https://doi.org/10.1186/s13059-025-03516-7

Genome Biology

RESEARCH

Open Access

Precise engineering of gene expression by editing plasticity

Yang Qiu[1,2†], Lifen Liu[2,3†], Jiali Yan[1,2†], Xianglei Xiang[2,3], Shouzhe Wang[1,2], Yun Luo[1,2], Kaixuan Deng[2,3], Jieting Xu[4], Minliang Jin[4], Xiaoyu Wu[4], Liwei Cheng[4], Ying Zhou[5,6], Weibo Xie[1,2], Hai-Jun Liu[7], Alisdair R. Fernie[8], Xuehai Hu[2,3*] and Jianbing Yan[1,2,7*]

†Yang Qiu, Lifen Liu, Jiali Yan have equal contributions.

*Correspondence:
huxuehai@mail.hzau.edu.cn;
yjianbing@mail.hzau.edu.cn

[1] National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China
[2] Hubei Hongshan Laboratory, Wuhan 430070, China
[3] College of Informatics, Agricultural Bioinformatics Key Laboratory of Hubei Province, Huazhong Agricultural University, Wuhan 430070, China
[4] WIMI Biotechnology Co., Ltd., Sanya, Hainan 572000, China
[5] Institute of Agricultural Sciences of Xishuangbanna Prefecture of Yunnan Province, Jinghong, Yunnan 666100, China
[6] The Expert Workstation of Jianbing Yan in Yunnan Province, Jinghong, Yunnan 666100, China
[7] Yazhouwan National Laboratory, Sanya 572024, China
[8] Department of Molecular Physiology, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm 14476, Germany

Abstract

Background: Identifying transcriptional *cis*-regulatory elements (CREs) and understanding their role in gene expression are essential for the precise manipulation of gene expression and associated phenotypes. This knowledge is fundamental for advancing genetic engineering and improving crop traits.

Results: We here demonstrate that CREs can be accurately predicted and utilized to precisely regulate gene expression beyond the range of natural variation. We firstly build two sequence-to-expression deep learning models to respectively identify distal and proximal CREs by combining them with interpretability methods in multiple crops. A large number of distal CREs are verified for enhancer activity in vitro using UMI-STARR-seq on 12,000 synthesized sequences. These comprehensively characterized CREs and their precisely predicted effects further contribute to the design of in silico editing schemes for precise engineering of gene expression. We introduce a novel concept of "editingplasticity" to evaluate the potential of promoter editing to alter expression of each gene. As a proof of concept, both exhaustive prediction and random knockout mutants are analyzed within the promoter region of *ZmVTE4*, a key gene affecting α-tocopherol content in maize. A high degree of agreement between predicted and observed expression is observed, extending the range of natural variation and thereby allowing the creation of an optimal phenotype.

Conclusions: Our study provides a robust computational framework that advances knowledge-guided gene editing for precise regulation of gene expression and crop improvement. By reliably predicting and validating CREs, we offer a tool for targeted genetic modifications, enhancing desirable traits in crops.

Keywords: CRE, Deep learning, UMI-STARR-seq, Precise regulation, Insilico editing scheme, Editing plasticity

Background

Phenotypes are largely associated with gene expression, which is mainly determined by two factors: *cis*-regulatory elements (CREs) and *trans*-acting factors [1]. Over the past two decades, quantitative trait locus (QTL) analysis and genome-wide association studies (GWAS) in plants have revealed that a substantial proportion of trait variation can be attributed to functional regulatory variants. For example, regulatory variation typically accounts for over half of the QTLs observed in maize and tomato [2]. This suggests that genetic engineering of *cis*-regulatory mutations holds great promise for elucidating how regulatory variants affect phenotypic variation and crop improvement.

Recently, researchers have begun to employ gene editing technology to target known CREs and in doing so have generated edited variants conferring altered phenotypes suitable for crop improvement [3–6]. Indeed, by fine-tuning target gene expression, regulatory edits generally induce subtle phenotypic changes, and are therefore often considered more feasible for crop breeding compared to coding mutations. For instance, a 4-bp deletion within a known silencer downstream of *SlWUS* gene increased fruit locule number in tomato [4], while knockout of a 698-bp promoter region of *ZmFCP1* increased maize yield [3].

Achieving such precise gene regulation in plants, however, requires a priori and high-resolution genome-wide map of CREs. While combined epigenomic signals such as open chromatin and histone modifications have been used to indirectly infer CREs [7], these approaches only provide broad genomic regions (hundreds to thousands of bases) that are insufficient for high-resolution identification and precision editing. Although an exhaustive search or random editing can be applied, such as a tiling-deletion-based CRISPR–Cas9 screen [5], this is currently infeasible and overly costly, especially for crops with large genomes including maize and tomato. A recent pioneering study used a convolutional neural network (CNN) to build a classification model with a 3K-bp input in order to predict if a gene was expressed or not [8]. This study thus opened up a new avenue for in silico identification of CREs. A number of deep learning architectures have been developed for predicting gene expression in humans and their predictive performance has been continuously improved with longer input sequences and corresponding modularization of model architecture [9–13]. Beyond accurately characterizing genome-wide CREs, it would be more valuable to determine whether this knowledge could enable in silico saturation mutagenesis of individual or combined CREs (beyond single base mutations), aiming to explore the extensive spectrum of potential expression changes for a given gene. In principle, the goal is to create predictive models that elucidate the general rules of CRE-mediated gene regulation, paving the way for rational and quantitative gene editing.

In this study, we provide a novel and systematic solution including both CRE identification and editing (Fig. 1). Specifically, we built two sequence-to-expression deep learning models with long (120K-bp) and short (3K-bp) inputs of genomic sequence in the four plant model species of maize, rice, tomato, and Arabidopsis thaliana, and combined the interpretability methods to accurately identify distal and proximal CREs in a genome-wide manner. These models further enabled us to introduce a new concept, “editing plasticity (EP),” as a means by which to theoretically estimate the potential for expression changes due to promoter editing. This newly developed tool was applied

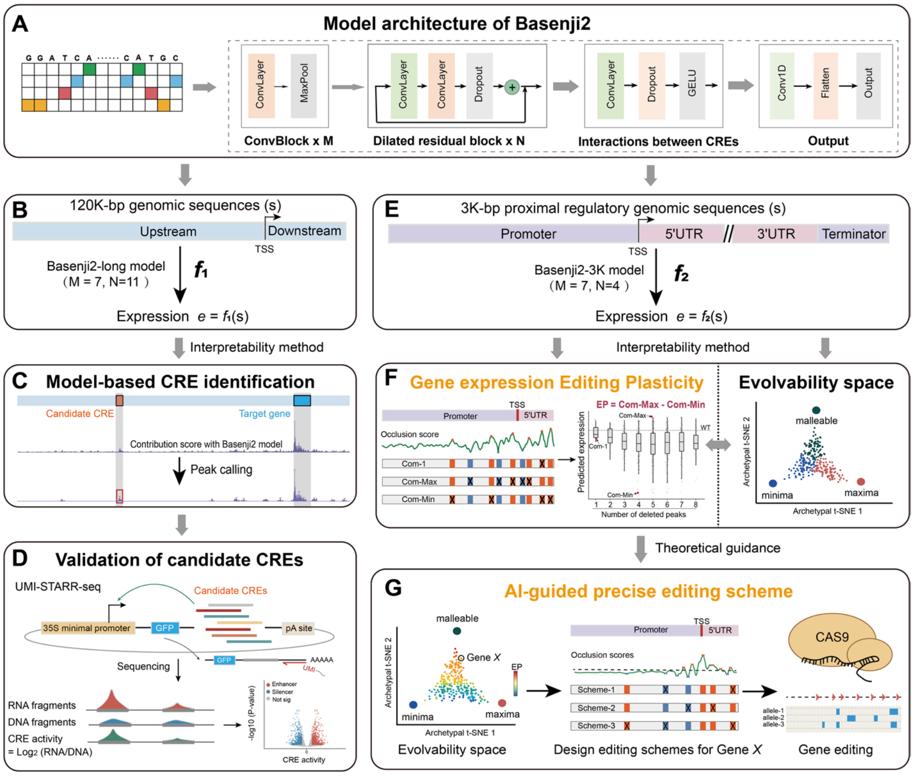

Fig. 1 Project overview. **A** Model architecture of Basenji2. The architecture of Basenji2 contains M convolution blocks and N dilated residual blocks in sequence. **B** Sequence-to-expression deep learning model with a long input. For each gene, genomic sequences were extracted 100-Kb upstream and 20-Kb downstream of its TSS. The Basenji2-long model successively contains seven convolution blocks (M = 7) and eleven dilated residual blocks (*N* = 11). **C** Model-based identification of CRE. For each genomic sequence, a deep interpretability method estimates a contribution score for each base and then obtains a contribution score vector of equal length as input. A peak-calling algorithm is used to identify candidate CREs from the vector. (D) Validation of candidate CREs. UMI-STARR-seq is used to measure the activities of model-identified candidate CREs. **E** Sequence-to-expression deep learning model with short input. For each gene, the proximal regulatory sequences were used as the input, including the promoter, 5'UTR, 3'UTR, and terminator sequences. The Basenji2-3K model successively contains seven convolution blocks (M = 7) and four dilated residual blocks (*N* = 4). **F** Theoretical guidance for gene editing. Editing plasticity estimates the expression changes of simulated deletions. Evolvability space estimates the expression changes of simulated single-nucleotide mutations and displays three distinct patterns. Both reflect the gene editing potential. **G** AI-guided precise editing scheme. Leveraging the tools of editing plasticity and evolvability space, AI designs precise editing schemes for genes with editing potential for precise regulation and crop genetic improvement with CRISPR-Cas9

in an empirical study of *ZmVTE4*, a key gene affecting α-tocopherol content in maize [14]. We generated an edited-allele with a novel (i.e. not present in the observed population) 4-bp deletion within its 5'UTR region under the guidance of the AI-guide editing scheme. This edited-allele was experimentally validated to significantly increase *ZmVTE4* gene expression and α-tocopherol content in kernels in vivo. Our work provides a quantitative estimate of the editing plasticity of each plant gene and provides a detailed and explicit roadmap to guide future gene editing experiments.

Results

Accurate prediction of gene expression from DNA sequence in multiple plant species

We first asked whether deep learning could accurately model the sequence-to-expression relationships in plants. For this purpose, we used an integrated gene expression dataset containing 421 RNA-seq datasets with the maximum expression level across multiple tissues as the prediction target (see the "Methods" section; Additional file 3: Table S1), which was used previously [8]. We adopted the model architecture of Basenji2 (Fig. 1A, Additional file 2: Fig S1), a top-performing framework for predicting transcriptional profiles in humans [10], for modeling long genomic sequence as input with an expectation of incorporating more regulatory elements. Using an independent test dataset (see the "Methods" section), prediction performance was evaluated via the Pearson correlation coefficients (PCC) between predicted and observed measurements across various input sequence lengths up to 140K-bp (Additional file 4: Supplementary Material 1).

Initially, we directly used the model architecture of Basenji2 as well as its original hyperparameters (Additional file 3: Table S2) [10] using the 10K-bp input data in maize (Additional file 4: Supplementary Material 1), and obtained a good prediction performance with a PCC of 0.660 on the independent testing dataset, preliminarily demonstrating the feasibility of prediction. To further improve prediction performance, we optimized an important hyperparameter of "Channel Number" (CM, other hyperparameters are mostly determined by CM) in Basenji2 [10] by using a 5-foldCV (five-fold cross-validation) on the training dataset, with all other hyperparameters retained. A small promotion of PCC on the independent testing dataset from 0.660 to 0.6776 was found in this optimization with the optimal CM of 720 (close to the original CM of 768 [10]), which was fixed in all the following modelings in this study (Additional file 3: Table S2). Discussions on its advancement and comparison with other existing methods can be found in Additional file 2: Fig S2 and Additional file 4: Supplementary Material 1.

