Genome Biology

METHODOLOGY

Open Access

DeepWheat: predicting the effects of genomic variants on gene expression and regulatory activities across tissues and varieties in wheat using deep learning

Zhigang Ma[1†], Jiazi Zhang[1†], Hongcui Pei[1†], Yanhong Liu[1], Hongning Tong[1], Lei Wang[2] and Zefu Lu[1*]

†Zhigang Ma, Jiazi Zhang, and Hongcui Pei contributed equally to this work.

*Correspondence: luzefu@caas.cn

[1] State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[2] Center for Agricultural Resources Research, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Shijiazhuang 050022, China

Abstract

Spatiotemporal gene expression shapes key agronomic traits, yet tissue-specific prediction remains challenging in complex crops. We present DeepWheat, a broadly applicable deep learning framework comprising DeepEXP and DeepEPI, for accurate, tissue-specific gene expression prediction. DeepEXP integrates sequence and epigenomic features to predict gene expression (PCC 0.82–0.88), while DeepEPI predicts epigenomic maps from DNA sequence to support model transfer across varieties. Validations in five wheat cultivars confirm robustness and accuracy. DeepWheat also identifies regulatory variants with strong expression effects, enabling targeted *cis*-regulatory elements editing and offering a powerful tool for crop functional genomics and breeding.

Background

Cis-regulatory elements (CREs) are pivotal in the precise regulation of gene expression [1, 2], yet their functional impacts remain challenging to characterize [3, 4]. The application of machine learning techniques has markedly advanced the prediction of regulatory activities and their influence on epigenomic modifications and gene expression [5, 6]. These approaches have improved our ability to characterize CRE function and support the development of trait-improving strategies [5, 7, 8].

Considerable progress has been achieved in model plants and certain crops, such as *Arabidopsis*, rice and maize, a diverse array of computational and experimental approaches has been developed and integrated by researchers to dissect the complexities of gene regulation and expression [6, 9]. These efforts have greatly enriched our understanding of the regulatory mechanisms underpinning gene function and the ways in which genetic variations influence phenotypic outcomes. By simulating virtual mutations, these models are able to predict the effects of alterations within CREs, thus

guiding the precise editing of these regions [10, 11]. However, progress in complex-genome species, such as wheat, remains limited, due to the additional complexity introduced by their unique genomic characteristics [12, 13].

The prediction accuracy of gene expression can be significantly improved by integrating 3D chromatin structures and epigenomic data, compared to models relying solely on sequence-based information [14]. These findings are consistent with that epigenomic modifications, such as DNA methylation and histone modifications, play a critical role in regulating gene expression [15–17]. However, obtaining high-quality epigenomic data remains expensive and challenging, particularly in plants, where data quality often lags behind that of animal models or *Arabidopsis* [18].

Another challenge is the low prediction accuracy across tissues [14], largely due to the similar overall expression patterns of most genes [19]. However, the spatiotemporal-specific expression of genes is crucial for the formation of key traits, for example, variations in the expression pattern of the *BRD3* gene lead to its tissue-specific expression in grain, resulting in a multi-grain phenotype that significantly enhances rice yield [20]. Therefore, accurate prediction of tissue-specific gene expression is urgently needed in the field of crop improvement. Since epigenomic features are closely linked to tissue-specific expression [21], their incorporation holds potential for enhancing the precise prediction of tissue-specific genes [14]. However, this strategy has yet to be widely adopted in plant species.

Wheat's large and complex genome, characterized by redundancy and structural variations [12, 22], presents significant challenges in accurately predicting gene expression and regulatory activities across tissues, developmental stages, and genetic backgrounds, making traditional sequence-based models less effective. To address these challenges, we developed DeepWheat, comprising two models: DeepEXP and DeepEPI. DeepEXP integrates epigenomic and transcriptomic data to predict gene expression in various wheat tissues and stages using a deep learning framework, achieving Pearson correlation coefficients (PCC) ranging from 0.82 to 0.88 and outperforming sequence-only models. DeepEPI predicts epigenomic features from DNA sequences, which are then integrated with sequence data to improve gene expression predictions. This integration allows the model transfer and enhances prediction accuracy across wheat varieties. Our models also evaluate genetic variations, revealing that indels have a stronger impact on gene expression than SNPs. Additionally, not only promoter regions but also the 5′UTR, 3′UTR, and introns play critical roles in gene regulation. These advancements provide a comprehensive toolkit for exploring gene regulation in wheat, with significant potential to enhance breeding strategies and functional genomics in this vital crop species.

Results

Method outline

Gene expression varies substantially across wheat tissues and developmental stages [17], yet most existing models predict expression based solely on genomic sequence, often using median or peak expression across tissues [23, 24]. These approaches lack resolution for tissue- or stage-specific predictions. To address this limitation, we first trained a model based on the Basenji2 [25, 26] framework to predict wheat gene expression from sequence alone. Despite extensive tuning, prediction accuracy remained limited

(PCC < 0.66 across tissues), and dropped to 0.25 in vernalized leaves (Additional file 1: Fig. S1A). Similarly, models built using the Xpresso [27] and PhytoExpr [23] frameworks failed to achieve high-accuracy predictions across tissues and stages (Additional file 1: Fig. S1A). Independent evaluation on 4700 randomly selected genes (Additional file 1: Fig. S1B) further revealed large discrepancies between predicted and experimental tissue-specificity indices (Additional file 1: Fig. S1C), suggesting poor performance in capturing spatiotemporal dynamics.

Considering epigenomic modifications are highly relative with gene expression (Additional file 1: Fig. S1D and S1E) and the feasibility of predicting gene expression using epigenomic and methylation data in human [14], we developed DeepWheat, a deep learning toolkit comprising two core modules: DeepEXP and DeepEPI (Fig. 1). DeepEXP integrates genomic sequence and experimental epigenomic data (e.g., chromatin accessibility and histone modifications) across multiple wheat tissues and developmental stages to predict tissue-specific gene expression. High-quality epigenomic profiles were first reconstructed using AtacWorks [28]. Features from proximal regulatory regions and partial genebodies were extracted using two parallel convolutional neural network (CNN) branches, followed by channel-wise concatenation and deep residual learning blocks. A

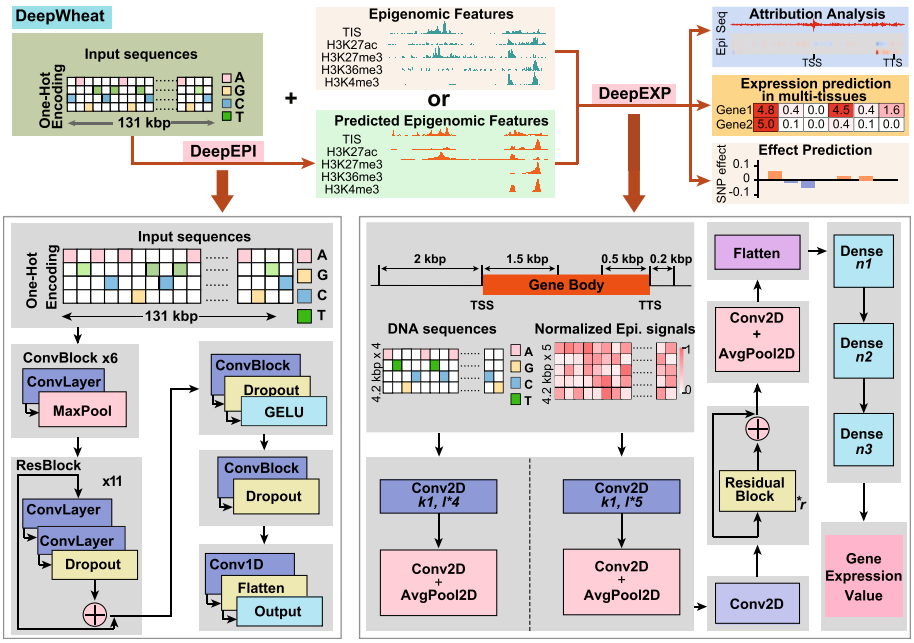

Fig. 1 Architecture and predictive framework of the DeepWheat. The DeepWheat suite comprises two complementary deep-learning models, DeepEPI and DeepEXP, that together predict wheat epigenomic modifications and gene expression. DeepEPI is built on the Basenji2 framework and accepts 131 kb genomic sequences surrounding each locus; through stacked convolutional and residual blocks it outputs genome-wide epigenomic signal tracks, including TIS (Tn5 transposome integration sites in ATAC-seq) and various histone modifications, across multiple tissues and predicts the effects of SNPs and INDELs on these marks. DeepEXP integrates DNA sequence windows flanking transcription start sites (TSSs) and transcription termination sites (TTSs) with epigenomic feature vectors—either experimentally measured or DeepEPI-predicted. Its architecture is defined by the number of convolutional filters (*k*), sequence length (*l*), number of residual layers (*r*) and number of fully connected neurons (*n*). DeepEXP subsequently predicts gene expression levels across tissues, quantifies the impact of sequence variants on tissue-specific expression, and, via attribution analysis, identifies both tissue-general and tissue-specific regulatory sequence elements

fully connected regression head outputs non-negative, continuous gene expression values (Fig. 1, Methods).

