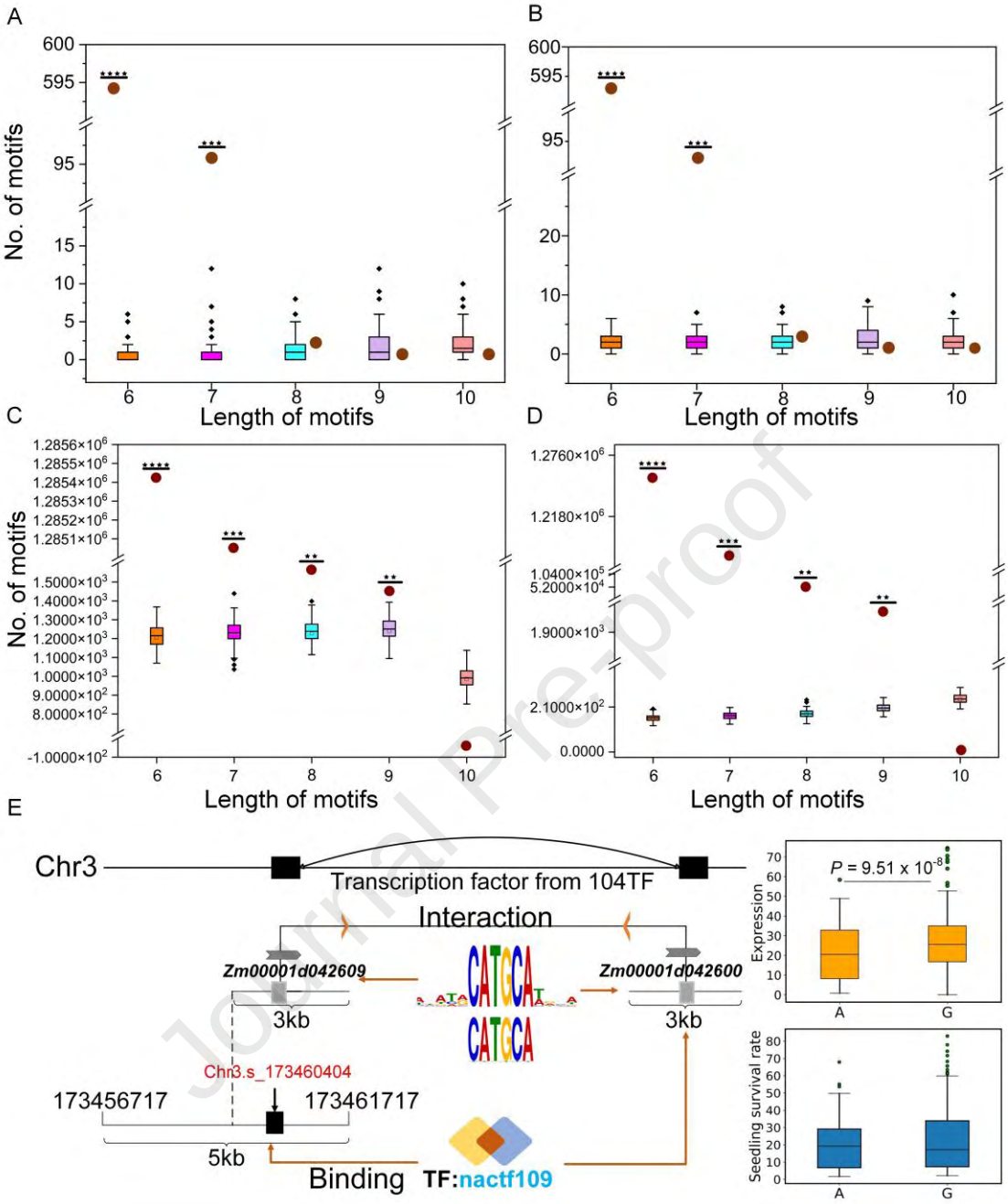

A
B
C
D
E
No. of motifs
Length of motifs
Chr3
Transcription factor from 104TF
Interaction
Zm00001d042609
Zm00001d042600
3kb
3kb
Chr3.s_173460404
173456717
173461717
5kb
Binding
TF:nactf109
P = 9.51 x 10-8
Expression
Seedling survival rate
A
G

