PNAS RESEARCH ARTICLE | PLANT BIOLOGY
BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Modeling 0.6 million genes for the rational design of functional *cis*-regulatory variants and de novo design of *cis*-regulatory sequences

Tianyi Li[a,1], Hui Xu[a,1], Shouzhen Teng[a,1], Mingrui Suo[a], Revocatus Bahitwa[a,b], Mingchi Xu[a], Yiheng Qian[a], Guillaume P. Ramstein[c], Baoxing Song[d,e], Edward S. Buckler[f,g], and Hai Wang[a,h,i,2]

Affiliations are included on p. 9.

Edited by Daniel Voytas, University of Minnesota Twin Cities, Saint Paul, MN; received November 12, 2023; accepted May 14, 2024

Rational design of plant *cis*-regulatory DNA sequences without expert intervention or prior domain knowledge is still a daunting task. Here, we developed PhytoExpr, a deep learning framework capable of predicting both mRNA abundance and plant species using the proximal regulatory sequence as the sole input. PhytoExpr was trained over 17 species representative of major clades of the plant kingdom to enhance its generalizability. Via input perturbation, quantitative functional annotation of the input sequence was achieved at single-nucleotide resolution, revealing an abundance of predicted high-impact nucleotides in conserved noncoding sequences and transcription factor binding sites. Evaluation of maize HapMap3 single-nucleotide polymorphisms (SNPs) by PhytoExpr demonstrates an enrichment of predicted high-impact SNPs in *cis*-eQTL. Additionally, we provided two algorithms that harnessed the power of PhytoExpr in designing functional *cis*-regulatory variants, and de novo creation of species-specific *cis*-regulatory sequences through in silico evolution of random DNA sequences. Our model represents a general and robust approach for functional variant discovery in population genetics and rational design of regulatory sequences for genome editing and synthetic biology.

transcriptional regulation | deep learning | *cis*-engineering | synthetic biology

A comprehensive understanding of the genetic commands embedded within the *cis*-regulatory sequences is crucial for accurate discovery and design of functional *cis*-acting variants, as well as de novo creation of *cis*-regulatory sequences. Nonetheless, the immense variation space of *cis*-regulatory sequences and relatively limited knowledge of their associated syntax rules present translational challenges (1). In the last 20 y, advances in molecular biology, population genetics, and comparative genomics have provided critical insights into the mechanism of *cis*-elements and their regulatory logic in plants.

In the field of molecular biology, high-throughput sequencing-based methods, such as chromatin immunoprecipitation followed by sequencing (ChIP-Seq), assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-Seq), and Hi-C, have been widely utilized in diverse plant species to annotate functional *cis*-elements (2, 3). However, the volume of data is relatively small compared to that of the ENCODE project in human and mouse genome research (4). Measuring the expression of synthetic regulatory sequences in massively parallel reporter assays has proved successful in plants, particularly crop species (5–8). These datasets offer valuable insights into the intricate architecture of *cis*-regulatory sequences in plants. Nevertheless, it is worth noting that several novel mechanisms initially discovered in animal and human cells are relatively underexplored in plants. These include enhancer–promoter interactions reliant on phase separation (9), transcriptional bursting (10), specificity and robustness conferred by suboptimal transcription factor binding sites (TFBSs) (11), and the trade-off between precision and economy in gene expression (12).

In the realm of crop population genetics, a considerable number of *cis*-regulatory variants underlying crop domestication and improvement traits have been cloned and functionally validated (13). *Cis*-expression quantitative trait loci (*cis*-eQTL) was also identified in several crop populations (14, 15). However, given finite population size, association and linkage analysis can only detect common alleles exerting large effects, leaving rare alleles or small-effect alleles largely unexplored. Furthermore, precise identification of causal variants is often challenging due to the presence of linkage disequilibrium. Comparative genomics has aided the discovery of *cis*-elements at a broader scale. Through comparisons of phylogenetically related plant species, conserved noncoding sequences (16, 17) and deleterious *cis*-variants (18, 19) have been identified. This technique however requires a substantial number of