DeepEPI is an optimized version of the Basenji2 [25, 26] architecture, trained to predict tissue- and stage-specific chromatin accessibility and histone modification profiles directly from DNA sequence. Given the high cost of generating experimental epigenomic datasets, we further designed a transfer strategy: DeepEPI-predicted regulatory features are combined with sequence and fed into DeepEXP to predict gene expression in a purely in silico setting, without requiring experimental epigenomic input (Fig. 1).

Together, DeepEXP and DeepEPI enable high-resolution, cross-tissue prediction of gene expression and regulatory activity in wheat. We further developed an analysis pipeline to perform attribution analysis, identify variants with strong influence on gene expression, and assess the effects of genomic variants on gene expression and epigenomic states across tissues and developmental stages (Fig. 1), providing a valuable tool for functional variant interpretation and CRE editing.

Integrating sequence and epigenomic data improves tissue-specific gene expression prediction in wheat

DeepEXP, a deep learning model that integrates genomic sequences and multi-omic epigenomic data, was developed to accurately predict gene expression across wheat tissues and developmental stages (Fig. 1). To identify the optimal length of proximal regulatory sequences, we first tested different intervals of sequences and epigenomic features around the transcription start site (TSS) as model inputs. The optimal region was 2000 bp upstream and 1500 bp downstream of the TSS, which yielded improved PCC and R^2 across six tissues (Fig. 2A, Additional file 2: Table S1A). Subsequently, with the TSS region fixed, we incorporated sequences and epigenomic maps within 500 bp upstream and 200 bp downstream of the transcription termination site (TTS). This addition further improved prediction accuracy, as validated in spike and vernalization stages (Fig. 2B, Additional file 2: Table S1B). Training on sequences and epigenomic data from 2000 bp upstream to 1500 bp downstream of TSS and 500 bp upstream to 200 bp downstream of TTS, DeepEXP achieved PCC values between 0.82 and 0.88, outperforming sequence-only models such as Basenji2 [25, 26], Xpresso [27], and PhytoExpr [23] (Fig. 2C). Notably, DeepEXP also surpassed these models in *Arabidopsis*, rice, and maize, demonstrating its broader applicability (Fig. 2D and Additional file 2: Table S2).

To assess tissue-specific predictive performance, we selected 1543 genes with high tissue specificity (tissue specificity index, Tau > 0.8) from the independent test set of 4,700 genes (Additional file 1: Fig. S1B). Sequence-only models exhibited a notable drop in performance for these genes, whereas DeepEXP showed only a minor reduction across all tissues (Fig. 2E, Additional file 1: Fig. S1C and S1F), highlighting the critical role of epigenomic features in capturing tissue-specific expression. A similar trend was observed in the prediction of expression levels for cloned genes in spike and leaf tissues (Additional file 1: Fig. S1G).

To assess the contributions of different epigenomic modifications, we integrated individual modification data to predict gene expression. Chromatin accessibility data had the highest contribution in most tissues, while H3K27me3 had the least, with other modifications in between (Additional file 1: Fig. S1H). We then evaluated

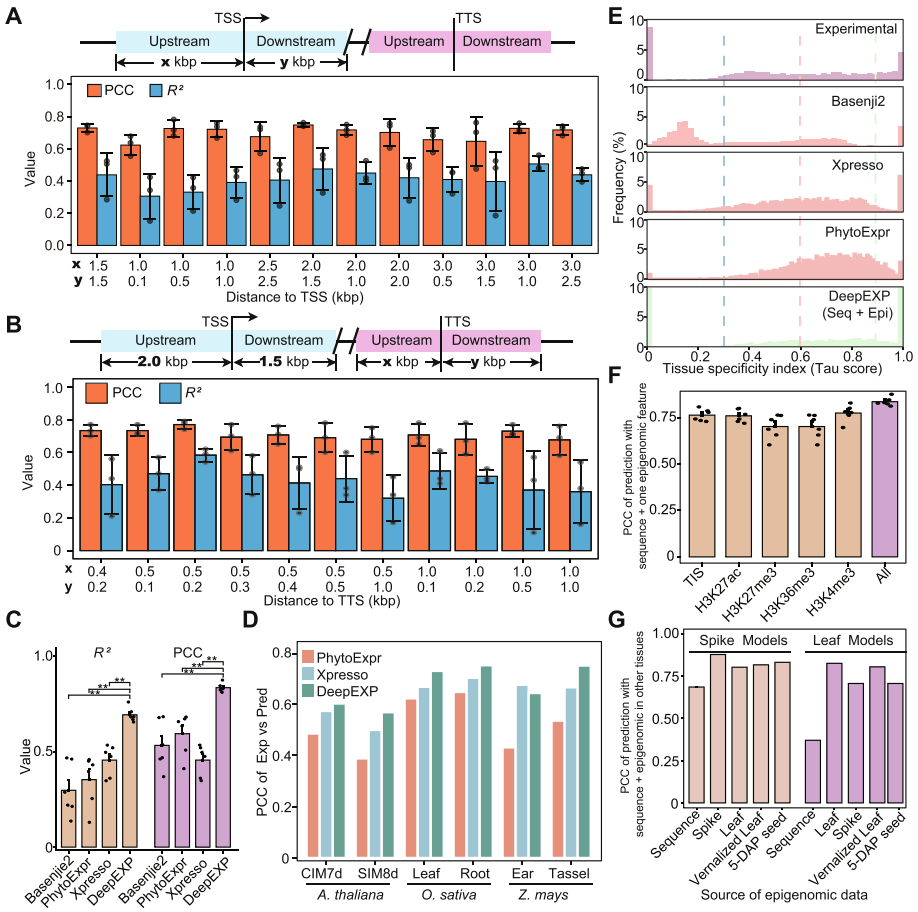

Fig. 2 Training of the DeepEXP gene expression prediction model using DNA sequence and epigenomic data. **A**–**B** Prediction performance using different length regions around TSS (**A**) and TTS (**B**). For TTS length performance inspection, the best length of TSS regions were used. Pearson correlation coefficient (PCC) and R^2 values of DeepEXP on testing datasets using sequence and epigenomic data from varying lengths of regions surrounding the TSS and TTS across different tissues. **C** PCC and R^2 values between experimental and predicted gene expression levels on the independent test set. Each dot indicated the value for prediction in one tissue. Two asterisks indicate $p < 0.01$ in *Wilcoxon* rank-sum test. **D** PCC between experimental and predicted gene expression levels on independent test sets across different tissues of *Arabidopsis*, rice and maize. CIM7d, 7-day callus induction medium; SIM8d, 8-day shoot induction medium. **E** Frequency distribution histograms illustrating the tissue specificity index (Tau) of predicted gene expression for 4700 genes in the independent test set. Predictions are shown for Basenji2, Xpresso, PhytoExpr and DeepEXP, and experimentally measured gene expression as a reference. Tau values range from 0 to 1, with values closer to 1 indicating stronger tissue-specific expression. All histograms share the same axis range and bin size (100 equally spaced bins). **F** PCC between experimental and predicted gene expression levels on independent test sets across multiple wheat tissues using DeepEXP, by integrating sequence with either one individual epigenomic feature or multiple epigenomic features. TIS indicates Tn5 transposome integration sites in ATAC-seq. **G** PCC between experimentally measured and predicted gene expression levels in spike and leaf tissues. Comparisons include models trained with different input types: those integrating epigenomic data from multiple tissues and developmental stages along with sequence information, and the sequence-only Basenji2 model

DeepEXP using only sequence plus single epigenomic data. Even with the addition of a single epigenomic data, model performance improved significantly over sequence-only inputs (Fig. 2F). We also find prediction accuracy varied across tissues and stages. Quality analysis of epigenomic data revealed moderate correlations

between data quality and prediction accuracy (Additional file 1: Fig. S1I), indicating that data quality significantly impacts accuracy. To further enhance prediction accuracy by improving the quality of epigenomic data, we employed Atac Works [28], a deep learning model for refining epigenomic tracks. This approach increases prediction accuracy and robustness, especially for low-quality samples (Additional file 1: Fig. S1J).