Figure 4

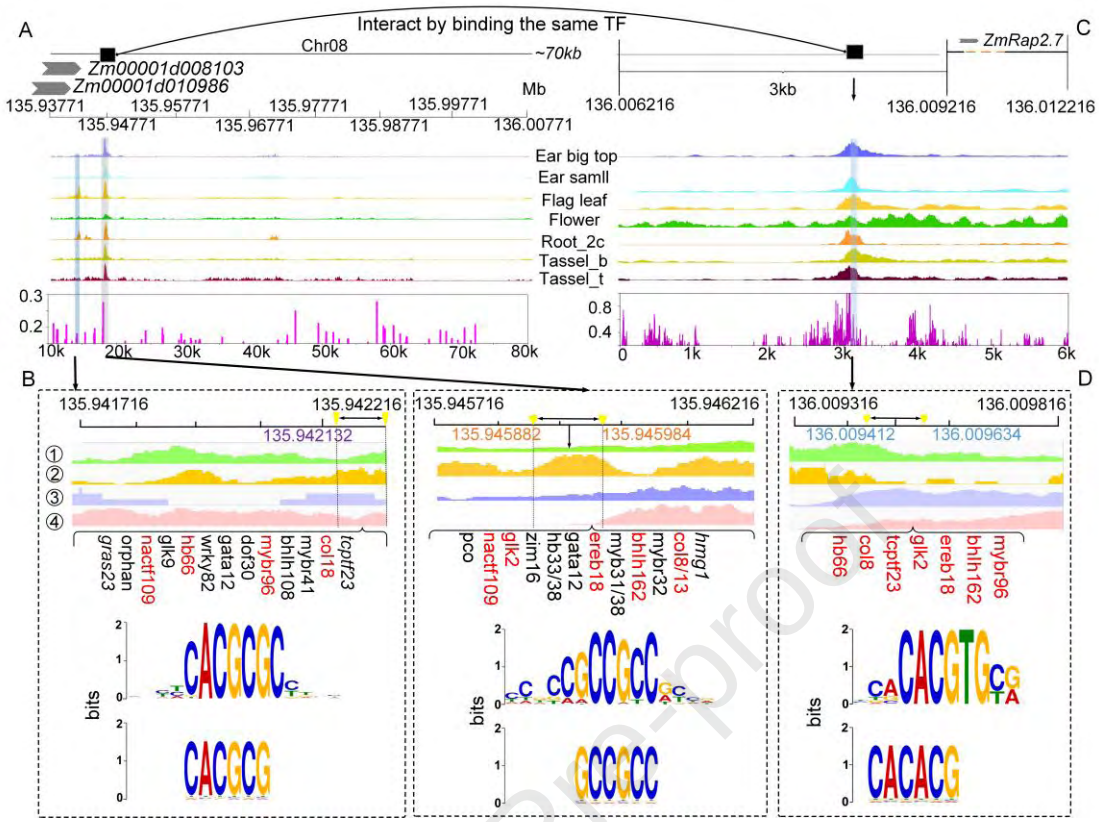


Figure 5

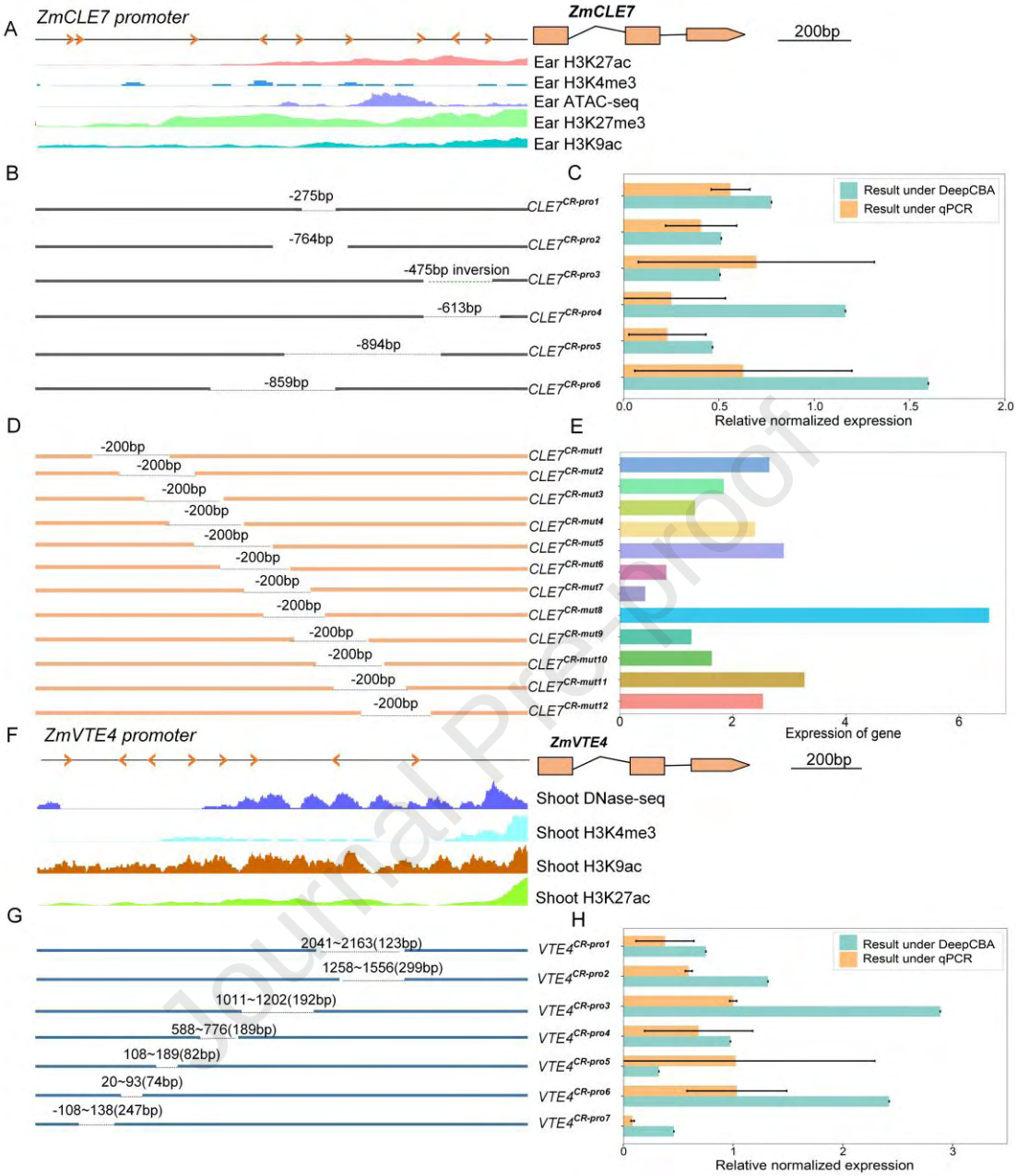

A
ZmCLE7 promoter
ZmCLE7
200bp
Ear H3K27ac
Ear H3K4me3
Ear ATAC-seq
Ear H3K27me3
Ear H3K9ac
B
-275bp
-764bp
-475bp inversion
-613bp
-894bp
-859bp
CLE7CR-pro1
CLE7CR-pro2
CLE7CR-pro3
CLE7CR-pro4
CLE7CR-pro5
CLE7CR-pro6
C
Result under DeepCBA
Result under qPCR
0.0
0.5
1.0
1.5
2.0
Relative normalized expression
D
-200bp
CLE7CR-mut1
CLE7CR-mut2
CLE7CR-mut3
CLE7CR-mut4
CLE7CR-mut5
CLE7CR-mut6
CLE7CR-mut7
CLE7CR-mut8
CLE7CR-mut9
CLE7CR-mut10
CLE7CR-mut11
CLE7CR-mut12
E
0
2
4
6
Expression of gene
F
ZmVTE4 promoter
ZmVTE4
200bp
Shoot DNase-seq
Shoot H3K4me3
Shoot H3K9ac
Shoot H3K27ac
G
2041~2163(123bp)
1258~1556(299bp)
1011~1202(192bp)
588~776(189bp)
108~189(82bp)
20~93(74bp)
-108~138(247bp)
VTE4CR-pro1
VTE4CR-pro2
VTE4CR-pro3
VTE4CR-pro4
VTE4CR-pro5
VTE4CR-pro6
VTE4CR-pro7
H
Result under DeepCBA
Result under qPCR
0
1
2
3
Relative normalized expression