Significance

The enormous variation space and obscure syntax rules of eukaryotic transcriptional regulatory DNA sequences hamper their rational design. Here, we developed PhytoExpr, a deep learning framework that reads regulatory DNA sequences to predict their messenger ribonucleic acid (mRNA) abundance and also the plant species they are from. PhytoExpr was trained over major clades of the plant kingdom to make predictions on unseen gene families from unseen species. The sequence features learned by PhytoExpr were enriched with conserved noncoding sequences, transcription factor binding sites, and eQTLs. We also fit PhytoExpr into two algorithms for the rational design of functional *cis*-regulatory variants for genome editing, as well as the de novo design of species-specific *cis*-regulatory DNA sequences for synthetic biology.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] T.L., H.X., and S.T. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: wanghai@cau.edu.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2319811121/-/DCSupplemental.

Published June 18, 2024.

genomes for optimal sensitivity and accuracy (20). In addition, deleterious *cis*-variants were found associated with dysregulation of gene expression and reduced overall fitness (15).

Despite the limited knowledge of *cis*-regulatory logic desired for precise manipulation of gene expression, promoter editing has been successfully applied in crop improvement (21–26). Commonly a large pool of diverse *cis*-regulatory alleles was generated, followed by identification of favorable alleles through resource-intensive phenotypic selection. De novo design of regulatory DNA sequences for plant synthetic biology presents yet another formidable challenge. Simple organisms such as *E. coli* and yeast exhibit short edit-distances between random sequences and functional promoters (27–29), facilitating efficient de novo promoter design by deep generative networks (30, 31). However, in more complex organisms such as higher plants, de novo design of *cis*-regulatory sequences is much more complex due to their extensive lengths and an absence of suitable computational tools.

In this study, we trained and tested a deep learning framework termed PhytoExpr over ~0.6 million regulatory sequences from 17 plant species to implicitly learn both general and species-specific regulatory syntax rules. Remarkably, PhytoExpr managed to identify sequence features aligning with functional annotations, such as conserved noncoding sequences, TFBSs, and *cis*-eQTL even though it had never been exposed on such datasets. Furthermore, we showed that PhytoExpr can be effectively used for functional *cis*-variant discovery and design, as well as de novo design of regulatory sequences.

Results

Modeling Proximal *cis*-Regulatory Sequences Using Solely Sequence Information and Gene Expression Data. In the case of the human and mouse, accurate prediction of gene expression levels from *cis*-regulatory sequences, as demonstrated by ExPecto (32), Basenji (33, 34), and Enformer (35), requires thousands of high-quality annotation datasets detailing noncoding regions, such as histone marks, TFBSs, and DNA accessibility profiles. Such strategies are not readily applicable to plants, as datasets of such scale, magnitude, and quality are yet to be developed. Therefore, we aimed to develop models capable of forecasting mRNA abundance solely using transcriptome datasets. To better capture the enormous variation space of *cis*-regulatory DNA sequence in the plant kingdom, ~0.6 million protein-coding genes from 17 phylogenetically diverse plant species (Fig. 1*A*, summarized in Dataset S1) were used to train and test the models in this study (referred to as PhytoExpr hereafter).