Recognizing the high cost and limited availability of multi-tissue epigenomic data, we further tested transferability across tissues: using chromatin data from one tissue to predict gene expression in another still outperformed sequence-only models, particularly when the donor and target tissues were developmentally similar (Fig. 2G). We hypothesize that sequence-only models capture static features like motif presence but fail to reflect dynamic chromatin states essential for gene regulation. By incorporating epigenomic signals-even from non-matching tissues—DeepEXP effectively filters out inactive regions and models complex sequence–chromatin interactions, thereby enhancing tissue-specific prediction accuracy.

Optimizing multi-tissue epigenetic profiling and sequence integration for improved gene expression prediction in wheat

Since integrating sequences with epigenomic maps improves gene expression prediction accuracy and obtaining epigenomic data is more costly than gene expression data [29, 30], we intend to necessitate a model to predict epigenomic features and integrated these predicted features and sequences to predict gene expressions (Fig. 1). We speculate that this approach will facilitate the transfer of the prediction model to different wheat varieties.

We developed DeepEPI to predict wheat epigenomics using multi-tissue (stage) epigenomic data, optimizing the Basenji2 [25, 26] framework (Fig. 1). On an independent test set, the PCC between predicted and experimental chromatin accessibility and histone modifications ranged from 0.65 to 0.79 and 0.30 to 0.80, respectively (Fig. 3A and Additional file 2: Table S3). The PCC for peak predictions ranged from 0.83 to 0.94 and 0.69 to 0.96 (Fig. 3B and Additional file 2: Table S3). Predicted and experimental epigenomic distributions showed high consistency (Fig. 3C), and the data quality was strongly correlated (Additional file 1: Fig. S2A–2C). These models also identified distal accessible chromatin regions (Fig. 3D), which could be used to discover putative enhancers.

We integrated sequence data with epigenomic features predicted by DeepEPI to predict gene expression (Fig. 1). The PCC between predicted and experimental values ranged from 0.71 to 0.74 across different tissues and stages (Fig. 3E). Although this accuracy was lower than that achieved using combined sequence and experimental epigenomic data, it was significantly higher than using sequence data alone (Fig. 3E). The frequency distribution of predicted gene expression specificity closely matches experimental values (Fig. 3F), outperforming sequence-only models. We further transferred this strategy to the Chinese Spring wheat variety and still achieved higher prediction accuracy compared to sequence-only models (Additional file 2: Table S4). Further analysis of tissue-specific genes (Tau > 0.8) [31] between spike, leaf and vernalized leaf tissues showed that integrating sequences with predicted epigenomic features enhanced prediction accuracy for tissue-specific gene expression (Additional file 1: Fig. S2D).

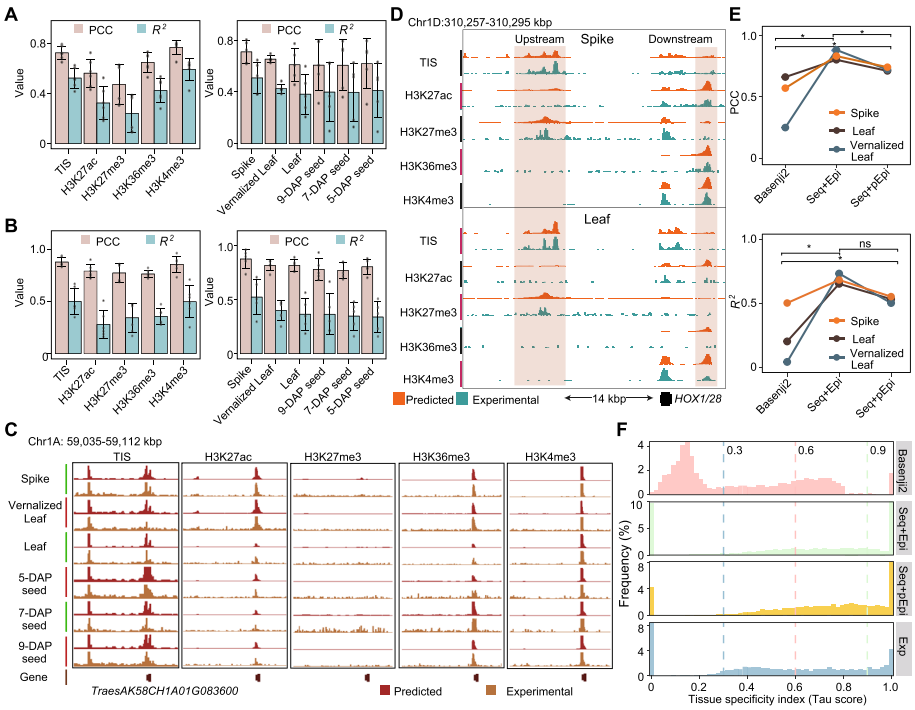

Fig. 3 Integration of DeepEPI and DeepEXP to optimize DNA sequence-based gene expression prediction model. **A**–**B** PCC and R^2 values for DeepEPI's prediction of various epigenomic marks, with **A** showing predictions based on track signals and **B** showing predictions based on modified coordinates. DAP, days after pollination. TIS indicates Tn5 transposome integration sites in ATAC-seq. **C** IGV screenshot showing predicted and experimental signals for various epigenomic modifications across different tissues. **D** Visualization of predicted putative distal regulatory regions through DeepEPI. **E** PCC and R^2 values of gene expression prediction using the DeepEPI + DeepEXP model (sequence + predicted epigenomic data, Seq + pEpi), compared with models based on sequence data alone (Seq) and sequence + experimental epigenomic data (Seq + Epi). Statistical significance between the two groups was tested using the *Wilcoxon* rank-sum test (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ns = not significant) **F** Gene's Tau distribution derived from experimental data and DeepEXP predictions using sequence data alone (Seq), sequence + experimental epigenomic data (Seq + Epi), and sequence + predicted epigenomic data (Seq + pEpi)

Validation of DeepWheat prediction accuracy across 5 wheat varieties

To validate the across varieties prediction accuracy of DeepWheat, we generated chromatin accessibility and transcriptomic data from 5 additional wheat varieties (Fig. 4A). In young spike, accessible chromatin regions (ACRs) differing from the AK58 cultivar accounted for 6.31% to 38.4% of the total ACRs (Additional file 1: Fig. S3A), in young spike tissues the number of differentially expressed genes compared to AK58 ranged from 8455 to 11,535 (Additional file 1: Fig. S3B). These variations indicated that the samples were suitable for evaluating model accuracy.

To predict the epigenetic landscape and gene expression levels across different wheat varieties, we employed the strategy outlined in Fig. 4B (Methods). We selected differentially expressed genes with high-quality INDELs/SNPs located within 2000 bp upstream and 1500 bp downstream of the TSS, as well as 500 bp upstream and 200 bp downstream of the TTS. A total of 4638 and 5145 genes were selected from spike and leaf tissues across 5 varieties, with an average of 7400 and 6193 INDELs, and 38,411 and 38,051 SNPs, respectively, between AK58 and each variety for spike and leaf tissues. A pseudo-genome centered on the selected genes was generated by modifying the AK58 reference