Figure 6

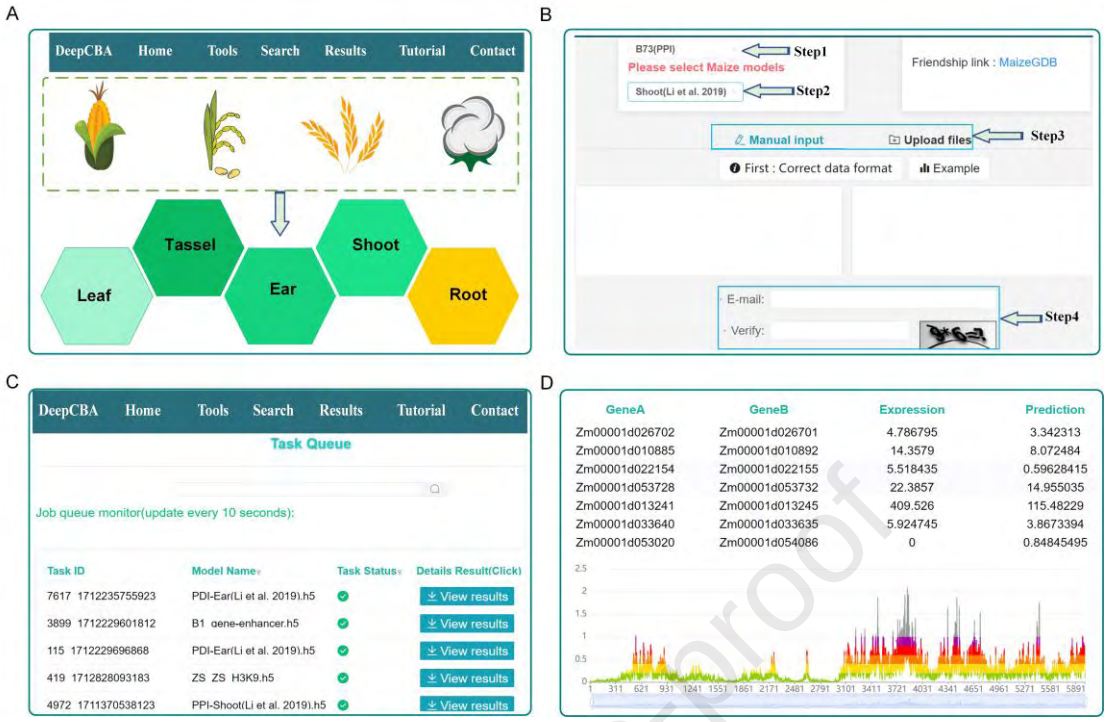
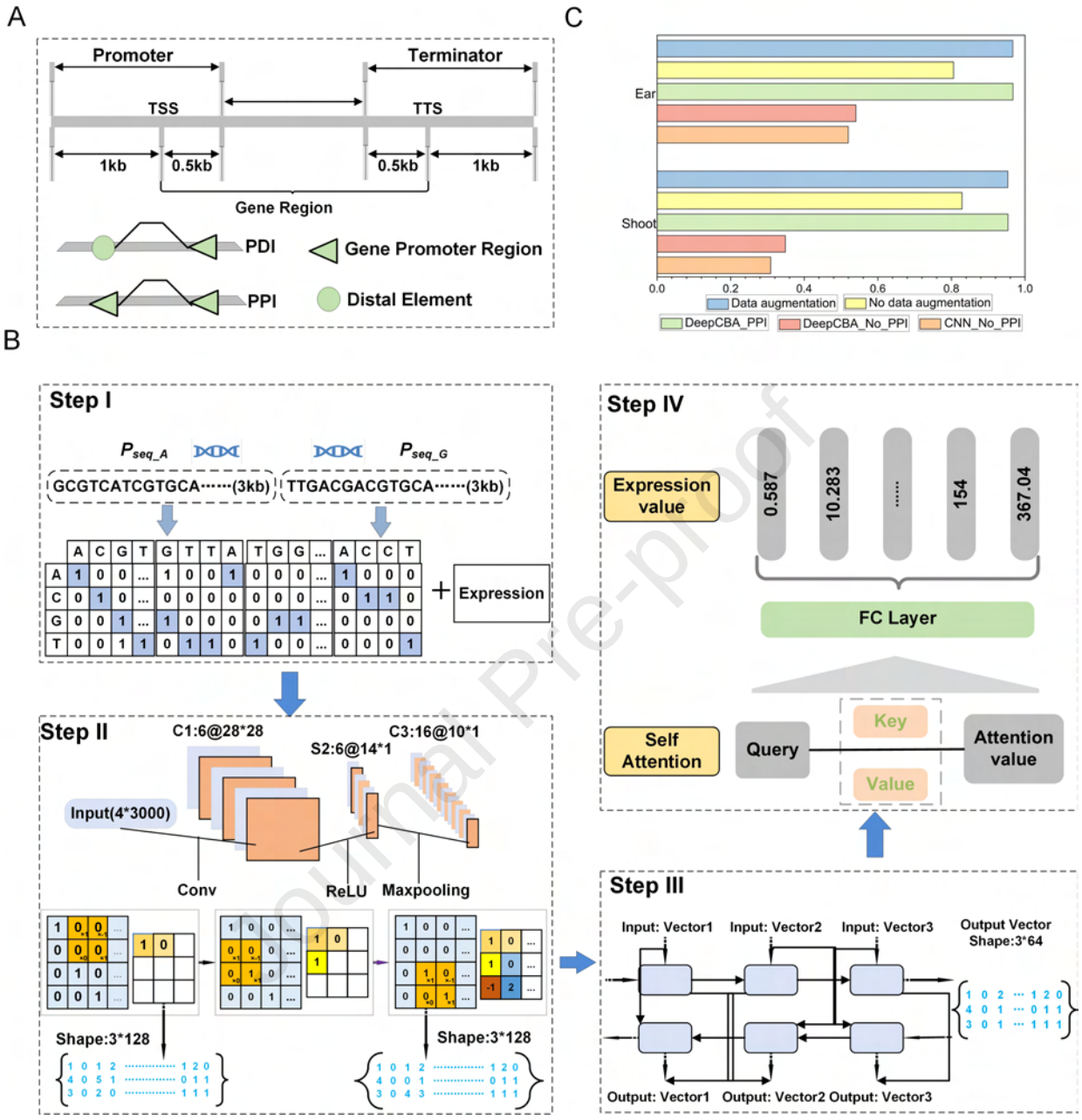