When involving longer inputs from 10K-bp to 140K-bp, prediction accuracy gradually increases and reaches an optimal at 120K-bp, 100K-bp upstream plus 20K-bp downstream of TSS (referred to Basenji2-long model), with a PCC of 0.733 (Fig. 1B and the blue curve in Fig. 2A). This demonstrates that complex regulation can be accurately captured within large genomic contexts. Different optimal lengths were also examined for the three other model plant species studied: rice (40K-bp), tomato (80K-bp), and Arabidopsis thaliana (5K-bp), and all achieved good prediction accuracy (PCC ranged between 0.641 and 0.816; Fig. 2A, Additional file 2: Fig S2).

Constructing a genome-wide regulatory map with deep interpretability

We next asked which sequence elements within the entire genomic sequence input are important for high-precision prediction, given that the regions with high interpretative importance are likely to reflect key regulatory elements. For this task, model interpretability is appropriate to prioritize important features during prediction and is essential in biological studies [13, 15]. We adopted a backpropagation-based interpretability method of input gradients [12, 13] and employed a peak-calling algorithm to form candidate CREs (see the "Methods" section; Fig. 1C, Additional file 2: Fig S3). A total of 745,684 candidate CREs (594.7M) associated with 45,564 target genes were identified,

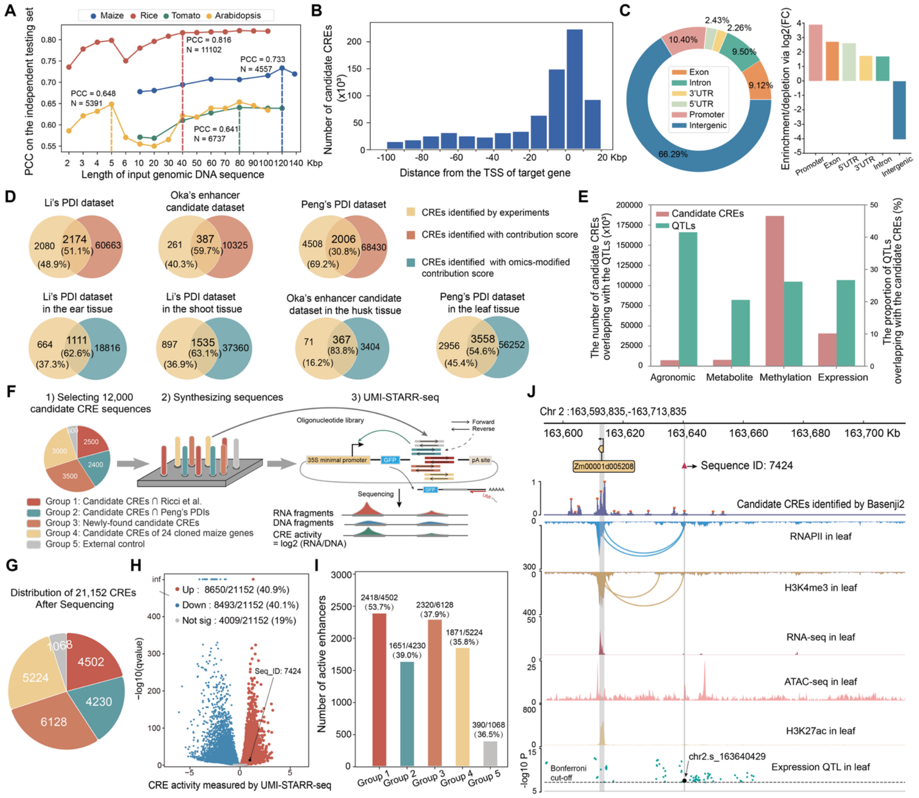

Fig. 2 Genome-wide regulatory map and experimental validations. **A** Basenji2-long model prediction performance evaluated by Pearson correlation coefficient (PCC) across four plant species, using different DNA input lengths to select the optimal length based on independent test sets (N denotes its sample size). **B** Distance distribution of candidate CREs from the TSS of its target genes identified by the Basenji2-long model of maize. **C** Overlap proportions and enrichment fold of candidate CREs across five genomic components. **D** Reappearance rates of candidate CREs in three experimental datasets, comparing CRE-gene pairs with experimentally identified CREs (top), and tissue-specific candidate CREs identified by the omics-modified gradient × input contribution score (see the "Methods" section) (bottom panel). **E** Overlap analysis of candidate CREs with four types of QTLs. Red bars: number of overlapping CREs, green bars: proportion of QTLs overlapping with CREs. **F** Selection, synthesis, and UMI-STARR-seq screening of candidate CRE sequences in maize protoplasts. **G** Distribution of 21,152 CREs after sequencing, showing total sequences (forward and reverse orientations) in each group, corresponding to Fig. 2F. **H** CRE activities of synthesized candidate CRE sequences. A volcano plot of log2FC scores (i.e., CRE activity) against –log10-transformed BH adjusted *P*-value from DESeq2 is shown. The CRE activity of a candidate CRE sequence (ID:7424) of *Zm00001d005208* is highlighted. **I** Number and percentage of 8650 enhancers in five groups. **J** A representative enhancer validated for *Zm00001d005208*. The first panel shows the genomic positions of *Zm00001d005208* and the enhancer (seq ID: 7424). The second panel displays the gradient × input contribution score from the Basenji2-long model, highlighting 14 candidate CREs (marked in orange at the peak summits). Subsequent panels show genomic tracks of chromatin interactions, RNA-seq, ATAC-seq, H3K27ac, and eQTLs for the region, with eQTL (chr2.s_163640429) within the candidate enhancer highlighted

accounting for nearly 28.3% of the maize genome (Additional file 5: Supplementary Material 2). We repeated this analysis for all four plant species and provided freely available genome-wide CRE maps at [16].

As expected, most of the candidate CREs (50.4%) are in TSS-proximal regions between ± 10K-bp (Fig. 2B). Interestingly, distal regions also exhibit a considerable number of candidate CREs, being slightly enriched in the upstream 70K-bp to 60K-bp. This region exactly harbors well-known distal enhancers including *Vgt1*, which regulates *ZmRap2.7* located ~70Kb downstream [17], and the enhancer derived from a TE insertion for *tb1*, which is also located 60–70Kb upstream [18]. The fact that both distal enhancers were effectively identified in our analysis acts as a strong proof-of-concept of our approach (Additional file 2: Fig S4).

We next collected the reference intervals of six main components, including promoter (2K-bp upstream of TSS), exon, intron, 5’UTR, 3’UTR, and intergenic region, and intersected the candidate CREs (total size of 594M) with the respective reference interval. Interestingly, although a large proportion of the candidate CREs (66.29%, 404.86M) were located in intergenic regions (Fig. 2C, Additional file 3: Table S3), they were significantly enriched in promoter (totally 63.54M, Enrichment fold = 14.82, p-value = 1.52E − 24), 5’UTR (14.83M, Enrichment fold = 6.07, p-value = 1.16E − 4), and 3’UTR (13.83M, Enrichment fold = 3.36, p-value = 4.31E − 3) regions implying an important regulatory function of these regions.

To systematically evaluate the biological relevance of the distal CRE candidates identified in this study, we compared them with experimentally inferred distal regulatory elements from three independent maize studies: (a) Oka’s enhancer dataset that was classified with features of low DNA methylation, high chromatin accessibility, and H3K9ac enrichment [7]; (b) Li’s proximal–distal interactions (PDI) from H3K4me3-ChIA-PET and H3K27ac-ChIA-PET [19]; (c) Peng’s PDI from RNAPII-ChIA-PET [20]. The PDI is commonly used to reflect enhancer-promoter-interactions. Generally, 59.72% (387/648) of Oka’s enhancers, 51.15% (2174 of 4254) of Li’s distal enhancer candidates, and 30.83% (2006/6514) of Peng’s distal regulatory elements were consistently identified in the present study (Fig. 2D, Additional file 3: Table S4, Additional file 5: Supplementary Material 2). In total, 6649 candidate distal elements were supported by at least one of these datasets. These results suggest that our model, trained only on DNA sequences, is effective in identifying distal CRE candidates.

We next propose to introduce tissue-specific epigenomic signals to mine tissue-specific candidate CREs. We collected available chromatin accessibility [7, 21], and DNA methylation data [22] and designed a modified score with a weighted sum of original contribution score, chromatin accessibility, and DNA methylation (see the “Methods” section) in order to identify tissue-specific distal elements (Additional file 6: Supplementary Material 3). This approach led to the reappearance rates being increased from 51.1 to 62.59% (1111 of 1775 in leaf) and 63.12% (1535 of 2432 in shoot), respectively (Fig. 2D). For the other two datasets, the reappearance rates were also increased (Additional file 3: Table S5). This implies that the model has learnt some tissue-specific information, as the maximum gene expression often comes from the tissue where it specifically functions for most tissue-specific genes. This information was, moreover, further enhanced with epigenomic data from the corresponding tissue.

Overlapping the candidate CREs with existing QTL information provides hints to their potential function

By examining the consistency between newly identified and existing epigenomic-inferred CREs [7, 17, 23, 24], we demonstrate that candidate CREs predominantly overlap with H3K9ac-marked distal enhancers, effectively replicating the majority of enhancers defined by H3K9ac epigenomic signals (Additional file 2: Fig S5A-B). To bridge these regulatory elements with functional traits, we next systematically investigated the potential biological functions of the candidate CREs by overlapping them with existing QTLs mapped in a pan-Zea population including 721 pan-Zea individuals [25]. Surprisingly, we observed that nearly half of the agronomic QTLs (13,987 of 33,679, or 41.53%) overlapped with the model-identified candidate CREs (Fig. 2E, Additional file 2: Fig S5C), supporting that many of the candidate CREs (7410, Additional file 3: Table S7) might affect agronomic traits. We also observed a 26.7% overlap of eQTL (168,941 out of 632,930 eQTLs; Fig. 2E), further highlighting the potential regulating function of the overlapping CREs (27.8%, 40,585 out of 146,042 CREs). These results provide substantial evidence for the involvement of a significant portion of candidate CREs in both trait variation and gene expression control, while implying that the remaining CREs that do not meet the standards of statistical testing might contribute in more subtle or context-specific ways.

Validating the candidate CREs with UMI-STARR-seq

In addition to the accurate identification of functional CRE sequences, a reliable estimation of their corresponding effects is essential for the prediction of gene expression. Here, we leveraged UMI-STARR-seq, a method that integrates unique molecular identifiers (UMIs) with self-transcribing active regulatory region sequencing (STARR-seq) to quantify CRE activities on a large scale [26], enabling us to experimentally verify the predicted effect of CREs (Fig. 1D). A batch of 12,000 candidate CREs (of 200-bp each) were selected and synthesized, and their activities were assessed in duplicate in maize protoplasts (see the "Methods" section; Fig. 2F, Additional file 7: Supplementary Material 4). Specifically, these candidates consist of five groups: group 1 containing 2500 candidate CREs overlapping with enhancers previously identified by STARR-seq [21]; group 2 with 2400 candidate CREs coinciding with existing PDIs [20]; group 3 with 3500 candidate CREs newly identified in this study consisting of 3000 distal (beyond 5K-bp from TSS) and 500 proximal (within 5K-bp of TSS) sequences with top contribution scores; group 4 contains 3000 candidate CREs associated with 24 cloned maize genes influencing key phenotypes, identified by tiling each 1K-bp candidate CRE into five 200-bp windows; and group 5 including 600 external controls with zero-value activity in a previous study [21] (see Additional file 1: Supplementary Methods).