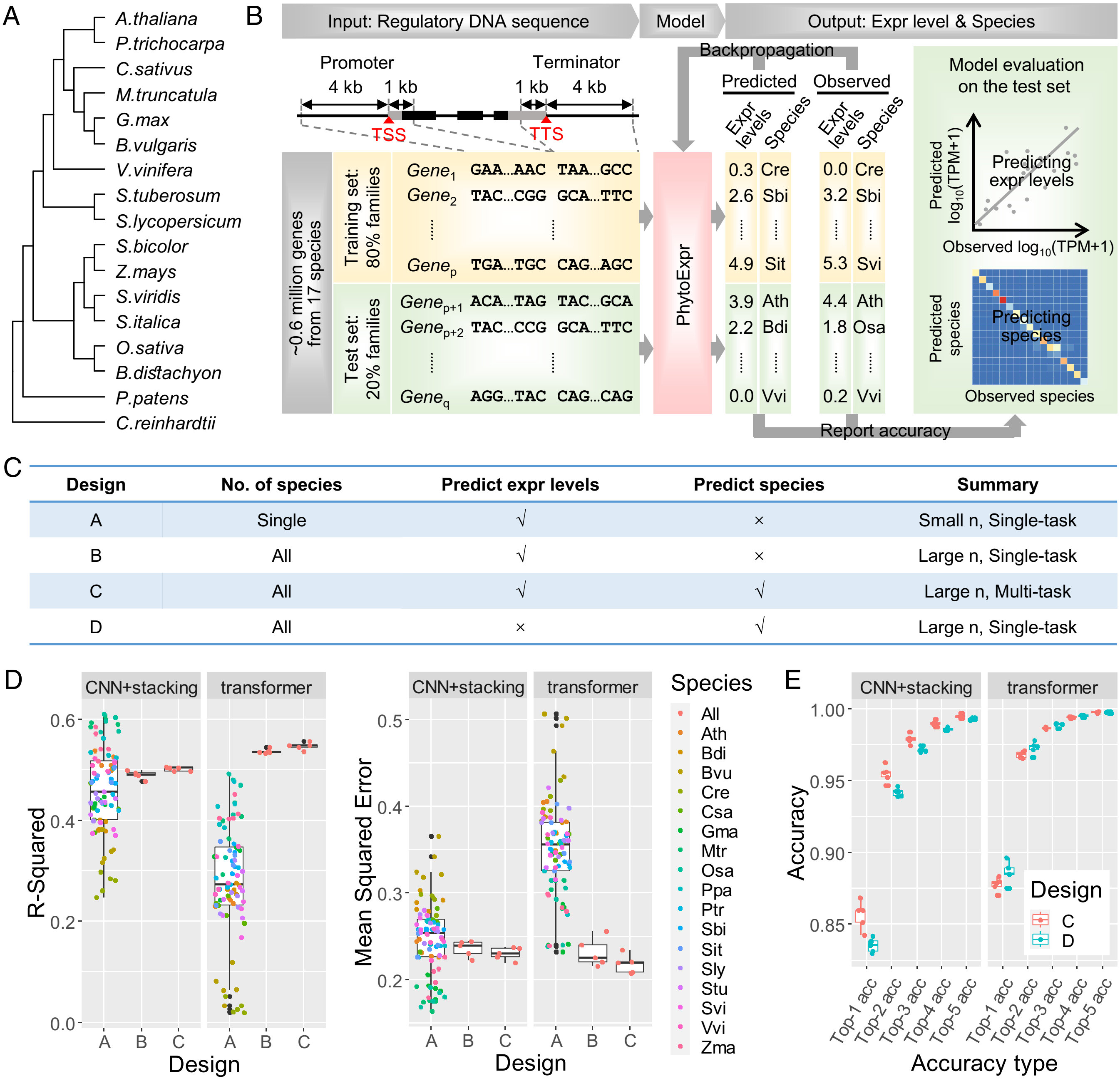
To prepare the RNA sequencing (RNA-seq) datasets for PhytoExpr, raw reads of 6,256 samples originating from 17 plant species (Dataset S2) were acquired from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database and analyzed using a unified pipeline. To simplify the complexities of spatiotemporal gene expression patterns, a virtual representation called “median sample” was created for each species by extracting the median expression level for each gene across all RNA-seq samples within that species. These “median samples” were positioned at the center of Principal Component Analysis (PCA) plots (*SI Appendix*, Fig. S1*A*) and could thus be considered as a composite of various tissue types. Notably, these “median samples” exhibited strong correlations with the actual RNA-seq samples utilized during their construction, with most Pearson correlation coefficients surpassing 0.8 (*SI Appendix*, Fig. S1*B*). To evaluate the robustness of these “median samples,” extensive simulations were conducted by incorporating different numbers of RNA-seq samples to establish them. The results showed that “median samples” stabilized after incorporating as few as 50 RNA-seq samples (*SI Appendix*, Fig. S1*C*).