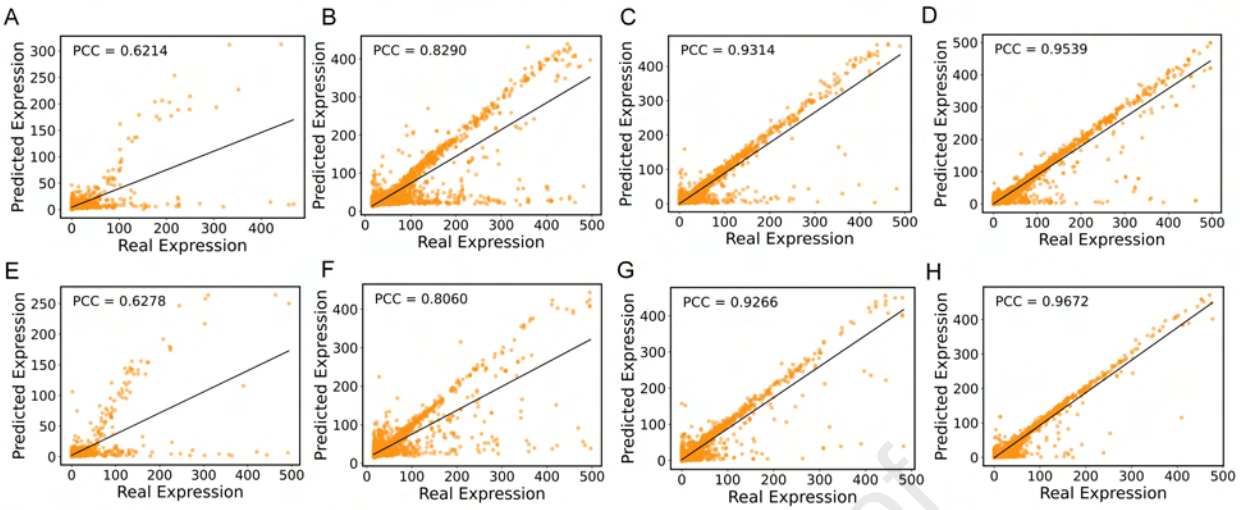