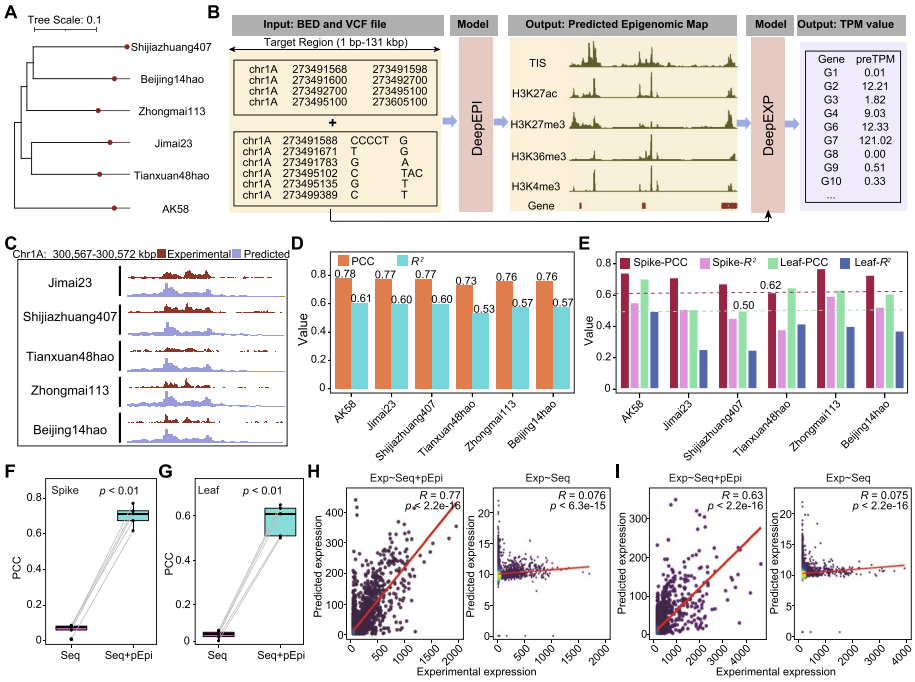

Fig. 4 Validation of the prediction accuracy of DeepEPI, DeepEXP, and DeepEPI + DeepEXP models in different wheat varieties. **A** Phylogenetic relationships of AK58 and the other 5 selected wheat cultivars. **B** Pipeline for predicting epigenomic modifications and gene expression across wheat cultivars. Epigenomic modification maps and gene expression predictions can be generated for different varieties by inputting SNPs/INDELs. TIS indicates Tn5 transposase integration sites in ATAC-seq. **C** Visualization of predicted and experimental chromatin accessibility signals in spike tissues across different wheat varieties. **D** PCC of spike chromatin accessibility predictions by DeepEPI, evaluated across various wheat cultivars and AK58. **E** PCC and R^2 of spike gene expression predictions with sequence + predicted epigenomic data (Seq + pEpi) as input, compared the differentially expressed genes between AK58 and other cultivars. **F**–**G** PCC of gene expression predictions for various cultivars using sequence (Seq) and sequence + predicted epigenomic data (Seq + pEpi) as input, for spike and leaf tissues, respectively. The statistical significance of the differences in prediction performance (measured by PCC values) between groups and were calculated using the *Wilcoxon* rank-sum test. **H**–**I** Scatter plots comparing experimental expression and predicted expression using sequence + predicted epigenomic data (Seq + pEpi, left panel) and sequence only (Seq, right panel) in Zhongmai113 spike (**H**) and leaf tissues (**I**), respectively

sequence based on these INDEL/SNPs. Predicted epigenomic features were generated with DeepEPI (Fig. 4C) and integrated with DNA sequences to predict gene expression. The PCC between experimental and predicted values in spike chromatin accessibility was only 0.01–0.05 lower than that of Aikang58 (Fig. 4D), with prediction accuracy for differentially expressed genes ranging from 0.62 to 0.77 in spike and 0.50 to 0.71 in leaf tissues (Fig. 4E).

We also tested the sequence only model's (using Basenji as an example) efficiency across 5 other wheat varieties, which showed that gene expression prediction accuracy was much lower than the DeepWheat models (Fig. 4F and G). And the scatter plot further demonstrated that predicted gene expression accuracies were significantly better than those based on sequence-only models (Fig. 4H and I).

To assess the model's robustness to structural variation (SV) and copy number variation (CNV), we transferred DeepWheat to Chinese Spring and used published SV/CNV [32] datasets to divide test genes into two groups: those overlapping with SV/CNV

regions within the proximal regulatory sequence (2 kb–gene body–200 bp), and those without overlap. Despite the presence of structural variation, DeepWheat performance showed only a slight, non-significant decrease in PCC ($\Delta PCC = -0.078$ in spike, -0.018 in leaf; *Wilcoxon* rank-sum test, no significance; Additional file 1: Fig. S3C), suggesting the model can tolerate regulatory sequence variation and learn SV/CNV-related features. These results not only highlight DeepWheat's ability to predict epigenetic modifications and gene expression across different wheat varieties, but also validate its capacity to accurately predict the effects of SNPs on gene expression.

Prediction of variant effects on regulatory sequence activity and gene expression using DeepWheat

Cis-regulatory elements (CREs), primarily located in non-coding regions, are challenging to evaluate their regulatory effects [33]. Through attribution analysis, DeepWheat can identify the nucleotide bases that have the greatest impact on gene expression (Additional file 1: Fig. S4A), which is crucial for elucidating the effects of genetic variants and uncovering key regulatory sites involved in gene expression (details in the "Methods" section). To assess the effects of SNPs/INDELs within a 2000 bp upstream to 1500 bp downstream of TSS, and 500 bp upstream to 200 bp downstream of TTS window, we transferred the model across different wheat varieties and compared the predicted gene expression levels with the observed expression of genes that have SNPs/INDELs within the aforementioned windows and exhibit differential expression compared to AK58. The effect value was defined as:

Effect value = (predicted expression for each SNP/INDEL − predicted expression of AK58)/(predicted expression for all SNPs/INDELs − predicted expression of AK58).

Most SNP effects ranged from −1 to 1, with few outliers (Additional file 1: Fig. S4B and S4C). Among SNPs with effects unequal to 0, 75–85% showed significant effects, compared to 85–95% of INDELs (Additional file 1: Fig. S4D and S4E). Notably, the effect of INDELs is significantly greater than that of SNPs (Fig. 5A). Based on published spike and leaf eQTL data in wheat [34], we found that approximately 10% of the reported *cis*-eQTLs overlapped with the predicted effective regulatory variants ($|effects| > 0$) in both spike and leaf tissues (Fisher's exact test, $p < 0.001$) (Fig. 5B), indicating significant enrichment. And the predicted regulatory effect sizes of *cis*-eQTL SNPs were significantly higher than those of randomly selected SNPs (Fig. 5C), supporting the biological relevance and interpretability of our model's predictions. Further analysis revealed that effective SNPs and INDELs were predominantly enriched in promoters, followed by introns, exons, downstream regions, 3′ UTR, and 5′ UTR (Fig. 5D). Interestingly, stronger regulatory effects were observed not only in promoter regions but also in the 5′ UTR, 3′ UTR, and introns, underscoring their critical roles in modulating gene expression (Fig. 5E and Additional file 1: Fig. S4F) [35, 36]. We also found that missense SNPs had a greater effect than synonymous mutations, which may be linked not only to gene expression but also RNA stability (Additional file 1: Fig. S4G) [37]. The strong effects observed in gene bodies suggest that DeepWheat can also be applied to the abundant WES (Whole Exome Sequencing) data in wheat to identify expression-associated variants [38].