For each sequence, the plasmid library allowed insertion in either the forward or reverse orientation. These two orientations were treated as independent sequences to calculate separate activity values, resulting in ~ 24,000 sequences to be analyzed. Both unique RNA (cDNA) reads and DNA input reads were mapped to the reference set, and the CRE activity was calculated as log2(cDNA/input) for each orientation separately using DEseq2 [27]. Out of the candidate CREs, 21,152 sequences with at least 10 uniquely aligned DNA reads (accounting for both orientations) were included in the analysis (Fig. 2G). Among these, 8650 sequences (40.9%) exhibited enhancer activity (CRE activity > 0 and adjusted P-value ≤ 0.05, Fig. 2H). When we examined the distribution of enhancers across the five groups (Fig. 2I), Group 1 exhibited the highest proportion of active enhancers (53.7%, 2418 out of 4502), consistent with prior studies [21]. Group 2 followed with 39.0% (1651 out of 4,230), suggesting these enhancers may regulate expression through a chromatin interaction mechanism. Group 3 showed 37.9% (2320 out of 6,128), representing a significant new finding in this study. Group

4, associated with important cloned maize genes, demonstrated 35.8% (1871 out of 5224) active enhancers, which are critical for regulating gene expression. This implies that genetic manipulation of these elements using gene editing tools could have desired effects on both gene expression and agronomic traits.

To demonstrate the utility of our framework, we highlight a candidate enhancer (Fig. 2J; seq ID: 7424) targeting *Zm00001d005208*, encoding a NAC transcription factor, a family known for their roles in regulating plant development and stress responses [28]. The model identified 14 peaks, and we selected the 200-bp sequence with the highest contribution score (Sequence ID: 7424) within the 12th peak (Zm00001d005208-CRE-3), supported by PDI evidence, for inclusion in Group 2 for experimental validation. This candidate CRE exhibited an activity score of 1.81 (Fig. 2G) and is supported by multiple epigenomic evidences: (a) strong interaction with the *Zm00001d005208* promoter, as indicated by ChIA-PET RNAPII data, suggesting active transcriptional regulation; (b) localization in an open chromatin region (ATAC-seq); (c) marked by the enhancer-associated H3K27ac modification; and (d) supported by an expression QTL (chr2.s_163640429) in leaf tissue, linking it to expression variation. These findings underscore the enhancer's potential to regulate *Zm00001d005208* expression, showcasing the strength of our framework in identifying key regulatory elements with high relevance for crop improvement.

High-resolution mapping of TSS-proximal CREs

Above, we constructed a genome-wide regulatory map by identifying 745,684 candidate CREs, however, the resolution of these CREs is limited to 1K-bp which falls short of the precision needed for accurate regulation. Since the genomic sequence proximal to TSS can account for the majority of the explained variance (EV) of gene expression, we narrowed our focus to the identification of TSS-proximal CREs with higher resolution in order to capture finer details within this critical region.

Different input combinations were tested (Additional file 3: Table S10) by evaluating the PCC and EV in relation to gene expression. The optimal configuration was found to be the 3K-bp input combination (Basenji2-3K-B73 model, Fig. 1E), with a PCC of 0.676 which is slightly lower than the 0.733 of the Basenji2-long model (Additional file 2: Fig S7B, Additional file 3: Table S10). This combination includes a 2K-bp promoter, 300-bp 5'UTR, 500-bp 3'UTR and 200-bp downstream of the transcription termination site (TTS) (Additional file 1: Supplementary Methods; Fig. 3A, Additional file 2: Fig S7A), which aligns with a previous finding that the combination of these regulatory regions defines gene expression levels [29]. Based on the same sequence composition, we also trained the Basenji2-3K models for three other model plants (Additional file 3: Table S11).

We next applied the Basenji2-3K-B73 model to identify TSS-proximal CREs using another interpretability method of "Occlusion" (see the "Methods" section) to compute the base contribution score for each gene in order to increase the resolution. As a general demonstration, the mean (purple curve) and variance (purple shade) of the contribution score at each base across all 37,979 genes are presented in Fig. 3A. The results showed the promoter and 5' UTR play critical roles in regulating gene expression, which might contribute to the maintenance or enhancement of expression

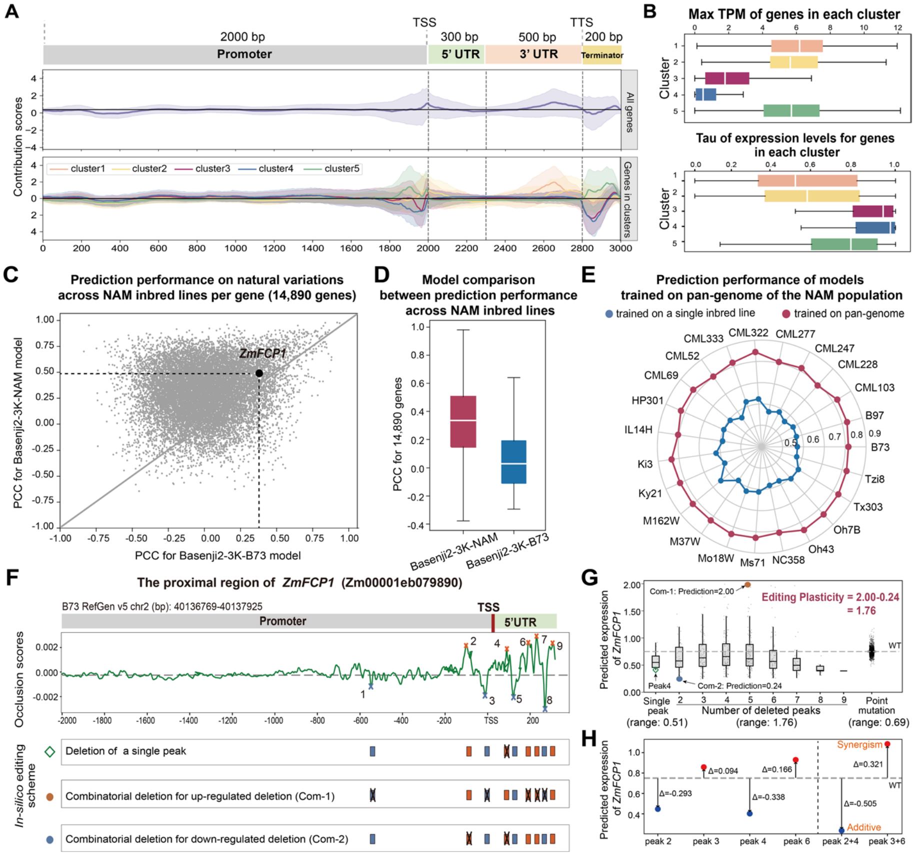

Fig. 3 Estimation of editing plasticity using the TSS-proximal model. **A** Base contribution scores from the Basenji2-3K model and gene cluster analysis. The input regulatory sequence (3K-bp) includes 2K-bp promoter, 300-bp 5'UTR, 500-bp 3'UTR and 200-bp terminator. The middle panel shows the mean (purple curve) and variance (purple shade) of the contribution scores across 37,979 genes. The bottom panel illustrates mean and variance values for each gene cluster. Genes were grouped into five clusters using K-medoids, revealing distinct regulatory patterns linked to expression levels and tissue specificity. This clustering approach helps identify gene groups with similar regulatory behaviors, providing insights into their functional roles in biological processes. **B** Expression and tissue specificity of clusters, showing maximum TPM (top) and Tau values (bottom). **C** Prediction performance for natural variations across NAM inbred lines per pan-gene, highlighting *ZmFCP1*. **D** Comparison of prediction performance between the Basenji2-3K B73 and Basenji2-3K-NAM models across NAM inbred lines. **E** Performance of single-line models versus the pan-genome-trained Basenji2-3K-NAM model evaluated on independent test sets. **F** High-resolution analysis of the TSS-proximal region of *ZmFCP1*. Occlusion contribution scores identified candidate CREs (with orange/blue color at the summit of each peak) within 2K-bp upstream and 300 bp downstream of the TSS (top panel). Peaks were selected as candidate regions for gene editing based on their occlusion contribution scores, which reflect the impact of perturbing these regions on gene expression. Three in silico editing schemes are presented (bottom panel). **G** Editing plasticity of *ZmFCP1*. The *x*-axis represents the three simulated scenarios when calculating editing plasticity: single peak deletion, combinatorial deletion of multiple peaks, and point mutation. The *y*-axis shows the predicted expression levels of simulated sequences. Predicted expression ranges are shown below the *x*-axis. **H** Interaction effects between CREs. Red dots indicate increased expression, blue dots indicate decreased expression, and Δ represents the deviation from wild-type expression

levels. Using cluster analysis with the K-Medoids method [30] and the silhouette coefficient as the optimization criterion, all genes were grouped into five distinct clusters (Additional file 2: Fig S8), exhibiting varying patterns of contribution scores (Fig. 3A). Cluster 1 (12,608 genes) and 2 (13,599 genes) have normal expression pattern with positive contribution in the promoter and 5’UTR, and negative contribution in the 3’UTR and terminator, correlating with relatively high expressions (Fig. 3B). Clusters 3 (5,194 genes), 4 (1,294 genes) and 5 (4,904 genes) show distinct patterns: clusters 3 and 4 are associated with high negative contributions, while cluster 5 has a high positive contribution in both TSS-proximal and TTS-proximal regions, leading to low (for cluster 3 and 4) or high (for cluster 5) expression levels. We further investigated the tissue-specific expression levels of the five clusters with the Tau metric to assess the tissue specificity of expression [31], where higher Tau values indicate more significant expression variability across different tissues. We found that genes from cluster 1 and 2 are likely housekeeping genes (displaying relatively low Tau values, Fig. 3B), while cluster 5 mainly contains tissue-specific genes.

The model can accurately predict expression change of natural variation

To demonstrate whether the Basenji2-3K-B73 model can computationally simulate the promoter editing events and predict their effects, we took the *ZmFCP1* gene, which acts in the CLAVATA-WUSCHEL feedback pathway, as an example, and five mutant alleles were obtained by CRISPR-Cas9 in the previous study [3]. The predicted expression on these five edited promoter alleles from our model showed high agreement (PCC = 0.88) with the actual experimental expression (Additional file 2: Fig S9). Another two examples of tomato *SlCLV3* gene [4, 32, 33] and rice *IPA1* gene [5] were also demonstrated (Additional file 2: Figs S10-11).

We next asked whether the Basenji2-3K-B73 model can accurately predict expression alteration of natural variation. For this purpose, we employed a pan-genome dataset (with full genome sequence and transcriptome data) from the 22 founders of the maize NAM (Nested Association Mapping) population [34], representing a wide breadth of maize genetic diversity. We first examined the consistency between predicted and observed expression of *ZmFCP1* across the 22 lines of the NAM population and found a low PCC of 0.38 (Fig. 3C), with the average PCC for all measured genes being only 0.011 (Fig. 3D).

With the expectation that integrating population-level genetic variation may improve the ability to predict gene expression across individuals, an enhanced model (Basenji2-3K-NAM) was trained by incorporating the 22 maize genomes (see the “Methods” section). This bolstered our training samples from 30,764 to 687,545 genes (Additional file 3: Table S12) and improved the PCC from a range of 0.5–0.6 to 0.79–0.84 (Fig. 3E, Additional file 3: Table S16). In terms of advantages, the Basenji2-3K-NAM model has a greater generalization ability by significantly promoting the predictability of natural variation for most (12,412/14,890 ≈ 83.36%) of the maize pan-genes annotated in all NAM lines (Fig. 3C, D) [35]. Furthermore, the Basenji2-3K-NAM model also has higher predictability for structural variants (SVs). For example, considering the expression of the *GL15* gene (encoding an AP2-like TF that promotes juvenile leaf transition of juvenile leaf and represses adult leaf in maize) [36] in 10 maize lines, the PCC of prediction jumped from 0.16 with the original Basenji2-3K-B73 model to 0.75 with the NAM-enhanced model (Additional file 2: Fig S13B). Systematic comparison across NAM inbred lines for 10,303 genes with SVs also showed a significant outperformance of the Basenji2-3K-NAM model (Additional file 2: Fig S13D). Interestingly, this is also the case for the majority of cloned genes controlling agronomic traits (Additional file 2: Fig S13E). Pan-genome-based training in rice additionally supports its benefit (Additional file 2: Fig S14).