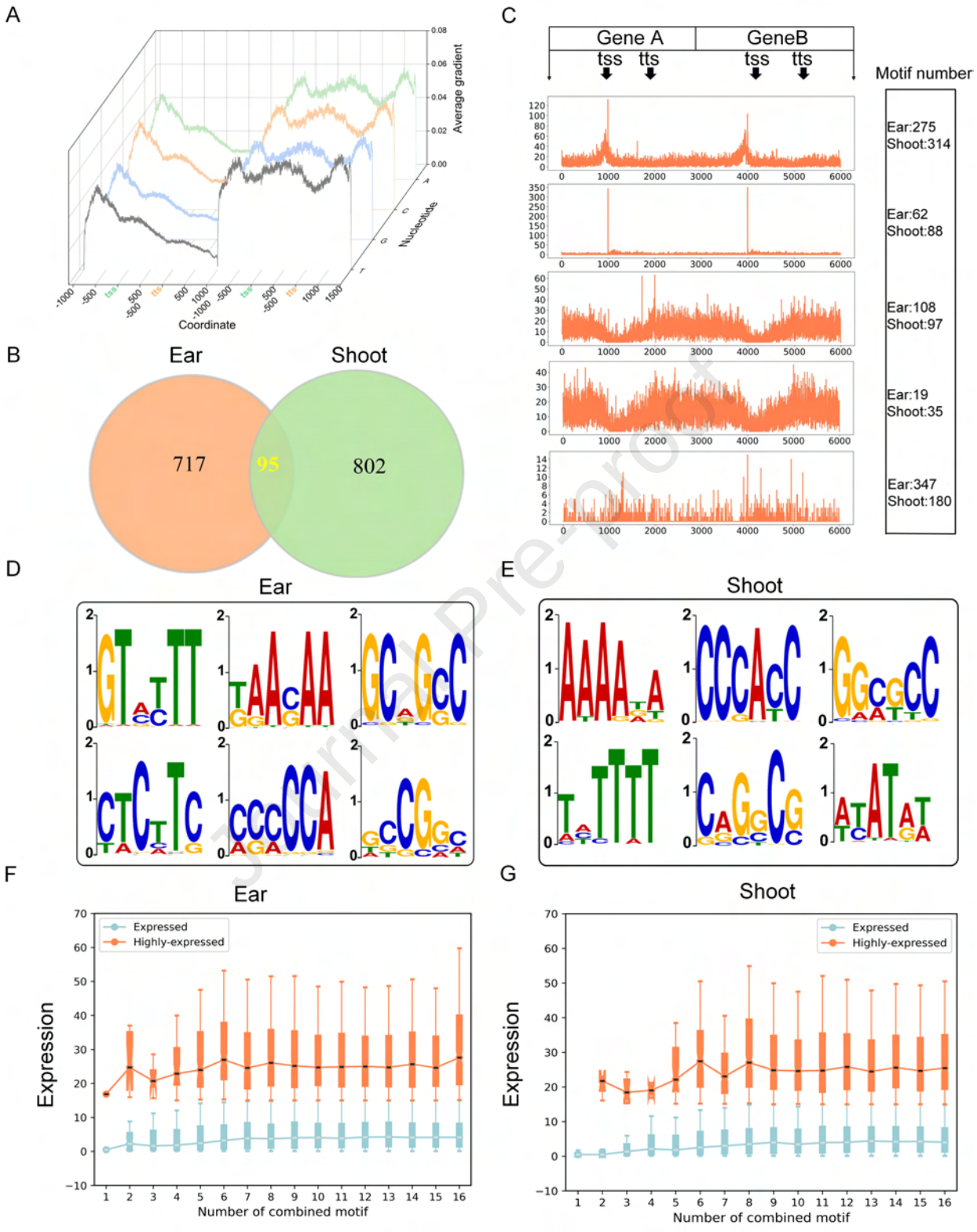
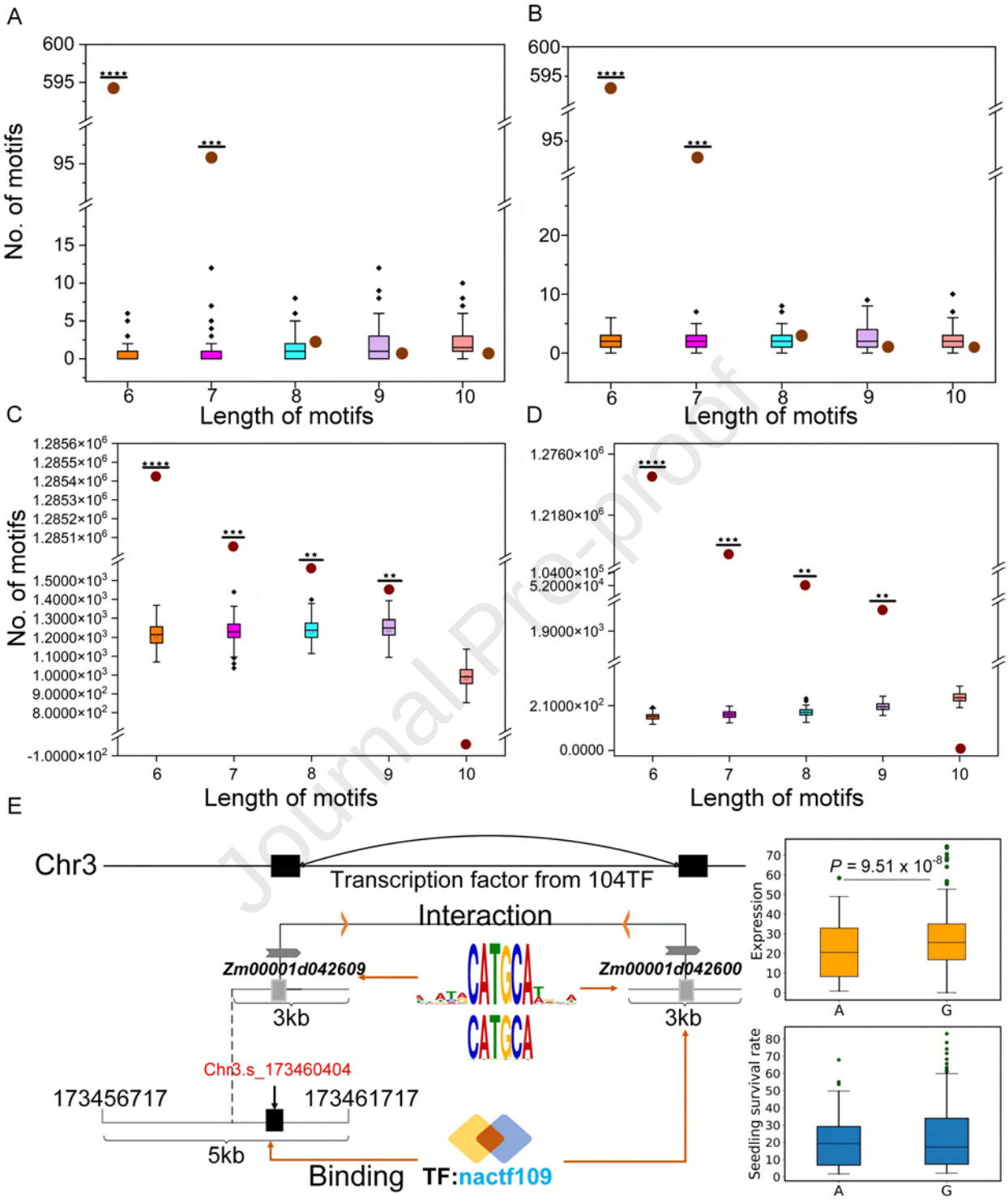