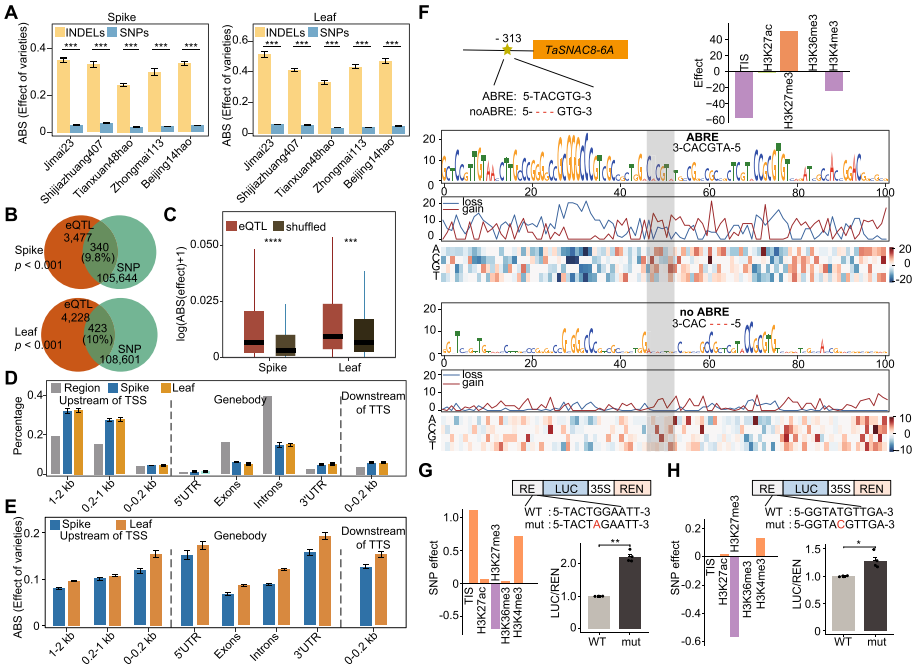

Fig. 5 Evaluation of sequence variation effects on gene expression through DeepEPI and DeepEXP. **A** Distribution of relative effects of SNPs and INDELs on gene expression in wheat varieties and tissues. Three asterisks indicate $p < 0.001$ in *Wilcoxon* rank-sum test. **B** Venn diagrams showing the overlap between published *cis*-eQTL loci [34] and SNPs with regulatory effects (|effect|> 0) in wheat spike and leaf tissues. All loci are confined to proximal regions: 2 kb upstream to 1.5 kb downstream of TSS, and 500 bp upstream to 200 bp downstream of TTS. Statistical significance was assessed using Fisher's exact test. **C** Boxplots showing the distributions of SNP effect values in wheat spike and leaf tissues for two SNP sets within proximal regions. SNPs overlapping *cis*-eQTL loci (|effect|> 0; light red) versus randomly sampled SNPs with |effect|> 0 (light blue). Boxes represent the median and interquartile range, and dots indicate outliers. Statistical significance between the two groups was tested using the *Wilcoxon* rank-sum test (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ns = not significant). **D** Percentage of effective (|effect|> 0) SNPs/INDELs in different genomic regions. **E** Effects (mean ± SEM) of SNPs/INDELs located in various genomic regions. **F** Changes in chromatin accessibility predicted by deep learning models for mutations in the promoter of *TaSNAC8-6A*. "Loss" represents reduced chromatin accessibility after the mutation, and "gain" represents increased accessibility. TIS indicates Tn5 transposase integration sites in ATAC-seq. **G**–**H** Prediction and validation of SNP effects in promoter of the *TraesAK58CH3B01G284500* (G) and *TraesAK58CH6B01G231000* (H). Transcriptional activation was assessed using a dual-luciferase reporter assay: firefly luciferase (LUC) driven by promoter variants, with Renilla luciferase (REN) as an internal control. RE, predicted regulatory element. Data are presented as mean ± SEM, and statistical significance was determined by the *Wilcoxon* rank-sum test (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ns, not significant)

Large-scale CRE editing studies, which evaluate the impact of each variant, remain a significant challenge in crops. To assess whether DeepWheat can guide CRE editing, we first evaluated the regulatory activity and gene expression changes of a known CRE associated with the drought-responsive gene *TaSNAC8-6A*. The TaABFs transcription factors can target a favorable allele, *TaSNAC8-6A*$^{In-313}$, within an inserted ABRE promoter motif (where C is replaced by CGTA), enhancing *TaSNAC8-6A* expression in drought-resistant genotypes [39]. When this variant was input into DeepWheat, the predictions indicated the formation of a regulatory motif that boosts the activity of regulatory elements and enhances gene expression (Fig. 5F). To further assess the broader applicability of DeepWheat in guiding CRE editing, we performed saturation mutagenesis on the upstream promoter regions of two differentially expressed genes,

TraesAK58CH3B01G284500 and *TraesAK58CH6B01G231000* (Additional file 1: Fig. S4H and S4I). DeepWheat predictions suggested that mutations at G-A and T-C sites could alter chromatin status and enhance gene expression, respectively, luciferase reporter assays confirmed that these mutations indeed increased the expression of both genes (Fig. 5G and H).

Discussion

Accurate gene expression prediction is a powerful tool for evaluating SNP/INDEL functions and guiding CRE editing [40]. Our study advances gene expression and regulatory sequence prediction in wheat, a species with a complex genome that has long posed challenges for genomic research. Through the development of the DeepWheat suite, which includes the DeepEXP and DeepEPI models, we have overcome many limitations of sequence-based prediction approaches.

Spatiotemporal-specific gene expression is of fundamental importance in the development of superior traits [41, 42]. Precise modulation of tissue-specific gene expression is imperative for gene editing in the creation of elite materials [43–45]. DeepEXP integrates diverse epigenomic and transcriptomic datasets, achieving impressive predictive performance with PCC over 0.82 across various tissues and developmental stages, notably outperforming sequence-only models. While models like DeepEPI excel in predicting epigenomic features, the strategy employed by DeepWheat further enhances prediction accuracy across different cultivars by integrating sequence data with epigenomic features inferred by DeepEPI. Overall, these integrative models have markedly improved the precision of gene expression predictions, providing a more robust tool for elucidating gene regulatory mechanisms.

In addition to gene expression prediction, DeepWheat models offer several key functions for functional genomic studies and crop improvement. These include identifying regulatory sequences, assessing the impact of sequence variations on regulatory activity and gene expression, and performing saturation mutagenesis to identify high-effect sites. These capabilities deepen our understanding of how genetic variations affect regulatory networks and phenotypic traits [46, 47]. Furthermore, by predicting the regulatory effects of noncoding variants across tissues, DeepWheat provides a useful tool for prioritizing candidate mutations beyond coding regions. Virtual mutagenesis based on DeepWheat can simulate regulatory edits in silico, reducing experimental workload and guiding precise genome editing designs. These features support the development of new phenotypes and efficient breeding strategies through targeted regulatory interventions.

While generating high-quality epigenomic data (e.g., ATAC-seq or ChIP-seq) is more costly than transcriptomic profiling, we found that even single-tissue epigenomic profiles, when integrated with sequence features, substantially improved prediction accuracy compared to sequence-only models. Cross-tissue applications resulted in a modest performance decline, yet still outperformed sequence-only baselines, particularly when the tissues were closely related. Notably, DeepWheat demonstrated strong generalizability across multiple wheat varieties, including those lacking reference-quality genomes. Nevertheless, for optimal accuracy, retraining or fine-tuning on tissue-specific data remains advisable.

Conclusions

Future work should focus on expanding the epigenomic datasets used for training these models, incorporating data from a broader range of wheat varieties and environmental conditions. Additionally, integrating 3D chromatin structure data could further enhance prediction accuracy and provide insights into the spatial organization of regulatory elements [48, 49].

These efforts will be essential for refining predictive models and enhancing their applicability across different genetic backgrounds and environmental conditions. Overall, DeepWheat provides a versatile framework for predicting gene expression and regulatory activity, prioritizing candidate mutations, and guiding precision genome editing, offering promising applications in wheat functional genomics and breeding.

Methods

ATAC-seq, ChIP-seq, and RNA-seq

ATAC-seq

ATAC-seq was performed as previously described with minor modifications [50]. Briefly, 1–2 g of flash-frozen wheat tissue (7 days seedlings and 1–2 cm young spike) was minced in 1 mL of ice-cold lysis buffer (15 mM Tris–HCl, pH 7.5; 20 mM NaCl; 80 mM KCl; 0.5 mM spermine; 5 mM 2-mercaptoethanol; 0.2% Triton X-100). The crude nuclei extract was filtered twice through a 40 μm filter, stained with DAPI (Sigma, D9542), and sorted using a BD FACSCanto flow cytometer. Nuclei were pelleted by centrifugation, washed with Tris-Mg buffer (10 mM Tris–HCl, pH 8.0; 5 mM $MgCl_2$), and incubated with 3.5 μL Tn5 transposomes in 40 μL TTBL buffer (TruePrep DNA Library Prep Kit V2, Vazyme, TD501) at 37 °C for 30 min. The DNA integration products were purified using the NEB Monarch™ DNA Cleanup Kit (T1030S) and amplified for 10–13 PCR cycles using NEBNext Ultra II Q5 Master Mix (M0544L). Amplified libraries were purified with Hieff NGS® DNA Selection Beads (Yeasen, 12601ES03).

ChIP-seq

ChIP-seq experiments were conducted following established protocols [51], using antibodies specific for H3K27ac (Abclonal, A7253), H3K27me3 (Abcam, ab6002), H3K4me3 (Millipore, 07–473), and H3K36me3 (Abcam, ab9050). Library preparation was performed with the TransGen Biotech Kit (KP201-02), The libraries were sequenced on the Illumina NovaSeq 6000 platform, generating 150 bp paired-end reads.

RNA-seq

Wheat samples were flash-frozen in liquid nitrogen, and total RNA was extracted using TRIzol™ Reagent (Invitrogen, 15,596–026) according to the manufacturer's instructions. RNA-seq libraries were prepared by Berry Genomics (Beijing, China) and sequenced on the Illumina NovaSeq 6000 platform, generating 150 bp paired-end reads.