The significance of proximal regulatory regions in plant *cis*-engineering and transcriptional regulation surpasses that of distal regions, as supported by three findings: 1) Across all the 17 plant species studied, genes positioned in close vicinity (<8 kb) are more likely to be coexpressed compared to those located farther apart, although this pattern is weaker in maize than in other plant species (*SI Appendix*, Fig. S2*A*). 2) An analysis of ChIP-Seq data from a prior study on maize cistrome (2) indicated that about 70% of TFBSs were positioned within 4 kb of transcription start sites (TSS) or transcription termination sites (TTS) (*SI Appendix*, Fig. S2*B*), despite the sparser distribution of *cis*-elements in maize relative to species with smaller and more compact genomes. 3) A former eQTL analysis (15) showed that about 50% of lead SNPs in *cis*-eQTL peaks were within 4 kb of the nearest gene (*SI Appendix*, Fig. S2*C*). Based on these observations, PhytoExpr considered 5 kb promoter (ranging from 4 kb upstream of TSS to 1 kb downstream of TSS) and 5 kb terminator (ranging from 1 kb upstream of TTS to 4 kb downstream of TTS) of each gene as the input (Fig. 1*B*), covering 14.57 to 85.29% of total genome sizes, depending on the species analyzed (Dataset S1).

PhytoExpr Achieved the Highest Accuracy When Employing Multispecies and Multitask Learning with a Transformer Architecture. To ensure an unbiased evaluation of the predictive accuracy of PhytoExpr, it was trained and tested using a fivefold cross-validation scheme in a gene family-aware fashion (36): The model was tested on gene families not used in training (Fig. 1*B*). Two model architectures were utilized (*SI Appendix*, Fig. S3). The first architecture is based on ensemble learning: A linear model was stacked on top of 27 parallel but architecturally distinct convolution neural networks (CNNs) to integrate their predictions (referred to as CNN+stacking hereafter). The second architecture was a transformer with multihead attention (referred to as transformer hereafter). Hyperparameters of both architectures were fine-tuned by random forests (Dataset S3).

For each architecture, four designs of the PhytoExpr framework were evaluated (Fig. 1*C*). We first tried training and testing PhytoExpr on individual species (design A) or all of the 17 species (design B). Notably PhytoExpr trained using design B showcased significant greater accuracy than that of design A, evident from significantly higher R-Squared (R^2) and lower mean squared error (MSE) (Fig. 1*D*). We further speculated that using multiple species, while beneficial for PhytoExpr to learn general *cis*-regulatory logic in the plant kingdom, may also incur a penalty in model performance, as the data are confounded by species-specific regulatory logic. Thus, we modified design B to predict both mRNA abundance and plant species (termed design C), resulting in a considerable improvement of PhytoExpr’s predictive accuracy of mRNA abundance (Fig. 1*D*). The transformer architecture outperformed CNN+stacking for design B and C, but not for A, likely due to the small size of each CNN in the CNN+stacking architecture being more suitable for handling small datasets (Fig. 1*D*).

We further evaluated whether the accuracy of the transformer architecture on design C was associated with the dispersion of gene expression levels across RNA-seq samples. As shown in *SI Appendix*, Fig. S4, although MSE is positively correlated with median absolute deviation (MAD) of gene expression, such positive association is absent between MSE and MAD divided by median expression level (MADM), indicating that the accuracy



Design	No. of species	Predict expr levels	Predict species	Summary
A	Single	√	×	Small n, Single-task
B	All	√	×	Large n, Single-task
C	All	√	√	Large n, Multi-task
D	All	×	√	Large n, Single-task

Fig. 1. Deep learning-based prediction of mRNA abundance and plant species with proximal regulatory sequence as the sole input. (*A*) 17 phylogenetically diverse plant species representing major branches in the plant kingdom were chosen. (*B*) A schematic representation of the architecture of the PhytoExpr framework, which takes proximal regulatory sequences as inputs to predict mRNA abundance as well as the plant species the DNA sequence is from. (*C*) A summary of the four designs of PhytoExpr. PhytoExpr was trained on individual species (Design A), or on all the 17 species (Design B, C, and D). The output of PhytoExpr can be mRNA abundance only (Design A and B), species only (Design D), or both (Design C, also illustrated in *B*). (*D*) The accuracy of PhytoExpr in predicting mRNA abundance, represented as R-squared (the *Left* panel) and MSE (the *Right* panel), respectively. (*E*) The accuracy of PhytoExpr in predicting species information.

of the transformer model is unaffected by the relative dispersion of gene expression levels. We also reported the mean absolute error and median absolute error, two metrics that are comparable to the target variable (i.e., log10-transformed median expression level) (*SI Appendix*, Fig. S5).