To evaluate the model's generalization and transferability, we performed transfer learning on wheat (Triticum aestivum) and cotton (Gossypium hirsutum) (Additional file 2: Fig S15). We collected annotated genomes and corresponding RNA-seq datasets for both species and applied eight pre-trained models, including single-genome and pan-genome models for maize, rice, tomato, and Arabidopsis. The results demonstrated that the models exhibited moderate transferability to wheat, with Pearson correlation coefficients (PCC) ranging from 0.004 to 0.35, where the maize pan-genome model achieved the highest PCC of ~ 0.35. In contrast, the models showed almost no transferability to cotton, with PCC values close to zero. Notably, pan-genome models consistently outperformed single-material models in transferability. These findings suggest that pan-genome models are particularly effective for species with complex genomes like wheat, though further optimization through fine-tuning or species-specific re-training may be necessary for challenging cases such as cotton.

These results underscore the substantial improvements in prediction accuracy and model predictability, robustness, and transferability achieved by training with diverse genetic data.

In silico editing scheme can evaluate editing effects across the full profile of promoter variants

Encouraged by the high predictive accuracy for natural variation, we next aimed to use the Basenji2-3K-NAM model to assess novel variation that can be created by gene editing. This is particularly significant since it will enable us to assess the potential effects of any editing event on gene expression beyond the limited scope of natural variation. We propose the term 'editing plasticity' as a measure of the expression change potential of a given promoter variant created by in silico gene editing. The occlusion score from the Basenji2-3K-NAM model helped to identify a total of 377,385 proximal CREs (3.7M) with high-resolution (10-bp) associated with some 35,179 annotated genes (Additional file 8: Supplementary Material 5). Taking the *ZmFCP1* gene for demonstration purposes again, nine peaks with significant expression impact were identified in its promoter and 5' UTR (Fig. 3F, Additional file 3: Table S17). In silico deletion of these peaks followed by Basenji2-3K-NAM model predictions revealed that deletion of a single peak only slightly alters *ZmFCP1* expression (Fig. 3G). Although the range of the observed changes (0.51) is comparable to that of seen with in silico point mutation (0.69), individual peaks contribute minimally to *ZmFCP1* expression.

We next examine the effect of peak combination (Com-) on gene expression (Figs. 1F and 3F). As expected, deletion of multiple peaks together can induce drastic changes in predicted expressions (such as Com-1 and Com-2 of *ZmFCP1*) with a notably expanded range from 0.51 to 1.76 (Fig. 3G). We defined this range of expression alterations as EP. We hypothesize that the broader predicted expression range and variability in combination deletions may result from interactions between regulatory elements as investigated in [33]. We further analyzed the relationships between peaks by examining the effects of both individual and combined peak deletions. Our findings indicate that the interactions between peaks are not necessarily additive (peak 2 + 4 in Fig. 3H), but also include synergistic effects (peak 3 + 6 in Fig. 3H). Thus, our model helps to unravel the hidden complexity among regulatory elements, providing valuable insights to achieve desired expression outcomes. Beyond this specific example, we next estimated the in silico editing plasticity of each gene in maize and the other three plant species, which would be helpful in assessing the consequences of promoter editing (Fig. 4A, Additional file 9: Supplementary Material 6).

EP is associated with gene evolvability

We next investigated the biological implications of EP. By assessing the EP of maize genes, we found that most of them exhibit an EP range of 0 to 7 (Fig. 4A). Focusing on genes at the extremes—those with an EP less than 1 (6th percentile) and those with an EP larger than 4 (92nd percentile), we discovered that genes having larger EPs tend to have higher tissue specificity (Fig. 4B). This suggests a potential association between EP and gene expression variability across tissues.

To further explore the possible connection between EP and gene evolvability, which takes the concept of "evolvability space" to visualize the spatial structure in gene expression prediction through in silico mutagenesis [37], we adapted the Basenji2-3K-NAM

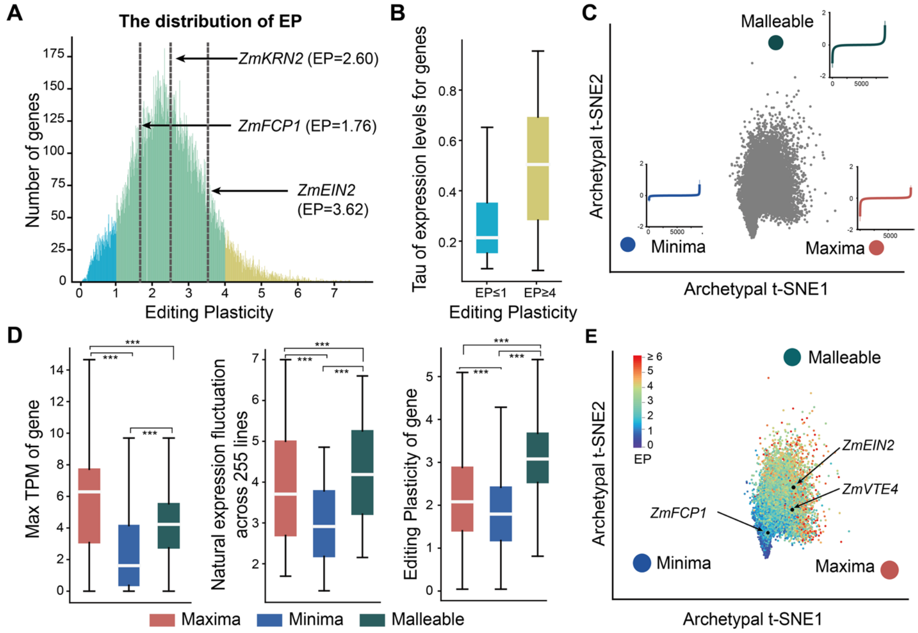

Fig. 4 Editing plasticity is associated with gene evolvability. **A** Distribution of editing plasticity of all genes in B73 reference genome. Some important genes are highlighted. **B** Tissue-specific expression pattern for two types of genes with extreme EP values. Tau values of the genes are shown to measure the tissue specificity of expression. **C** The evolvability space of Zea mays. Evolvability vectors for the 3-Kbp regulatory sequences of each gene are projected onto the evolvability space. Expression change patterns (colored curve) of three archetypes are demonstrated. **D** Expression patterns of three groups in the evolvability space. The maximum TPM values (left panel) and the natural expression fluctuation across 255 maize inbred lines for three groups (middle panel) are demonstrated. EP values of genes in three groups (right panel) are shown. Student's *t* test *P* values are presented. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. **E** Landscape of editing plasticity in the evolvability space. Evolvability vectors (points) are projected onto the evolvability space and are colored by editing plasticity

model to construct an evolvability space for all maize genes (see the "Methods" section; Fig. 4C, Additional file 2: Fig S16, Additional file 2: Figs S17-19 for other species). Within this space, we identified genes nearest to three anchor points—maxima, minima, and malleable—and calculated their average evolutionary vectors (Fig. 4C). Genes at the minima are more likely to exhibit expression increases after mutations (blue curve in Fig. 4C), whereas maxima genes are more prone to show expression decreases after mutation (red curve in Fig. 4C). Malleable genes have the potential for significant expression changes in both directions (green curve in Fig. 4C). Examination of the maximum expression (measured in transcript per million, TPM) of the three groups revealed that the maxima, minima, and malleable genes correspond to the highest, lowest, and intermediate levels, respectively (Fig. 4D). This pattern raises a link between gene evolvability and gene expression abundance.

We next wanted to explore if the malleable genes also exhibit greater expression variability in natural population. Utilizing another independent dataset of transcriptomes from seven tissues across 255 diverse maize lines [38], we observed that malleable genes do indeed demonstrate higher expression fluctuations (fluctuation, calculated as the difference between maximum and minimum expression values of this gene across 255 lines, with expression values taken as the maximum across multiple tissues of each line) (Fig. 4D). This supports the notion that, in addition to the gene expression level, gene evolvability is also associated with expression variability within the population. Interestingly, malleable genes additionally displayed significantly higher EP, a trend consistent with their natural expression fluctuations (Fig. 4D), reinforcing the association between EP and gene evolvability. By mapping EP values onto the maize evolvability space, we found that genes with larger EP values tend to be closer to the "malleable" anchor point (Fig. 4E). In conclusion, our analyses suggest that EP may serve as a useful indicator of gene evolvability, which could potentially be applied for pre-screening genes to assess their responsiveness to editing interventions. Specifically, the characteristics of the malleable genes include intermediate expression and increased EP and expression fluctuation, suggesting a potentially greater impact from promoter editing.

AI-guided promoter editing for ZmVTE4 gene enhancement

We next sought to assess the application of editing plasticity to guide a promoter editing experiment in the *ZmVTE4* gene (Fig. 1G), which is crucial for α-tocopherol variation and thereby associated with vitamin E activity [14]. Promoting *ZmVTE4* expression in conjunction with increasing α-tocopherol content in maize, a primary source of vitamin E for humans, is an important goal. In the KN5585 maize line [39], we first used in silico EP to predict the outcomes of promoter editing. Analysis identified ten peaks indicative of regulatory importance (Additional file 3: Table S18), all residing in the region spanning TSS − 200 to + 200bp (Fig. 5A). Our in silico deletions of each individual peak revealed that removal of peaks 4, 7, and 9 could increase *ZmVTE4* expression (Fig. 5B), with removal of peak 4 (TSS + 12 to + 21) presenting the highest negative contribution (Fig. 5A), potentially increasing expression by nearly 10% (Fig. 5B, Additional file 3: Table S18). By contrast, the removal of peaks 3, 6, and 10 could be critical for reducing expression (Additional file 3: Table S18). Further analysis of combined peak deletions showed that the maximum in silico EP value