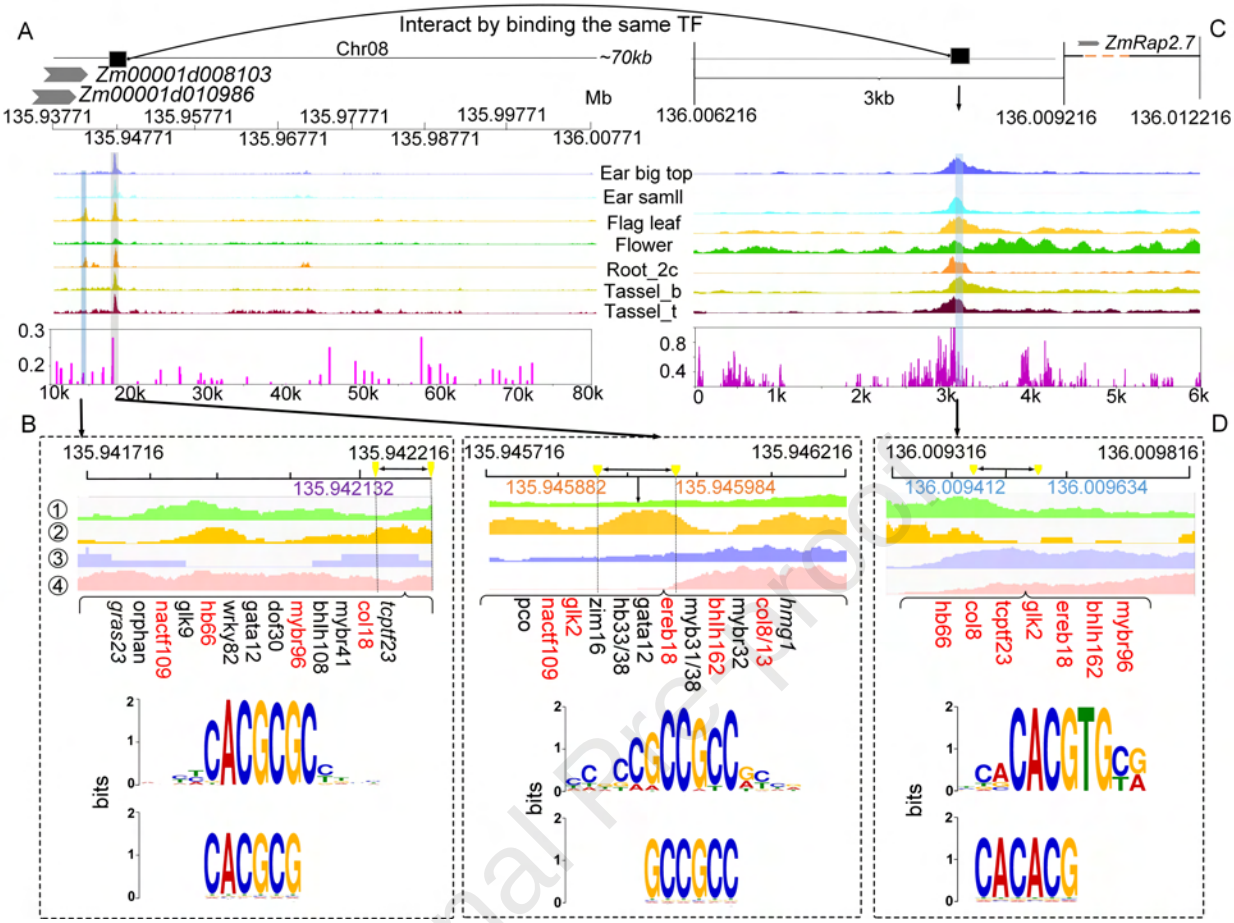
Figure 7

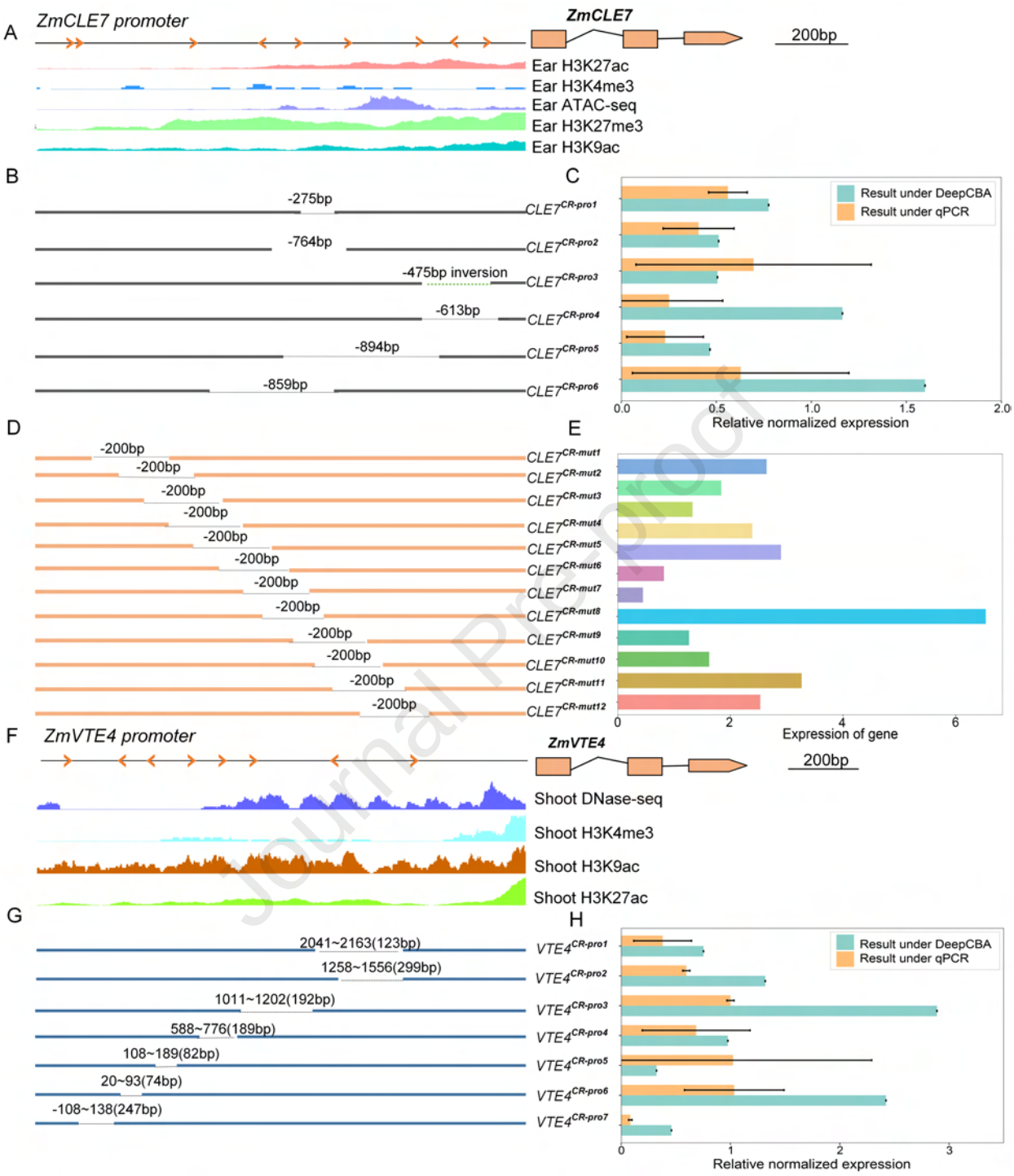




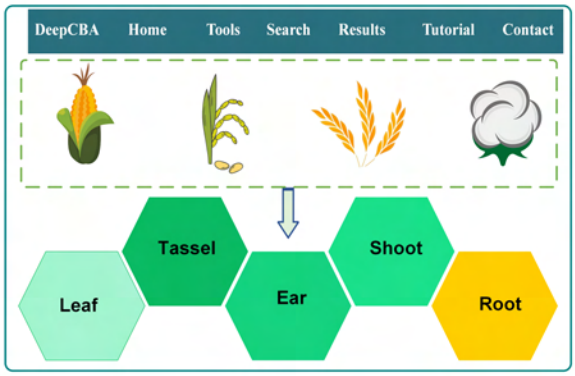