ATAC-seq, ChIP-seq, and RNA-seq data analysis

ATAC-seq data processing

Raw sequencing reads were subjected to quality control using fastp v0.21.0 [52]. High-quality reads were then aligned to the wheat reference genome using Bowtie2 v2.3.5

[53] with parameters -X 1000 and –very-sensitive. The resulting alignments were sorted and filtered (MAPQ > 10) using SAMtools v1.3.1 [54]. Clonal duplicates were removed with Picard v2.16.0 (http://broadinstitute.github.io/picard/). Peak calling was performed using MACS2 v2.2.6 [55] with the following parameters: –keep-dup all –nomodel –extsize 150 –shift −75. Differential peaks were found using MACS2 v2.2.6 [55] with the following parameters: bdgdiff –d1 60 –d2 120 -c 3.

ChIP-seq data processing

Raw sequencing reads underwent quality control with fastp v0.21.0 [52]. Filtered reads were aligned to the wheat reference genome using Bowtie2 v2.3.5 [53] with parameters -X 1000 and –very-sensitive. Aligned reads were sorted and filtered (MAPQ > 10) using SAMtools v1.3.1 [54], and clonal duplicates were removed using Picard v2.16.0 (http://broadinstitute.github.io/picard/). Peaks were identified using MACS2 v2.2.6 [55] with the parameters –keep-dup all –nomodel –extsize 150 –shift −75. Differential peaks were found using MACS2 v2.2.6 with the following parameters: bdgdiff –d1 60 –d2 120 -c 3.

RNA-seq data processing

Quality control of raw reads was conducted using fastp v0.21.0 [52]. Clean reads were aligned to the wheat reference genome using HISAT2 (https://daehwankimlab.github.io/hisat2/) with default settings. Sorted BAM files were generated using SAMtools v1.3.1 [54], and read quantification was performed using featureCounts [56]. Transcripts Per Million values (TPM) were calculated using TPMCalculator [57]. Differentially expressed genes (DEGs) were identified using DESeq2 [58], with significance thresholds set at p-value < 0.01 and fold change > 2.

DeepEXP architecture and training

Model input and output

The model takes as input DNA sequences flanking the transcription start site (TSS) and transcription termination site (TTS) of genes, along with various epigenomic modification signals. The output is the gene expression level, represented by log1p (TPM). To ensure model reliability, the epigenomic and gene expression data are derived from the same tissue sample batch. DNA sequences are encoded using one-hot encoding, resulting in an $L \times 4$ matrix, where L represents the sequence length. Epigenomic data, including accessible chromatin regions and peaks with histone modifications identified by MACS2 [55] (https://github.com/macs3-project/MACS) and BEDTools [59] were used; if an interval falls within an ACR or peaks, it will be assigned a modified coordinates value, or zero otherwise. The data are then normalized to fall within the range of 0 to 1.

Model architecture

DeepEXP employs a multi-branch architecture where epigenomic and DNA sequence data are processed independently through convolutional layers and pooled before concatenation along the channel dimension (Fig. 1). The epigenomic branch reshapes the input into a 4D tensor, applies two convolutional layers with kernel sizes (kernel_size, 5) and (kernel_size, 1), and concludes with batch normalization

(BN), ReLU activation, dropout, and average pooling. Similarly, the DNA sequence branch uses convolutional layers with kernel sizes (kernel_size, sequence_input_data.shape) and (kernel_size, 1), followed by identical normalization and pooling steps. The extracted features are concatenated and passed through r residual modules, each containing two convolutional layers with kernel size (res_block_kernel_size, res_block_kernel_size), and employing skip connections to preserve lower-level features while extracting higher-dimensional representations. Flattened concatenated features are fed into three fully connected layers with 256, 64, and 16 neurons, respectively, each equipped with BN, ReLU activation, and dropout, and finally a single-neuron output layer predicts gene expression values. The model uses Optuna (https://optuna.org/) for hyperparameter optimization, tuning learning rate, kernel sizes, filter numbers, residual block configurations, L2 regularization, and dropout rates, with the objective to maximize R^2 on the validation set. Training is performed using the Adam optimizer with dynamically adjusted learning rates, and early stopping (patience = 40) and model checkpointing are employed to save the best weights. The model's performance is evaluated on the test set using R^2 and Pearson correlation coefficients (PCC), demonstrating its ability to effectively integrate epigenomic and DNA sequence data for accurate gene expression prediction.

Model training For each dataset corresponding to a specific tissue or developmental stage, the data is randomly shuffled and divided into training (60%), validation (20%), and test (20%) sets. To achieve optimal predictive performance, Optuna (https://optuna.org/) is utilized to perform hyperparameter optimization and training for each tissue and time point individually, leveraging an NVIDIA RTX 4090 GPU. The detailed code for the model architecture and training process can be accessed in the following repository: https://github.com/WheatEpigenomics/DeepWheat.

DeepEPI model analysis

The DeepEPI model is based on the previously published epigenetic and transcriptomic prediction framework, Basenji2 [25, 26], as shown in Fig. 1. For input preparation, epigenomic data in BigWig format were generated using the bam_cov.py script and processed with the basenji_data.py script. To ensure valid sampling intervals, genomic regions with gaps larger than 10 bp were annotated in the wheat AK58 reference genome.

The epigenomic data used in this study include ATAC-seq and four types of histone modifications data from six tissues and developmental stages. We trained separate models for ATAC-seq and histone modification data using basenji_train.py. The training set comprised 70% of the total samples, while the validation and test sets made up 15% each. Due to limited computational resources (using an NVIDIA RTX 3090 for model training), we referred to hyperparameters from previously published high-precision human epigenomic prediction models based on Basenji2 [25, 26] and adjusted these parameters accordingly. Testing was conducted on an independent test set using the basenji_test.py script, with the optimal hyperparameter configuration available at https://github.com/WheatEpigenomics/DeepWheat.

Comparison with other methods

Gene expression prediction using Basenji2

Initially, we trained a model to predict gene expression profiles using the Basenji2 [25, 26] framework. Subsequently, we used BEDTools [59] to intersect the gene annotation files from the high-quality AK58 dataset with the independent test set to obtain the gene expression profiles for the test set. The gene expression profiles were then converted into BAM files, similar to RNA-seq data, using BEDTools [59] and SAMtools [54]. TPM values were calculated using TPMCalculator [57]. Finally, we evaluated the model's accuracy by calculating the R^2 and Pearson correlation coefficient (PCC) between the experimental and predicted gene expression values.

Gene expression prediction with Xpresso and PhytoExpr

Scripts and model definitions for Xpresso [27] were obtained from the authors' GitHub repository, and the PhytoExpr code was downloaded from the Zenodo archive. Each model was re-trained using the default hyperparameters recommended in its original publication on multi-tissue gene expression datasets from four plant species: wheat (*Triticum aestivum*), *Arabidopsis thaliana*, rice (*Oryza sativa*), and maize (*Zea mays*). Input features and preprocessing strictly followed the published pipelines: Xpresso [27] leveraged promoter-based and sequence-derived features, while PhytoExpr used sequence and GC-content features. Models were trained independently for each species with their recommended learning rates, batch sizes, and regularization settings, and early stopping based on validation loss was applied to prevent overfitting. Finally, predictive performance was evaluated on held-out test sets for each species by calculating Pearson's correlation coefficient (PCC) and coefficient of determination (R^2) between predicted and observed expression levels.

Gene expression prediction by combination of DeepEPI and DeepEXP

The DeepEPI model takes input DNA sequences of length 131,072 bp (based on the reference genome sequence and input BED files). If the submitted BED interval is shorter than 131,072 bp, it is expanded symmetrically to 131,072 bp. This allows for the generation of various epigenomic profiles within the interval. These epigenomic profiles, along with the sequences of the target regions upstream and downstream of gene TSS and TTS, are then input into the DeepEXP model to obtain the predicted gene expression values.

Epigenomic profiling and gene expression prediction across different wheat varieties

For different wheat varieties, high-quality VCF files and the AK58 reference genome sequence were used with g2gtools v0.2 (https://github.com/churchill-lab/g2gtools/) to generate pseudo-reference genome sequences for each variety. These pseudo-reference sequences, along with the BED files of the target prediction regions, were then used as inputs for DeepEPI to obtain the epigenomic profiles of the target prediction regions in different varieties. To predict gene expression, the epigenomic profiles and sequences from the target regions upstream and downstream of the gene TSS and

TTS are input into DeepEXP. This allows for the prediction of gene expression levels for the target genes across different wheat varieties.