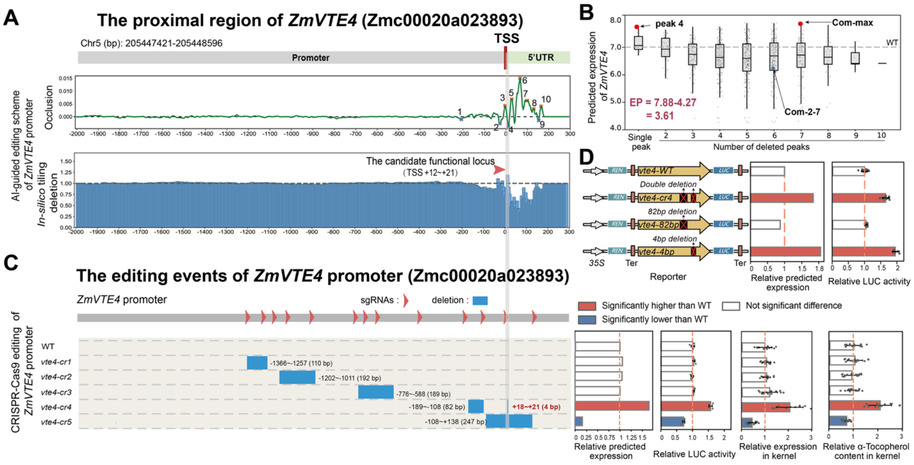

Fig. 5 AI-guided promoter editing for *ZmVTE4*. **A** The editing scheme of the *ZmVTE4* promoter. A total of 10 peaks were identified as the candidate CREs within the region of 2-Kbp upstream and 200-bp downstream of the TSS of *ZmVTE4* (top panel). An interpretability method of in silico tiling deletion (see the "Methods" section) was employed to highlight potential regulatory regions and proposed a candidate functional locus, the deletion of which could increase *ZmVTE4* expression level (12 ~ 21-bp downstream of the TSS, grey shade and red arrow). **B** Editing plasticity of *ZmVTE4*. The model simulated the predicted expression changes on deletion with a single peak and deletion with each combination of all peaks. The number of deleted peaks (*x*-axis) and predicted expression levels (*y*-axis) are demonstrated. **C** Experimental editing results for AI-guided editing scheme of *ZmVTE4* promoter. The left panel shows five promoter-edited alleles of *ZmVTE4*. The right panel shows relative predicted expression levels, LUC activities, *ZmVTE4* expression levels, and α-tocopherol contents of WT and five edited alleles. Repeated experimental results about LUC activities, *ZmVTE4* expression levels, and α-tocopherol contents in the right panel are significantly higher (*vte4-cr4* allele) or lower (*vte4-cr5* allele) compared to WT, determined by two-sided Student's *t* test at $P < 0.05$. **D** Luciferase activity validation of the *vte4-cr4* allele. The schematic diagrams of the *ZmVTE4* constructs (left), relative predicted expression, and relative LUC activity are displayed (right). LUC, firefly luciferase; REN, Renilla luciferase

(achieved by removing seven peaks: peaks 1, 2, 4, 5, 6, 9, and 10, termed Com-max) was comparable to the effect of deletion of peak 4 alone. This suggests that targeting peak 4 could be an optimal editing strategy for promoting *ZmVTE4* expression.

Guided by the in silico EP analysis, we conducted gene editing experiments (with 13 sgRNAs) to target the promoter and 5'UTR regions of *ZmVTE4* (Fig. 5C), including regions predicted by our model to affect gene expression. Five edited alleles were identified and designated *vte4-cr1* to *vte4-cr5* (Fig. 5C, Additional file 10: Supplementary Material 7). We tested these edited alleles against the Basenji2-3K-NAM model to predict changes in gene expression. The model showed no change for three alleles (*vte4-cr1, 2, 3*), but predicted increased gene expression for *vte4-cr4* and decreased for *vte4-cr5* compared to the WT (Additional file 11: Supplementary Material 8). These predictions were confirmed by LUC activities of the edited alleles in maize protoplasts (Fig. 5C, Additional file 11: Supplementary Material 8). As the co-occurrence of the 82-bp deletion and 4-bp deletion in the *vte4-cr4* allele, we further isolated the effect of each individual deletion on the observed increase in expression. We constructed two promoter variants with only a 4-bp deletion or an 82-bp deletion respectively and tested their LUC activities. The 4-bp deletion promoter variant showed a similar trend to the *vte4-cr4*, consistent with our predicted results (Fig. 5D). However, the 82-bp deletion promoter variant has no significant improvement (the third line in

Fig. 5D). Therefore, the critical region leading to increased activity in the *vte4-cr4* is the 4-bp deletion (located in peak 4) rather than the 82-bp deletion.

We then measured *ZmVTE4* expression and α-tocopherol content in maize using RT-qPCR and UPLC (ultra-performance liquid chromatography). *ZmVTE4* expression and α-tocopherol contents in kernels displayed a high correlation among the mutants, while the *vte4-cr4* allele displaying both significantly increased *ZmVTE4* expression and α-tocopherol levels (Fig. 5C). In summary, our "AI-guided editing scheme" effectively identified and validated the influence of specific genomic regions, demonstrating the utility of the strategy in streamlining gene editing approaches.

Discussion

Precise regulatory control is essential in modern plant breeding. Genetic manipulation of CREs offers the potential for such regulation, but often suffers from limited knowledge of a high-resolution regulatory map of CREs and no effective assessment of the editing potential of each gene. An AI-based expression prediction model promises to be a useful tool to overcome these difficulties.

Here, we performed an above proof-of-concept exploration of genome-wide CRE identification and AI-guided editing scheme. We constructed a genome-wide regulatory map of CREs, part of which was validated with UMI-STARR-seq. More importantly, we proposed the new concept of "editing plasticity" to theoretically analyze and estimate the editing effect of every possible promoter variant. The core tool that enables us to achieve this is an accurate expression prediction model, which displays stable and robust performance across multiple plant species, and which is essential for the implementation of a series of large-scale editing simulation experiments. This confirms the adage: "precision in prediction begets precision in engineering." In principle, this AI-based strategy could be applied to the high-throughput study of any other intermediate omics datasets, including those for open chromatin and metabolite abundances.

Another significant contribution of the current study is the innovative integration of evolvability space and editing plasticity. The evolvability space, colored by EP value, provides a clear navigation map for promoter editing experiments for any gene from the four plant species studied here. The EP value provides a theoretical estimate of the maximum effect of the edited variants, and the AI-guided precise editing scheme provides a systematic editing plan to achieve the given goal. We believe that this new tool will greatly accelerate the plant breeding process by facilitating the fine-tuning expression of target genes by gene editing.

Finally, looking to the future, we discuss some potential applications of the current study for synthetic biology. We found that a critical 4-bp deletion of the *vte4-cr4* allele is responsible for upregulating *ZmVTE4* expression, yet this never appeared in the promoter sequence of the pan-Zea population (721 individuals), demonstrating that it is not a natural variant which has occurred during evolution and was newly created by the gene editing experiment. Notably, the 247-bp deletion in *vte4-cr5* spans several peaks, resulting in a significant reduction in gene expression, suggesting higher-order *cis*-regulatory interactions between these peaks (peaks 2–7, termed Com-2–7). Although the model captured additive and synergistic effects between regulatory elements in *ZmFCP1*, it struggled to resolve more complex interactions including some higher-order *cis*-regulatory interactions and showed limited accuracy in predicting gene editing outcomes in rice and tomato (Additional file 3: Table S10-11). CRISPR editing is a powerful tool for validating the model's predictions on individual peaks and their combinations. However, precise deletion of a single peak at 10-bp resolution remains technically challenging, limiting experimental verification with the same resolution as the model prediction. This also implies that the synthesis of new DNA fragments, which do not exist in nature, has a broader sequence space to meet the demands of precise regulation of plant breeding. We believe that both the accurate prediction model and the evolvability space colored with EP will be very helpful for primary screening of massive synthetic DNA libraries, and that this will be a critical step in future synthetic biology activities.

Our study highlights the potential of AI-guided precise gene editing for improving crop traits, yet translating computational predictions into practical applications remains challenging due to environmental variability. While our current model effectively identifies CREs, it does not account for environmental influences. To address this, we propose a systematic framework integrating computational predictions with experimental validation in a cyclical process for the future works. This involves (i) utilizing RNA-seq data from over 500 accessions across five geographically diverse regions in China to identify environmentally responsive genes; (ii) applying our in silico prediction model to identify CREs in their promoter regions; (iii) using light as a model environmental factor to establish a light-response experimental platform and employing Massively Parallel Reporter Assays (MPRA) to validate and quantify light-responsive CREs; and (iv) building a predictive model based on MPRA data to estimate CRE activity, enabling the precise selection of CREs that drive desirable traits in specific environments. By iteratively refining the model with experimental results, we aim to enhance its accuracy and reliability, providing actionable insights for breeding crops with stable performance under diverse environmental stresses. This approach integrates computational predictions with practical breeding, contributing to agricultural sustainability in changing environments.

In addition, although our model incorporates expression data from multiple tissues, it currently relies on maximum expression across tissues, limiting its ability to identify tissue-specific CREs. A promising direction would be to predict gene expression separately for each tissue, potentially using multi-task or multi-label modeling approaches. Recent advances in machine-guided approaches have demonstrated the feasibility of designing tissue-specific CREs [ref NATURE], which could significantly enhance our framework's utility for crops with complex regulatory patterns and enable precise tissue-targeted gene expression designs.

Addressing these challenges is essential for improving the applicability and accuracy of AI-guided editing schemes. With continued integration of environmental and tissue-specific data, we anticipate that this framework will evolve into a more powerful tool for crop improvement and precision breeding.

Conclusions

Deciphering the roles of transcriptional cis-regulatory elements and their regulatory mechanisms in gene expression is pivotal for achieving targeted modulation of gene activity and associated phenotypic traits. Here, we present an AI-driven approach to precisely modulating gene expression by identifying CREs in plants and quantitatively estimating their effects. Through two deep learning models, we effectively predicted the functional roles of CREs: one incorporating distal elements with higher predictive power, and the other focusing on the proximal region to offer high-resolution CRE predictions. A large number of these CREs were verified for enhancer activity in vitro using UMI-STARR-seq, providing valuable resources for plant breeding and synthetic biology.

A notable finding of our work is the identification of a small functional region, typically overlooked by conventional genetic approaches due to a lack of natural variation. With the AI-guided editing scheme, we successfully engineered novel variants beyond natural variability using gene editing tools, thereby enabling precise control of gene expression.

Overall, our research provides both innovative tools and deeper insights into the molecular mechanisms governing gene expression regulation, while simultaneously laying a strong foundation for the precise genetic improvement of crop traits and the advancement of synthetic biology. Moving forward, we seek to extend the application of this methodology to other crops and more complex environmental conditions, further advancing innovation in agriculture and biotechnology.

Methods

Data preprocessing

Input sequences of the Basenji2-long model

For Z. mays, B73 reference genome of RefGen_v4 and annotation files were downloaded from https://download.maizegdb.org/Zm-B73-REFERENCE-GRAMENE-4.0/. For each gene, we extracted the genomic DNA sequence upstream and downstream of TSS from the reference genome with continuously increasing length from 10 K-bp to 140 K-bp (Additional file 4: Supplementary Material 1). Considering the double-stranded DNA, sequences from negative strand were reverse-complemented. In this way, we obtained the genomic sequences for 45,564 genes and then transformed them into one-hot encoded format. For the other three species, *O. sativa*, *S. lycopersicum*, and *A. thaliana*, we extracted genomic DNA sequences as above from their reference genomes according to Additional file 3: Table S1.

Output of maximum gene expression for the Basenji2-long model and the Basenji2-3K-B73 model

We employed an integrated gene expression dataset containing 421 RNA-seq datasets initially curated by [8]. To process each RNA-seq dataset, we downloaded it using Fasterq-dump (https://hpc.nih.gov/apps/sratoolkit.html) from the NCBI SRA (see the website of https://github.com/liulifenyf/plantCRE to find the list of SRA used here). Sickle (https://github.com/najoshi/sickle) was then used for quality-trimming and checking. The cleaned reads were aligned to the maize B73 RefGen_v4 reference using HISAT2 [40]. The aligned read counts were normalized to TPM using Stringtie [41]. Forty out of 421 RNA-seq datasets exhibiting abnormal in read count and TPM calculations for some genes were removed, because of alternative splicing and multi-mapping reads of these genes. The TPM values of remaining 381 RNA-seq datasets were log2-scaled, and the maximum value for each gene across all experiments was regarded as its expression level. For other model plant species, we took an identical set of processing steps. The datasets and their corresponding TPM values can be found at [16].

Splitting the training and testing datasets

To verify the generalizability of the trained model, we used *train_test_split* function from sklearn to randomly partition the genes, with 90% (41,007 genes) for training and 10% (4557 genes) for independent testing.