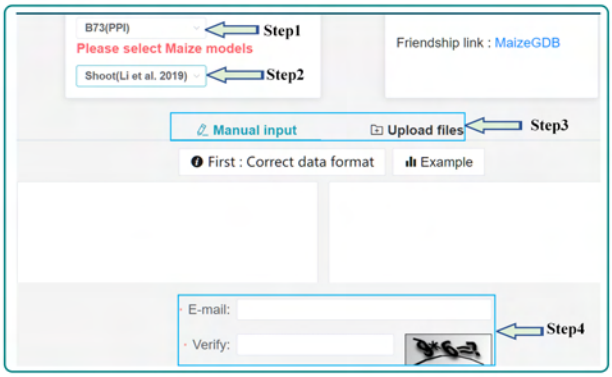




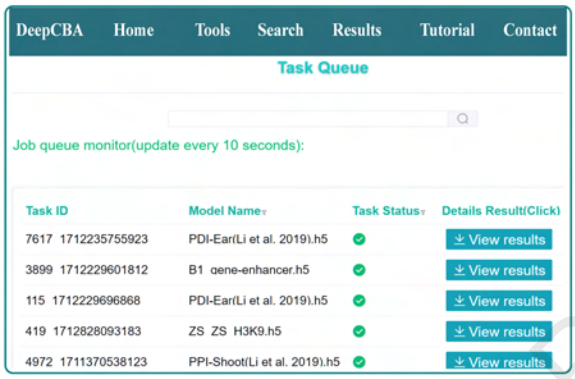
A



B



C



D