Epigenomic profiling and gene expression prediction for a single gene

For single-gene analysis, users provide the gene ID list to DeepEXP. The script "predict_gene_expression-IG_analysis.py" predicts gene expression levels and identifies key sequence or epigenomic regions contributing to expression using integrated gradients attribution (default 100 steps: python predict_gene_expression-IG_analysis.py –seq gene.seq.tsv –epi_dir epigenomic_data –predict_list predict_gene.list –attrib_list AA_gene.list –model_dir model –out_pred pred_results –out_ig IG_results –ig_steps 100). Inputs include sequence files (gene.seq.tsv), multi-tissue epigenomic data (epi_dir), prediction and attribution gene lists, and trained models.

Epigenomic profiling and gene expression prediction in incompletely annotated genomes

For cases where the reference genome annotation is incomplete, we provide the "incompletely_annotated_genome_prediction" module. Users supply the target DNA sequences in FASTA format along with corresponding relative gene position information in BED format. These inputs are fed into the DeepEPI model to predict epigenomic modifications within the specified regions. Additionally, DeepEPI's in silico saturation mutagenesis function enables assessment of the impact of variants on epigenomic features in these regions. Subsequently, the DeepEXP model can be used to predict gene expression levels and evaluate the contribution of proximal regulatory regions to gene expression.

Assessment of variant effects on regulatory sequence activity and gene expression through saturation mutagenesis analysis

DeepEPI, based on Basenji2 [25, 26], is used to assess the impact of variants on epigenomic modifications. To evaluate variant effects, we input VCF files containing variants into DeepEPI. The Basenji_sad.py script predicts the epigenomic signals for both reference and variant-containing sequences, with the difference representing the variant effect. Given that different epigenomic modifications are linked with regulatory sequence activity [60, 61], the magnitude of the variant effect can provide insights into its impact on regulatory sequence activity.

For scoring the impact of variants on gene expression, we input the epigenomic profiles and sequences of both reference and variant sequences into DeepEXP. The difference in gene expression levels between these sequences provides the variant's effect score on gene expression. When performing saturation mutagenesis analysis of regulatory sequences, we use BED files or VCF files containing the regulatory intervals. By calculating and plotting the differences between reference and variant signals predicted by DeepEPI at each site, we generate a heatmap of variant effects from saturation mutagenesis.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03809-x.

Additional file 1: All supplementary figures included in this article. Fig. S1. Benchmarking model performance and evaluating epigenomic feature contributions to gene expression prediction. Fig. S2. Evaluation of Epigenomic

Feature Prediction and Its Application to Tissue-Specific Gene Expression Modeling. Fig. S3. Impact of chromatin accessibility and structural variants on gene expression and model prediction across wheat varieties. Fig. S4. Integrated Gradientsanalysis and the effects of SNP/INDEL variants on gene expression and epigenomic features.

Additional file 2: All supplementary tables included in this article. Table S1. Optimization of Input Regions for Gene Expression Prediction Around TSS and TTS. Table S2. Pearson correlation coefficientsand coefficients of determinationbetween predicted and experimentally measured gene expression levels across different plant tissues. Table S3. Pearson correlation coefficientsand coefficients of determinationfor predictions and tests of epigenomic signal in dependent test sets. Table S4. Performance of different models in predicting gene expression across multiple tissues in the Chinese Spring wheat variety.

Acknowledgements
We thank Dr. Jizeng Jia, Dr. Guangyao Zhao, and Dr. Lifeng Gao from the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, for their valuable suggestions on the project design, and Ting Li from the flow cytometry core of the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, for supporting the nuclear sorting.

Peer review information
Qingxin Song and Wenjing She were the primary editors 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
Z.L. conceptualized and supervised the project. Z.M. designed and trained the deep learning models, performed performance evaluations, and conducted additional bioinformatic analyses. J.Z. and H.P. carried out the molecular biology experiments. J.Z., Y.L. and L.W. were responsible for sample collection and data acquisition. The manuscript was written and finalized by Z.L. and Z.M., L.W. and H.T. contributed to the revision of the manuscript. All authors read and approved the final manuscript.

Funding
This project was financially supported by the National Key Research and Development Project (2023YFF1000403 and 2022YFF1002903), the Outstanding Young Scientist Foundation of NSFC (Overseas), Chinese Academy of Agricultural Sciences Young Talent Scientist Program and Agricultural Science and Technology Innovation Program, the China Agricultural Research System (Grant No. CARS-03).

Data availability
The AK58 reference genome and annotation were obtained from the NCBI Genome Database (https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_025895885.1/). ChIP-seq datasets from AK58 spike tissue, ATAC-seq datasets from spike tissue of AK58 and five additional wheat varieties, and RNA-seq datasets from spike and leaf tissues of AK58 and these five varieties, together with ChIP-seq datasets from developing seeds at DAP-9, DAP-7, and DAP-5 and RNA-seq and ATAC-seq datasets from developing seeds at DAP-7, have been deposited in the Gene Expression Omnibus (GEO) and the Sequence Read Archive (SRA) under accession numbers GSE289179 [62], GSE287695 [63], and PRJNA1320958 [64], respectively. In addition, datasets comprising ATAC-seq, ChIP-seq, and RNA-seq from vernalized leaf and leaf tissues of AK58, as well as ATAC-seq and RNA-seq from developing seeds at DAP-9 and DAP-5, were obtained from previous studies [17, 65], with corresponding GEO accession numbers GSE232430 [66] and GSE214739 [67].
All scripts used for data processing, model construction, training, and prediction in this study are publicly available on GitHub (https://github.com/WheatEpigenomics/DeepWheat) [68], under the MIT License. To ensure long-term accessibility and reproducibility, the trained models have also been archived on Zenodo under the DOI: https://doi.org/10.5281/zenodo.15761553 [69].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 6 February 2025 Accepted: 22 September 2025
Published online: 29 September 2025

References

1. Wittkopp PJ, Kalay G. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet. 2011;13:59–69.
2. Marand AP, Eveland AL, Kaufmann K, Springer NM. *Cis*-regulatory elements in plant development, adaptation, and evolution. Annu Rev Plant Biol. 2023;74:111–37.
3. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers. Nat Biotechnol. 2012;30:265–70.