Model architecture

We adopted the Basenji2 model based on its proven success in human and mouse studies [10]. The classic Basenji2 model consists of seven convolutional blocks, followed by a number of dilated convolutional blocks, the count of which is automatically determined by the length of the input sequence, and a final convolutional layer. A notable hyperparameter of *channel number* (CM) [10] determines the structure of the Basenji2 model. For the convolutional block, let $C_i, i = 1, \cdots, 7$ be the filter number of the ith convolutional block, then C_i can be computed as:

$$\begin{cases} C_1 = 0.375 \times CM \\ C_i = 1.17759 \times C_{i-1}, i = 2, \cdots, 7 \end{cases}$$

For the dilated convolution blocks, each block has a filter count equal to CM, and the dilation rate $D_i, i = 1, \cdots, 11$ can be computed as:

$$\begin{cases} D_1 = 1 \\ D_i = round(1.5 \times D_{i-1}), i = 2, \cdots, 11 \end{cases}$$

where *round* means non-integer values must be taken as the nearest even number. The last dilated convolution block will produce a feature map matrix, which will be further fed into a convolution block (without maxpooling) containing $2 \times CM$ filters with a width of 1 to summarize all its information, setting the dropout probability to 0.05 to prevent overfitting. Finally, we employed a 1D convolution layer with 1 filter for dimensionality reduction. The output was then flattened and was inputted to a dense layer with 1 neuron for making final predictions (Additional file 2: Fig S1).

Training and hyperparameter optimization

We performed a fivefold-CV on the training set to optimize the Basenji2 model, with a particular emphasis on the CM hyperparameter. We tested four channel configurations: 360, 540, 720, and 900 with a fivefold-CV. After determining the optimal CM, we fixed this parameter and then proceeded to optimize the input length, using genomic sequence from 10K-bp to 140K-bp (Additional file 4: Supplementary Material 1, Additional file 3: Table S2). To assess the variance of the model, we conducted three independent runs using different splitting of the training and testing dataset (Additional file 4: Supplementary Material 1).

For training details, we initialized weights using He Normal [42] and regularized them using L_2 norm. The loss function is optimized using the Adam optimizer [43]. In the first three epochs, we set the learning rate as 0.001 and then set it as 0.00001 for the following epochs. We minimized the loss on a batch size of 32 and performed an early stop of three consecutive epochs. The relevant operations during model training were implemented using python3 using Keras2 with a Tensorflow2 backend.

We adopted the similar strategy to build models for three other model plant species (Additional file 2: Fig S2). We also compared Basenji2 with the other two mainstream models named Xpresso [11] and ExpResNet [44].

Deep interpretability methods

To investigate the importance of different features among all input features during the model prediction, we adopted the following four interpretation methods of deep learning to quantify the importance of each base of an input sequence.

Gradient-based method

The gradient × input method is one of the gradient-based methods, and it estimates contribution scores using the back-propagation procedure through the network [12]. Specifically, given a one-hot encoded input sequence, we first calculated the gradient vector and then employed an element-wise product between the gradient vector and the input. Subsequently, we took an average of contribution scores on four types of bases. Finally, we obtained a contribution score for each base with the same length as the input sequence.

Occlusion

The occlusion method is a perturbation-based interpretation method of deep learning [45]. Specifically, given a one-hot encoded input sequence, Occlusion scanned it with a sliding window of length $\boldsymbol{l}$ and stride $\boldsymbol{s}$ and then replaced the one-hot encoding within the sliding window with zero values ($N=[0{,}0{,}0{,}0]$). Finally, Occlusion measured the changes in the model predictions before and after the perturbation on the input to reflect the impact of the perturbed region on the output. The contribution score based on Occlusion is defined as:

$$Contribution\ score = (Y - Y_{0}ccluded)/Y$$

$\boldsymbol{Y}$ and $\boldsymbol{Y_{Occluded}}$ denote the model prediction for the original input sequence and the perturbed sequence respectively. Finally, we obtained a contribution score for each base with the same length as the input sequence.

In silico mutagenesis

To measure the expression change on every possible point mutation, a widely used interpretability method called “in silico mutagenesis” [46] was adopted. Specifically, this method mutates the current nucleotide into the other three nucleotides at each base and then measures the prediction change between before and after point mutation.

In silico tiling deletion

To assess the influence of larger stretches of the input sequence on predicted expression, we employed an interpretability method called “in silico tiling deletion” [47]. Different from in silico mutagenesis, in silico tiling deletion removes a small sliding window with an overlapping stride from the input sequence, and then computes the prediction change between before and after removal.

CRE identification by peak calling algorithm

CRE identification of the Basenji2-long model

To identify CREs for each gene, we here developed a peak-calling algorithm based on the base contribution score. We selected a threshold *t*, by taking the q^{th} percentile of the scores. Bases with scores higher than *t* were further examined for neighboring scores whether larger than *t*. This was done by computing the sum of scores within a window of length *w* around given base and filtering for bases with mean values exceeding *t*. The selected bases were kept, while the other base scores were set to 0. We then used the "*find_peaks*" function in SciPy [48] to identify the highest peak summit within a window length of *d*. These peak summits, along with ± *width*-bp genomic fragments, constituted the candidate CREs. The flowchart of the peak-calling algorithm is shown in detail in Additional file 2: Fig S3. Additionally, we used the FIMO [49] tool to scan the identified peaks to discover potential binding sites of transcription factors.

Tissue-specific CRE identification with omics data of the Basenji2-long model

For tissue-specific CRE-gene pair identification, we proposed an omics-modified contribution score modified by epigenetic patterns including chromatin accessibility and DNA methylation. The omics-modified contribution score was computed using a weighted sum formula: modified score = 0.9 × abs (gradient × input) + 0.05 × Cscore-0.05 × Mscore. Cscore and Mscore respectively denote the chromatin accessibility and CG methylation of each base in different tissues.

Validation of the candidate CREs

Selection of candidate CRE sequences for validation

We selected 12,000 sequences with a length of 200-bp from 745,684 candidate CREs (1K-bp) for synthesis and measured their activities using UMI-STARR-seq. We used a window size of 200-bp with a step size of 1 to scan for the contribution scores of 1K-bp long CREs, and selected the 200-bp long sequence with the maximum value of the sum of contribution score (see Additional file 1: Supplementary Methods).

UMI-STARR-seq library cloning

To perform STARR-seq in maize protoplasts, a screening vector was constructed by incorporating a CaMV 35S minimal promoter (−50 to + 5 bp), a *cat-1* intron, a GFP coding sequence and a *ccdB*-containing sequence into pMD18-T vector (Takara, catalog no. 6011), showed in Additional file 7: Supplementary Material 4.

Two hundred thirty-eight-mer oligonucleotide libraries, consisting of two Tn5 mosaic ends in inverted orientation at both ends and a forward CRE sequence in the center, were synthesized by Twist Bioscience for library cloning. Fragments were amplified following Twist oligo pool amplification guidelines. Two PCR reactions (95°C for 3 min (min); followed by 14 cycles of 98°C for 20 s (s), 55°C for 15 s, 72°C for 15 s; ended with 72°C for 1 min) with 10 ng Twist oligo pool as template were performed, using KAPA 2X HiFi HotStart ReadyMix Kit (Roche, catalog no. KK2602) and a single primer (referred as lib cloning primer in Additional file 7: Supplementary Material 4). The two PCR reactions were pooled and purified with 1X VAHTS DNA Clean Beads (Vazyme, catalog no. N411-01).

The STARR-seq plasmid was double-digested with the restriction enzymes SacI and KpnI and the upper band was gel-purified. The sticky ends of the linearized plasmid were blunted by incubation with Large (Klenow) Fragment (New England Biolabs, catalog no. M0210S), followed by DNA column purification. The oligonucleotide libraries and the vector were assembled using the NEBuilder HiFi DNA Assembly Master mix (New England Biolabs, catalog no. E2621S) in a total of five 20 μL reactions. The reactions were pooled, column-purified, and eluted in 10 μL of ultrapure water.

Five aliquots (20 μL each) of MegaX DH10B T1 Electrocompetent Cell (Invitrogen, catalog no. C640003) were transformed with 1.5 μL DNA each, according to the manufacturer's instructions. Five transformation reactions were pooled, transferred to 1 L LB-Amp medium, and incubated. Bacterial cultures were harvested at OD600 1.2–1.5. The plasmid libraries were extracted using GoldHi EndoFree Plasmid Maxi Kit Plus (CWBIO, catalog no. CW2113M) and the purified product was dissolved in ultrapure water to a concentration exceeding 1 μg/μl. The vector containing the CaMV 35S enhancer insertion, which served as a positive control, was constructed separately. Plasmids with the 35S enhancer inserted in both forward and reverse orientations were generated using the aforementioned method. Subsequently, 100 mL of bacterial cultures were cultivated for the extraction of the positive control plasmid. Positive controls were added at an appropriate ratio of 1:10,000 to the sequence-synthesized plasmid libraries.

Protoplast isolation and transient transformation

Protoplasts from second leaves of maize etiolated seedlings were isolated and transfected as described previously [23] with minimal modification. For transfection, 80 μg of plasmid DNA was mixed with 1 mL of protoplasts (at a concentration of 1×10^6 cells per mL) in a 14 mL round-bottom tube containing 1 mL of PEG-$CaCl_2$ solution. Each replicate transfected roughly 2 million protoplasts. The transfected protoplasts were incubated at 25°C in the dark for 16 h. This experiment was conducted using four biological replicates, each executed at different times (Additional file 2: Fig S6).

RNA and DNA isolation from transfected protoplasts

Two biological replicates of protoplasts were divided into two halves and subsequently pelleted to separately extract RNA and DNA. Total RNA was extracted from each biological replicate using TRIzol plus RNA purification kit (Invitrogen, catalog no. 1218355), while DNA was extracted using EasyPure plasmid miniPrep kit (TransGen Biotech, catalog no. EM101-02) and the CTAB method. Additionally, RNA and DNA from the other two biological replicates were extracted simultaneously using Allprep RNA/DNA kit (Aidlab, catalog no. RN29), which is shown in Additional file 7: Supplementary Material 4. The DNA re-isolated from transfected protoplast as STARR-seq input.

The polyA + RNA was isolated from 70 μg total RNA of each biological replicate using Dynabeads mRNA purification kit (Invitrogen, catalog no. 61006) and eluted with 30 μL 10 mM Tris–HCl (pH 7.5) buffer, and subjected to the treatment of TURBO DNase (Ambion, catalog no. AM2238) at 37°C for 30 min. Each replicate was purified using

1.8X VAHTS RNA clean beads (Vazyme, catalog no. N412-01) to inactivate TURBO DNase and eluted the RNA in 45 μL nuclease-free water.

First-strand reporter cDNA synthesis and second strand synthesis

UMI-STARR-seq was performed as described previously [26, 50]. Each biological replicate of first-strand cDNA synthesis was performed with SuperScript IV first-strand synthesis system (Invitrogen, catalog no. 18091050) using a reporter transcript-specific reverse transcription (RT) primer (5'-TAATCATCGCAAGACCGGCAACAG-3', referred as NOS rev primer in Additional file 7: Supplementary Material 4) and 350 ng polyA + RNA in every reaction. Each replicate performs 3 reactions for RT reactions and 1 reaction for minus RT control (replace SuperScript IV reverse transcriptase with nuclease-free water). Three RT reactions in each replicate were pooled and 0.2 μL of RNase A (Thermo Scientific, catalog no. EN0531) was added per RT reaction or minus RT reaction followed by magnetic bead-based purification (1.8X DNA clean beads) and was eluted in 12 μL per RT (or minus RT) reaction. Each replicate of second DNA strand of the reporter cDNA was synthesized by a linear PCR reaction (98°C for 1 min; 61°C for 30 s, 72°C for 1 min) without amplification using KAPA 2X HiFi HotStart ReadyMix Kit and a *cat-1* intron-spanning forward primer (referred as junction fw primer in Additional file 7: Supplementary Material 4) followed by magnetic bead-based purification (1.4X DNA clean beads) and was eluted in 11 μL per RT (or minus RT) reaction.