4. Park YJ, Kellis M. Deep learning for regulatory genomics. Nat Biotechnol. 2015;33:825–6.
5. Avsec Z, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18:1196–203.
6. Peleke FF, Zumkeller SM, Gültas M, Schmitt A, Szymański J. Deep learning the *cis*-regulatory code for gene expression in selected model plants. Nat Commun. 2024;15:3488.
7. Eraslan G, Avsec Z, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20:389–403.
8. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12:931–4.
9. Zhang TQ, Xu ZG, Shang GD, Wang JW. A single-cell RNA sequencing profiles the developmental landscape of Arabidopsis root. Mol Plant. 2019;12:648–60.
10. Zhao H, Tu Z, Liu Y, Zong Z, Li J, Liu H, et al. PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. Nucleic Acids Res. 2021;49:W523–9.
11. Wang Z, Peng Y, Li J, Li J, Yuan H, Yang S, et al. DeepCBA: a deep learning framework for gene expression prediction in maize based on DNA sequences and chromatin interactions. Plant Commun. 2024;5:100985.
12. Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 2018;361:eaar7191.
13. Uauy C, Wulff BBH, Dubcovsky J. Combining traditional mutagenesis with new high-throughput sequencing and genome editing to reveal hidden variation in polyploid wheat. Annu Rev Genet. 2017;51:435–54.
14. Gao S, Rehman J, Dai Y. Assessing comparative importance of DNA sequence and epigenetic modifications on gene expression using a deep convolutional neural network. Comput Struct Biotechnol J. 2022;20:3814–23.
15. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92.
16. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128:693–705.
17. Liu Y, Liu P, Gao L, Li Y, Ren X, Jia J, et al. Epigenomic identification of vernalization *cis*-regulatory elements in winter wheat. Genome Biol. 2024;25:200.
18. Schmitz RJ, Marand AP, Zhang X, Mosher RA, Turck F, Chen XM, et al. Quality control and evaluation of plant epigenomics data. Plant Cell. 2022;34:503–13.
19. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science. 2015;348:660–5.
20. Zhang X, Meng W, Liu D, Pan D, Yang Y, Chen Z, et al. Enhancing rice panicle branching and grain yield through tissue-specific brassinosteroid inhibition. Science. 2024;383:eadk8838.
21. He C, Bi S, Li Y, Song C, Zhang H, Xu X, et al. Dynamic atlas of histone modifications and gene regulatory networks in endosperm of bread wheat. Nat Commun. 2024;15:9572.
22. Jiao C, Xie X, Hao CY, Chen LY, Xie YX, Garg V, et al. Pan-genome bridges wheat structural variations with habitat and breeding. Nature. 2025;637:384–93.
23. Li T, Xu H, Teng S, Suo M, Bahitwa R, Xu M, et al. Modeling 0.6 million genes for the rational design of functional *cis*-regulatory variants and *de novo* design of *cis*-regulatory sequences. Proc Natl Acad Sci U S A. 2024;121:e2319811121.
24. Qiu Y, Liu L, Yan J, Xiang X, Wang S, Luo Y, et al. Precise engineering of gene expression by editing plasticity. Genome Biol. 2025;26:51.
25. Kelley DR. Cross-species regulatory sequence activity prediction. PLoS Comput Biol. 2020;16:e1008050.
26. 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.
27. Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. Cell Rep. 2020;31(7):107663.
28. Lal A, Chiang ZD, Yakovenko N, Duarte FM, Israeli J, Buenrostro JD. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun. 2021;12:1507.
29. Wang G, Li X, An Y, Zhang J, Li H. Transient ChIP-seq for genome-wide *in vivo* DNA binding landscape. Trends Plant Sci. 2021;26:524–5.
30. Zhu D, Wen Y, Tan Y, Chen X, Wu Z. A simple, robust, cost-effective, and low-input ChIP-seq method for profiling histone modifications and Pol II in plants. New Phytol. 2024;244:1658–69.
31. Lüleci HB, Yilmaz A. Robust and rigorous identification of tissue-specific genes by statistically extending tau score. Biodata Min. 2022;15:31.
32. Jia J, Zhao G, Li D, Wang K, Kong C, Deng P, et al. Genome resources for the elite bread wheat cultivar Aikang 58 and mining of elite homeologous haplotypes for accelerating wheat improvement. Mol Plant. 2023;16:1893–910.
33. Lyu R, Gao Y, Wu T, Ye C, Wang P, He C. Quantitative analysis of *cis*-regulatory elements in transcription with KAS-ATAC-seq. Nat Commun. 2024;15:6852.
34. He F, Wang W, Rutter WB, Jordan KW, Ren J, Taagen E, et al. Genomic variants affecting homoeologous gene expression dosage contribute to agronomic trait variation in allopolyploid wheat. Nat Commun. 2022;13:826.
35. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. Science. 2016;352:1413–6.
36. Wissink EM, Fogarty EA, Grimson A. High-throughput discovery of post-transcriptional *cis*-regulatory elements. BMC Genomics. 2016;17:177.
37. Wu H, Yu H, Zhang Y, Yang B, Sun W, Ren L, et al. Unveiling RNA structure-mediated regulations of RNA stability in wheat. Nat Commun. 2024;15:10042.
38. Cagirici HB, Akpinar BA, Sen TZ, Budak H. Multiple variant calling pipelines in wheat whole exome sequencing. Int J Mol Sci. 2021;22:10400.
39. Mao H, Li S, Wang Z, Cheng X, Li F, Mei F, et al. Regulatory changes in *TaSNAC8-6A* are associated with drought tolerance in wheat seedlings. Plant Biotechnol J. 2019;18:1078–92.
40. Benegas G, Batra SS, Song YS. DNA language models are powerful predictors of genome-wide variant effects. Proc Natl Acad Sci U S A. 2023;120:e2311219120.

41. Xu X, Crow M, Rice BR, Li F, Harris B, Liu L, et al. Single-cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery. Dev Cell. 2021;56:557-68.e6.
42. Zhang TQ, Chen Y, Liu Y, Lin WH, Wang JW. Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root. Nat Commun. 2021;12:2053.
43. Gao CX. Genome engineering for crop improvement and future agriculture. Cell. 2021;184:1621–35.
44. Liu L, Gallagher J, Arevalo ED, Chen R, Skopelitis T, Wu Q, et al. Enhancing grain-yield-related traits by CRISPR-Cas9 promoter editing of maize genes. Nat Plants. 2021;7:287–94.
45. Rodríguez-Leal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB. Engineering quantitative trait variation for crop improvement by genome editing. Cell. 2017;171:470-80.e8.
46. Wang H, Cimen E, Singh N, Buckler E. Deep learning for plant genomics and crop improvement. Curr Opin Plant Biol. 2020;54:34–41.
47. Akagi T, Masuda K, Kuwada E, Takeshita K, Kawakatsu T, Ariizumi T, et al. Genome-wide *cis*-decoding for expression design in tomato using cistrome data and explainable deep learning. Plant Cell. 2022;34:2174–87.
48. Khunsriraksakul C, McGuire D, Sauteraud R, Chen F, Yang L, Wang L, et al. Integrating 3D genomic and epigenomic data to enhance target gene discovery and drug repurposing in transcriptome-wide association studies. Nat Commun. 2022;13:3258.
49. Ballard JL, Wang Z, Li W, Shen L, Long Q. Deep learning-based approaches for multi-omics data integration and analysis. BioData Min. 2024;17:38.
50. Lu Z, Hofmeister BT, Vollmers C, DuBois RM, Schmitz RJ. Combining ATAC-seq with nuclei sorting for discovery of *cis*-regulatory regions in plant genomes. Nucleic Acids Res. 2017;45:e41.
51. Wang H, Liu C, Cheng J, Liu J, Zhang L, He C, et al. *Arabidopsis* flower and embryo developmental genes are repressed in seedlings by different combinations of polycomb group proteins in association with distinct sets of *cis*-regulatory elements. PLoS Genet. 2016;12:e1005771.
52. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:884–90.
53. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
54. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
55. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
56. Liao Y, Smyth GK, Shi W. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
57. Vera Alvarez R, Pongor LS, Mariño-Ramírez L, Landsman D. TPMcalculator: one-step software to quantify mrna abundance of genomic features. Bioinformatics. 2019;35:1960–2.
58. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
59. Quinlan AR, Hall IM. BEDtools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
60. Preissl S, Gaulton KJ, Ren B. Characterizing *cis*-regulatory elements using single-cell epigenomics. Nat Rev Genet. 2022;24:21–43.
61. Lane AK, Niederhuth CE, Ji L, Schmitz RJ. pENCODE: a plant encyclopedia of DNA elements. Annu Rev Genet. 2014;48:49–70.
62. Ma Z, Zhang J, Pei H, Liu Y, Tong H, Wang L, et al. DeepWheat: predicting the effects of genomic variants on gene expression and regulatory activities across tissues and varieties in wheat using deep learning. Datasets. Gene Expression Omnibus. 2025. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE289179.
63. Ma Z, Zhang J, Pei H, Liu Y, Tong H, Wang L, et al. DeepWheat: predicting the effects of genomic variants on gene expression and regulatory activities across tissues and varieties in wheat using deep learning. Datasets. Gene Expression Omnibus. 2025. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE287695.
64. Ma Z, Zhang J, Pei H, Liu Y, Tong H, Wang L, et al. DeepWheat: predicting the effects of genomic variants on gene expression and regulatory activities across tissues and varieties in wheat using deep learning. Datasets. NCBI Sequence Read Archive. 2025. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1320958/.
65. Pei H, Li Y, Liu Y, Liu P, Zhang J, Ren X, et al. Chromatin accessibility landscapes revealed the subgenome-divergent regulation networks during wheat grain development. aBIOTECH. 2023;4:8–19.
66. Liu Y, Liu P, Gao L, Li Y, Ren X, Jia J, et al. Epigenomic identification of vernalization cis-regulatory elements in winter wheat. Datasets. Gene Expression Omnibus. 2024. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE232430.
67. Pei H, Li Y, Liu Y, Liu P, Zhang J, Ren X, et al. Chromatin accessibility landscapes revealed the subgenome-divergent regulation networks during wheat grain development. Datasets. Gene Expression Omnibus. 2023. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE214739.
68. Ma Z, Zhang J, Pei H, Liu Y, Tong H, Wang L, et al. DeepWheat: predicting the effects of genomic variants on gene expression and regulatory activities across tissues and varieties in wheat using deep learning. GitHub. 2025. https://github.com/WheatEpigenomics/DeepWheat.
69. Ma Z, Zhang J, Pei H, Liu Y, Tong H, Wang L, et al. DeepWheat: predicting the effects of genomic variants on gene expression and regulatory activities across tissues and varieties in wheat using deep learning. 2025. Zenodo. https://doi.org/10.5281/zenodo.15761553.

Publisher's Note