Unique molecular identifier (UMI) introduction

Each biological replicate of resulting double-strand reporter cDNA (including 3 RT reactions and 1 minus RT reaction), 100 ng of DNA isolated using a plasmid miniPrep kit per reaction (2 reactions), 400 ng of DNA per reaction (2 reactions) isolated using the CTAB method or Allprep RNA/DNA kit, and 100 ng of sequence-synthesized plasmid libraries without protoplast transformation per reaction (2 reactions) were synthesized through a linear PCR reaction (98°C for 1 min; 65°C for 30 s, 72°C for 1 min) without amplification using a KAPA 2X HiFi HotStart ReadyMix Kit and a modified Illumina i7 index primer containing 10 random nucleotides at the position of the Illumina i7 index (referred as P7-UMI primer in Additional file 7: Supplementary Material 4). Reporter DNA was purified using 1.4X DNA clean beads and was eluted in 20 μL per reaction.

Amplification of cDNA and input DNA library

Each UMI-introduced reporter cDNA obtained was amplified for Illumina sequencing by a 2-step nested PCR strategy. In the first PCR (98°C for 45 s; followed by 16 cycles of 98 °C for 20 s, 65C for 30 s, 72°C for 1 min, and then 72°C for 2 min), cDNA was amplified with KAPA 2X HiFi HotStart ReadyMix Kit and two primers, junction fw primer, and P7-junction rev primer (Additional file 7: Supplementary Material 4).

All PCR products from each replicate, except for the minus RT reactions, were pooled and purified using 0.8X DNA clean beads. The purified PCR products were then divided into two equal portions to serve as templates for the second PCR (98°C for 45 s; followed by 7–9 cycles of 98 °C for 20 s, 65C for 30 s, 72°C for 30 s, and then 72°C for 2 min) with KAPA 2X HiFi HotStart ReadyMix Kit, Illumina i5 primer, and P7-SeqReady rev primer (Additional file 7: Supplementary Material 4). Each UMI-introduced plasmid DNA was divided into two equal portions for use as templates and amplified with KAPA 2X HiFi HotStart ReadyMix Kit, Illumina i5 primer and P7-SeqReady rev primer (Additional file 7: Supplementary Material 4). Except for the minus RT reaction, the two halves of each replicate were indexed with different Illumina i5 primers. PCR products were purified using 1X DNA clean beads, then size-selected using 0.5X and 0.3X DNA clean beads, and eluted in 10 μL per reaction.

Illumina sequencing

Next-generation sequencing was performed at Shanghai Personal Biotechnology Co.,Ltd on an Illumina NovaSeq X Plus platform, following the manufacturer's protocol, using standard Illumina i5 indexes as well as UMIs at the i7 index.

UMI-STARR-seq data analysis

For sequence orientation, our plasmid libraries contain two 19 bp Tn5 mosaic ends in inverted orientation at both ends of the synthetic fragments (Additional file 7: Supplementary Material 4), allowing for insertion in either forward or reverse orientation. Here, we treated the orientations as independent sequences and calculated separate activity values for each, resulting in a pool of 24,000 sequences to be analyzed. We adopted the process in [47]. First, we mapped DNA input reads and cDNA reads to a reference file composed of 12,000 sequences using Bowtie2. We retained only the reads with MAPQ > 30, correct length, and with no mismatches. We collapsed both DNA and RNA reads by UMIs. Then we retained sequences with more than 10 reads in both replicates and added one read pseudocount to sequences with zero RNA counts. Then we calculated CRE activity for each sequence by the log2 fold-change over input based on all replicates using DESeq2. We defined the sequences with CRE activity greater than 0 and with adjusted p-value less than or equal to 0.05 as enhancers. Additional file 2: Fig S6 shows the consistency of the four replicates for both input (Additional file 2: Fig S6A) and cDNA (Additional file 2: Fig S6B), demonstrating the consistency and reliability of the experimental data.

Metric of Tau for expression tissue specificity

To measure the expression tissue specificity of a gene, we adopted a recommended metric [31], called Tau, which is defined as:

$$\tau = \frac{\sum_i^n \left(1 - \widehat{x_i}\right)}{n-1};\ \widehat{x_i} = \frac{x_i}{max\ (x_i)1 \le i \le n}$$

$\boldsymbol{x_i}$ is expression of the gene in tissue $\boldsymbol{i}$, $\boldsymbol{n}$ is the number of tissues. We computed the expression tissue specificity for each gene across 381 RNA-seq experiments (https://github.com/liulifenyf/plant-basenji2).

Constructing the evolvability space in maize

We generally followed the preprocessing pipeline described in a prior study [37]. The process of constructing an evolvability space in the current study is described as follows:

(1) Based on the Basenji2-3K-NAM model, we performed in silico mutagenesis on the 3K-bp regulatory sequence of each gene and sorted the prediction results to obtain a monotonically increasing vector D with length 9000, which was named the evolvability vector of this gene.
(2) We perform archetypal analysis on evolvability vectors of all genes with AANet [51]. AAnet accepts the evolvability vector (with dimension of 9000) as input, which is fully connected to the first encode layer. The encoder consists of five fully connected layers, with node numbers [1024, 512, 256, 128, 64], while the decoder architecture mirrors the encoder's structure with reverse order of five fully connected layers [64, 128, 256, 512, 1024]. Between encoder and decoder, there is a latent lay containing two nodes of latent variables, which were directly used to be the first two dimensions of archetypal triplets. The third dimension of archetypal triplets is calculated as one minus the sum of the first two dimensions. Consequently, the output of decoder with dimension of 1024 was then fully connected to the last layer, which gets the final output with dimension of 9000 (the restored evolvability vectors having dimension of 9000 with the same shape as the input). The AAnet autoencoder was trained using a learning rate of 1e−4, a batch size of 1024, and 10,000 batches.
(3) With the archetypal analysis, AAnet would get three anchor points (archetypal triplets) in the latent layer. We then used three anchor points with dimension of 9000 to perform t-SNE analysis and obtained the transformer matrix. The 2D t-SNE representations of the three anchor points were first generated and demonstrated in Additional file 2: Fig S16. After that, the evolvability vector of each gene was processed with the same encoder and t-SNE transformer matrix, and was also drawn as a point in Additional file 2: Fig S16. The evolvability space was successfully constructed when all genes were drawn. Finally, we colored all genes with their editing plasticity values. For three other model plant species, we adopted the same strategy with their own Basenji2-3K pan-genome model (Additional file 2: Figs S17-19).

CRISPR-Cas9 editing experiment on ZmVTE4

Cis-regulatory region editing of maize ZmVTE4 by CRISPR-Cas9

CRISPR-Cas9 technology was employed to induce mutations in the promoter and 5' UTR regulatory regions of *ZmVTE4* (Zmc00020a023893). The sgRNAs were designed based on the KN5585 genome using CRISPR-P (http://crispr.hzau.edu.cn/CRISPR2/) [52] with predictions from in silico EP (Additional file 3: Table S18). Subsequently, sgRNA arrays were synthesized, cloned into pCPBZmUbi-hspCas9, and transformed into KN5585 with *Agrobacterium tumefaciens* EHA105 (Weimi Biotechnology) [39, 53]. Genomic edits were screened through PCR-amplifying and Sanger sequencing of the target regions. The guide RNA sequences for *ZmVTE4* and primers for *ZmVTE4* allele genotyping are listed in Additional file 11: Supplementary Material 8.

Gene expression analysis

For the promoter-edited alleles of *ZmVTE4*, designated as *vte4-cr1* to *vte4-cr5*, homozygous edited plants and their wild type controls were selected from segregating populations. Mature leaves were collected and stored in liquid nitrogen with 8–10 plants for each biological replicates. Subsequently, approximate 0.1 g of maize leaf tissue was used to extract total RNA using Quick RNA Isolation Kit (Huayueyang Biotechnology CO., LTD. Beijing, China). A reverse transcription reaction was performed by the cDNA Synthesis SuperMix (TransGen Biotech), and expression levels were assessed through real-time fluorescence quantitative PCR with SYBR Green Master Mix (Vazyme Biotech) on a CFX96 Real-Time System. Each set of experiments was conducted three times, and the maize *ACTIN* gene (*Zm00001d010159*) was served as the internal control. The primers utilized for quantitative real-time PCR are listed in Additional file 11: Supplementary Material 8.

LUC activity assay

To assess the promoter activity of *ZmVTE4* edited alleles, a dual-LUC transient expression assay was conducted in maize protoplasts. Approximately 1.7 K-bp cis-regulatory region, including promoter and 5'UTR sequences of *ZmVTE*4, were amplified from WT and *vte4-cr1* to *vte4-cr5* respectively. These promoter sequences were then cloned into upstream of the mpCaMV of pGreen II 0800-LUC vector to generate the reporter constructs. Mesophyll protoplasts were isolated from the leaves of 10-day-old etiolated B73 seedlings. Subsequently, the prepared plasmids were transformed into the isolated mesophyll protoplasts using polyethylene glycol-mediated transformation [54]. After 16 h of dark cultivation, firefly LUC and REN activities were measured using the Dual-Luciferase Reporter Assay System (Promega, Madison, WI, USA) following the manufacturer's instructions. Each LUC activity experiment comprised three biological replicates, with each replicate having three technical replicates. Relative LUC activity was calculated by normalizing the firefly LUC activity to the Renilla LUC activity. The primers for amplifying the *ZmVTE4* cis-regulatory sequences of WT and *vte4-cr1* to *vte4-cr5* are listed in Additional file 11: Supplementary Material 8.

Tocopherol content measurement

Fresh mature leaves from both WT and these editing lines were harvested, promptly frozen in liquid nitrogen, and stored at − 80 ℃. Subsequently, according to the previously described method with minor modifications [14], tocopherols from a 0.1-g frozen sample were extracted and quantified using an ultra-performance liquid chromatography (UPLC) from the Waters Corporation (Milford, MA), employing a reverse-phase BEH C18 column (1.7 µm particle, 2.1 × 100 mm). The mobile phase consisted of 100% methanol containing 0.05% triethylamine (TEA) and 0.0028% butylated hydroxytoluene (BHT). The flow rate was 0.4 mL/min. Three tocopherol components, DT, GT, and AT, were measured. The ratio of AT to total tocopherol content, which was obtained by summing the three forms of tocopherol, was calculated and designated as the final phenotype of WT and editing lines. The three standards were sourced from the Sigma-Aldrich Company (St. Louis, MO). Both standards and samples were dissolved in the mobile phase. And absorbance detection at 295 nm was performed using a photo-diode array detector.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03516-7.

Additional file 1. Supplemental Methods

Additional file 2: Supplemental Figures (Fig S1-S19)

Additional file 3: Supplemental Tables (Table S1-S18)

Additional file 4: Supplementary Material 1. Prediction performance via PCC on four model plant species

Additional file 5: Supplementary Material 2. Details of candidate CREs identified with deep interpretability method combined with the trained Basenji2 model

Additional file 6: Supplementary Material 3. Distal elements identified by experiments and by Basenji2-long model

Additional file 7: Supplementary Material 4. Details of 12,000 candidate CREs for UMI-STARR-seq validation

Additional file 8: Supplementary Material 5. Details of high-resolution candidate CREs for B73

Additional file 9: Supplementary Material 6. Statistical evaluation of editing plasticity and evolvability spaces of four model plant species

Additional file 10: Supplementary Material 7. Traces of *ZmVTE4* edited alleles and WT.

Additional file 11: Supplementary Material 8. Details of *ZmVTE4* edited alleles.

Acknowledgements
We would like to thank Mr. Hao Liu from the high-throughput computing platform of the National Key Laboratory of Crop Genetic Improvement.

Peer review information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
J-B. Yan and X-H. Hu conceived, designed and supervised the whole study. Y. Qiu, L-F. Liu, X-L. Xiang and K-X. Deng collected and preprocessed the RNA-seq data and performed sequence-to-expression model analysis. Y. Qiu and L-F. Liu performed CRE identification, collected and preprocessed the pan-genome data, performed analyses of editing plasticity and evolution space. Y. Qiu collected and analyzed the distal element experimental data. J-L. Yan performed the UMI-STARR-seq and NGS data collection experiments and Y. Qiu analyzed the data. Y. Luo and S-Z. Wang performed the gene expression analysis, LUC activity assay and tocopherol content measurement experiments of ZmVTE4 variations. W-B. Xie provided valuable suggestions for the Basenji2 model. J-T. Xu, M-L. Jin, X-Y. Wu, L-W. Cheng and Y. Zhou performed the CRISPR mutant validation. Y. Qiu and L-F. Liu prepared the figures and tables. X-H. Hu, Y. Qiu, L-F. Liu, Y. Luo, H-J. Liu, A-R. Fernie and J-B. Yan drafted the manuscript with input from all authors. All authors read and approved the final manuscript.

Funding
This research was supported by the National Key Research and Development Program of China (2023YFD1202903), the National Natural Science Foundation of China (32321005, 32070689), the 111 Project Crop Genomics and Molecular Breeding (B20051), the Science and Technology Major Program of Hubei Province (2021ABA011), and Hubei Hongshan laboratory.

Data availability
The UMI-STARR-seq sequencing data generated in this study are deposited in the NCBI Sequence Read Archive (SRA) under accession number PRJNA1211828 [55]. The source code, data, and pre-trained models are publicly accessible in the GitHub repository https://github.com/liulifenyf/PlantCRE [56]. The repository is distributed under the MIT License. A permanent DOI for the archived version of the code is available via Zenodo: https://doi.org/10.5281/zenodo.14898181 [57].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors affiliated with WIMI Biotechnology Co., Ltd., Sanya, China (Jieting Xu, Minliang Jin, Xiaoyu Wu, Liwei Cheng) declare that their participation in this study was solely for scientific research purposes. All the authors declare no competing interests.

Received: 9 August 2024 Accepted: 25 February 2025
Published online: 10 March 2025

References

1. Schmitz RJ, Grotewold E, Stam M. Cis-regulatory sequences in plants: their importance, discovery, and future challenges. Plant Cell. 2022;34:718–41.
2. Liang Y, Liu H-J, Yan J, Tian F. Natural variation in crops: realized understanding, continuing promise. Annu Rev Plant Biol. 2021;72:357–85.
3. Liu L, Gallagher J, Arevalo ED, Chen R, Skopelitis T, Wu Q, Bartlett M, Jackson D. Enhancing grain-yield-related traits by CRISPR–Cas9 promoter editing of maize CLE genes. Nat Plants. 2021;7:287–94.
4. Rodríguez-Leal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB. Engineering quantitative trait variation for crop improvement by genome editing. Cell. 2017;171:470-480.e478.
5. Song X, Meng X, Guo H, Cheng Q, Jing Y, Chen M, Liu G, Wang B, Wang Y, Li J, Yu H: Targeting a gene regulatory element enhances rice grain yield by decoupling panicle number and size. Nat Biotechnol. 2022.
6. Liu N, Du Y, Yan S, Chen W, Deng M, Xu S, Wang H, Zhan W, Huang W, Yin Y. The light and hypoxia induced gene ZmPORB1 determines tocopherol content in the maize kernel. Science China Life Sci. 2024;67:435–48.
7. Oka R, Zicola J, Weber B, Anderson SN, Hodgman C, Gent JI, Wesselink JJ, Springer NM, Hoefsloot HCJ, Turck F, Stam M. Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. Genome Biol. 2017;18:1.
8. Washburn Jacob D, Mejia-Guerra Maria K, Ramstein G, Kremling Karl A, Valluru R, Buckler Edward S, Wang H. Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proc Natl Acad Sci. 2019;116:5542–9.
9. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. 2018;28:739–50.
10. Kelley DR. Cross-species regulatory sequence activity prediction. PLoS Comput Biol. 2020;16:e1008050.
11. Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. Cell Rep. 2020;31:107663.
12. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, Assael Y, Jumper J, Kohli P, Kelley DR. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203.
13. Hu X, Fernie AR, Yan J. Deep learning in regulatory genomics: from identification to design. Curr Opin Biotechnol. 2023;79:102887.
14. Li Q, Yang X, Xu S, Cai Y, Zhang D, Han Y, Li L, Zhang Z, Gao S, Li J. Genome-wide association studies identified three independent polymorphisms associated with α-tocopherol content in maize kernels. PLoS One. 2012;7:e36807.
15. Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. Brief Bioinform. 2021;22:bbaa177.
16. Liu L, Qiu Y, Hu X. PlantCRE. 2023. http://www.hzau-hulab.com/plantCRE.
17. Salvi S, Sponza G, Morgante M, Tomes D, Niu X, Fengler KA, Meeley R, Ananiev EV, Svitashev S, Bruggemann E, et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci. 2007;104:11376–81.
18. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene tb1. Nat Genet. 2011;43:1160–3.
19. Li E, Liu H, Huang L, Zhang X, Dong X, Song W, Zhao H, Lai J. Long-range interactions between proximal and distal regulatory regions in maize. Nat Commun. 2019;10:2633.
20. Peng Y, Xiong D, Zhao L, Ouyang W, Wang S, Sun J, Zhang Q, Guan P, Xie L, Li W, et al. Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. Nat Commun. 2019;10:2632.
21. Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, Noshay JM, Galli M, Mejía-Guerra MK, Colomé-Tatché M. Widespread long-range cis-regulatory elements in the maize genome. Nature plants. 2019;5:1237–49.
22. Noshay JM, Anderson SN, Zhou P, Ji L, Ricci W, Lu Z, Stitzer MC, Crisp PA, Hirsch CN, Zhang X. Monitoring the interplay between transposable element families and DNA methylation in maize. PLoS Genet. 2019;15:e1008291.
23. Tu X, Mejía-Guerra MK, Valdes Franco JA, Tzeng D, Chu P-Y, Shen W, Wei Y, Dai X, Li P, Buckler ES. Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. Nat Commun. 2020;11:5089.
24. Zhao H, Tu Z, Liu Y, Zong Z, Li J, Liu H, Xiong F, Zhan J, Hu X, Xie W. PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. Nucleic Acids Res. 2021;49:W523–9.
25. Gui S, Wei W, Jiang C, Luo J, Chen L, Wu S, Li W, Wang Y, Li S, Yang N, et al. A pan-Zea genome map for enhancing maize improvement. Genome Biol. 2022;23:178.
26. Neumayr C, Pagani M, Stark A, Arnold CD. STARR-seq and UMI-STARR-seq: assessing enhancer activities for genome-wide-, high-, and low-complexity candidate libraries. Curr Protoc Mol Biol. 2019;128:e105.
27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
28. Zhao Y, Xu L, Huang Y, Wu H, Zhang X, Hu X, Ma Q. Identification and characterization of the core region of ZmDi19–5 promoter activity and its upstream regulatory proteins. Int J Mol Sci. 2022;23:7390.
29. Zrimec J, Börlin CS, Buric F, Muhammad AS, Chen R, Siewers V, Verendel V, Nielsen J, Töpel M, Zelezniak A. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat Commun. 2020;11:6141.
30. Schubert E, Lenssen L. Fast k-medoids clustering in rust and python. J Open Source Software. 2022;7:4183.
31. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform. 2017;18:205–14.
32. Ciren D, Zebell S, Lippman ZB. Extreme restructuring of *cis*-regulatory regions controlling a deeply conserved plant stem cell regulator. PLoS Genet. 2024;20(3):e1011174.
33. Wang X, Aguirre L, Rodríguez-Leal D, Hendelman A, Benoit M, Lippman ZB. Dissecting cis-regulatory control of quantitative trait variation in a plant stem cell circuit. Nat Plants. 2021;7:419–27.
34. Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. 2021;373:655–62.

35. MaizeGDB. MaizeGDB: a community-oriented, federally funded informatics service for maize research. 2025. https://maizegdb.org. Accessed 20 Feb 2025.
36. Li C, Li Y, Song G, Yang D, Xia Z, Sun C, Zhao Y, Hou M, Zhang M, Qi Z, et al. Gene expression and expression quantitative trait loci analyses uncover natural variations underlying the improvement of important agronomic traits during modern maize breeding. Plant J. 2023;115:772–87.
37. Vaishnav ED, de Boer CG, Molinet J, Yassour M, Fan L, Adiconis X, Thompson DA, Levin JZ, Cubillos FA, Regev A. The evolution, evolvability and engineering of gene regulatory DNA. Nature. 2022;603:455–63.
38. Kremling KAG, Chen SY, Su MH, Lepak NK, Romay MC, Swarts KL, Lu F, Lorant A, Bradbury PJ, Buckler ES. Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature. 2018;555:520–3.
39. Liu H-J, Jian L, Xu J, Zhang Q, Zhang M, Jin M, Peng Y, Yan J, Han B, Liu J. High-throughput CRISPR/Cas9 mutagenesis streamlines trait gene identification in maize. Plant Cell. 2020;32:1397–413.
40. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
41. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
42. He KM, Zhang XY, Ren SQ, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015:1026–1034.
43. Kingma DP, Ba J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference for Learning Representations. 2015:1-15.
44. Zhang YL, Zhou X, Cai XD. Predicting gene expression from DNA sequence using residual neural network. bioRxiv. 2020; 2020.06.21.163956.
45. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. Proceedings of the International Conference for Learning Representations. 2018:1-16.
46. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33:831–8.
47. de Almeida BP, Reiter F, Pagani M, Stark A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat Genet. 2022;54:613–24.
48. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17:261–72.
49. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8.
50. Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339:1074–7.
51. Dijk Dv, Burkhardt DB, Amodio M, Tong A, Wolf G, Krishnaswamy S: Finding archetypal spaces using neural networks. In 2019 IEEE International Conference on Big Data (Big Data); 9–12 Dec. 2019. 2019:2634–2643.
52. Liu H, Ding Y, Zhou Y, Jin W, Xie K, Chen L-L. CRISPR-P 2.0: an improved CRISPR-Cas9 tool for genome editing in plants. Mol Plant. 2017;10:530–2.
53. Li C, Liu C, Qi X, Wu Y, Fei X, Mao L, Cheng B, Li X, Xie C. RNA-guided Cas9 as an in vivo desired-target mutator in maize. Plant Biotechnol J. 2017;15:1566–76.
54. Yoo SD, Cho YH, Sheen J. Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nat Protoc. 2007;2:1565–72.
55. Yan J, Qiu Y, Liu L, Yan J, Xiang X, Wang S, Luo Y, Deng K, Xu J, Jin M, et al.: High-throughput validation of Cis-regulatory elements in maize using UMI-STARR-seq. Datasets. NCBI Sequence Read Archive. 2025. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1211828.
56. Yan J, Qiu Y, Liu L, Yan J, Xiang X, Wang S, Luo Y, Deng K, Xu J, Jin M, et al: PlantCRE: identification of plant Cis-regulatory elements using deep learning. GitHub. 2025. https://github.com/liulifenyf/PlantCRE.
57. Yan J, Qiu Y, Liu L, Yan J, Xiang X, Wang S, Luo Y, Deng K, Xu J, Jin M, et al: PlantCRE: identification of plant Cis-regulatory elements using deep learning. Zenodo. 2025. https://doi.org/10.5281/zenodo.14898181.

Publisher's Note