In a previous study in yeast, codon frequencies and mRNA features (lengths of 5′UTR, ORF, and 3′UTRs, GC content of 5′UTR and 3′UTRs, and GC content at each codon position in the ORF), when used together with *cis*-regulatory sequences as predictors of gene expression levels, significantly enhanced the model accuracy, compared with using *cis*-regulatory sequences as the sole predictor (37). In plants, we found that although codon frequencies and mRNA features each explained a small proportion of the variation in gene expression, they did not enhance model accuracy when used together with regulatory sequences (*SI Appendix*, Fig. S6).

In design C, the model assigned a probability to each species for every input DNA sequence, indicating the likelihood of the input sequence originating from that species. For both architectures, the averaged top-1 accuracy (i.e., the true species received the highest probability) was above 85%, and the averaged top-2 accuracy (i.e., the true species was among the top-2 guesses made by the model) was above 95% (Fig. 1*E*). These outcomes indicate that PhytoExpr has effectively learned species-specific signals in input sequences. Once again, the transformer architecture exhibited superior performance in comparison to the CNN+stacking architecture. Interestingly, design C displayed significantly higher accuracy than design D, where PhytoExpr was exclusively trained to predict species information, highlighting the effectiveness of multitask learning. In design C, most top-1 errors were attributed to phylogenetic relatedness. For example, *Setaria italica* sequences that were incorrectly predicted to be originating from its wild ancestor, *Setaria viridis*, comprised 30% of all top-1 errors (*SI Appendix*, Fig. S7).

PhytoExpr Is Generalizable to Unseen Gene Families from Unseen Plant Species. Given the generalizability of PhytoExpr to unseen gene families (gene family-guided data splitting in Fig. 1*B*), we further asked whether PhytoExpr can be generalizable to unseen gene families from unseen species not used in model training. We retrained the design C models (i.e., predicting both mRNA abundance and plant species) using 16 of the 17 species and then tested the models on held-out gene families from the held-out species. In most cases, the predictive accuracies for mRNA abundance were similar to those achieved while using all the 17 species (Fig. 2). The only exception was *Chlamydomonas reinhardtii*: Models that were not trained on *C. reinhardtii* had no predictive accuracy for this species, probably due to its very early divergence from the other 16 species and unique *cis*-regulatory logic. Across all scenarios, the transformer architecture consistently outperformed the CNN+stacking architecture. Design C models also provided species information as part of output. Notably, when making prediction on sequences from a new species, the design C models predominantly suggested closely related species (*SI Appendix*, Fig. S8), indicating that PhytoExpr implicitly learned the phylogenetic relationship among the species incorporated during model training. In summary, PhytoExpr generalizes well to unseen species that share evolutionary relations with the species included in the initial model training.

PhytoExpr-Derived Nucleotide Importance Scores Associated with Conserved Noncoding Sequences, TFBSs, and *cis*-eQTLs. As the design C model with a transformer architecture was most accurate, all the downstream analysis was conducted based on this model (referred to as PhytoExpr hereafter). To evaluate the importance of each nucleotide, we employed PhytoExpr (with the transformer architecture and design C) to predict mRNA abundance from the original sequence and a sequence in which that nucleotide was occluded (i.e., represented as four zeros in on-hot encoding). The importance score of the nucleotide was calculated as the difference between the two predicted expression levels, and nucleotides with strong positive or negative effects on gene expression were considered as high-impact nucleotides (Fig. 3*A*). We computed the importance score of each nucleotide in the *cis*-regulatory regions of each maize gene, and noted an enrichment of high-impact nucleotides near TSS or TTS (Fig. 3*B*). We hypothesized a potential association between the importance score and functional annotation data for regulatory regions although PhytoExpr had never been trained on such datasets. We first compared the importance scores with pan-Andropogoneae conserved noncoding sequences (pan-And-CNS). The maize pan-And-CNS was curated by aligning the noncoding sequences of

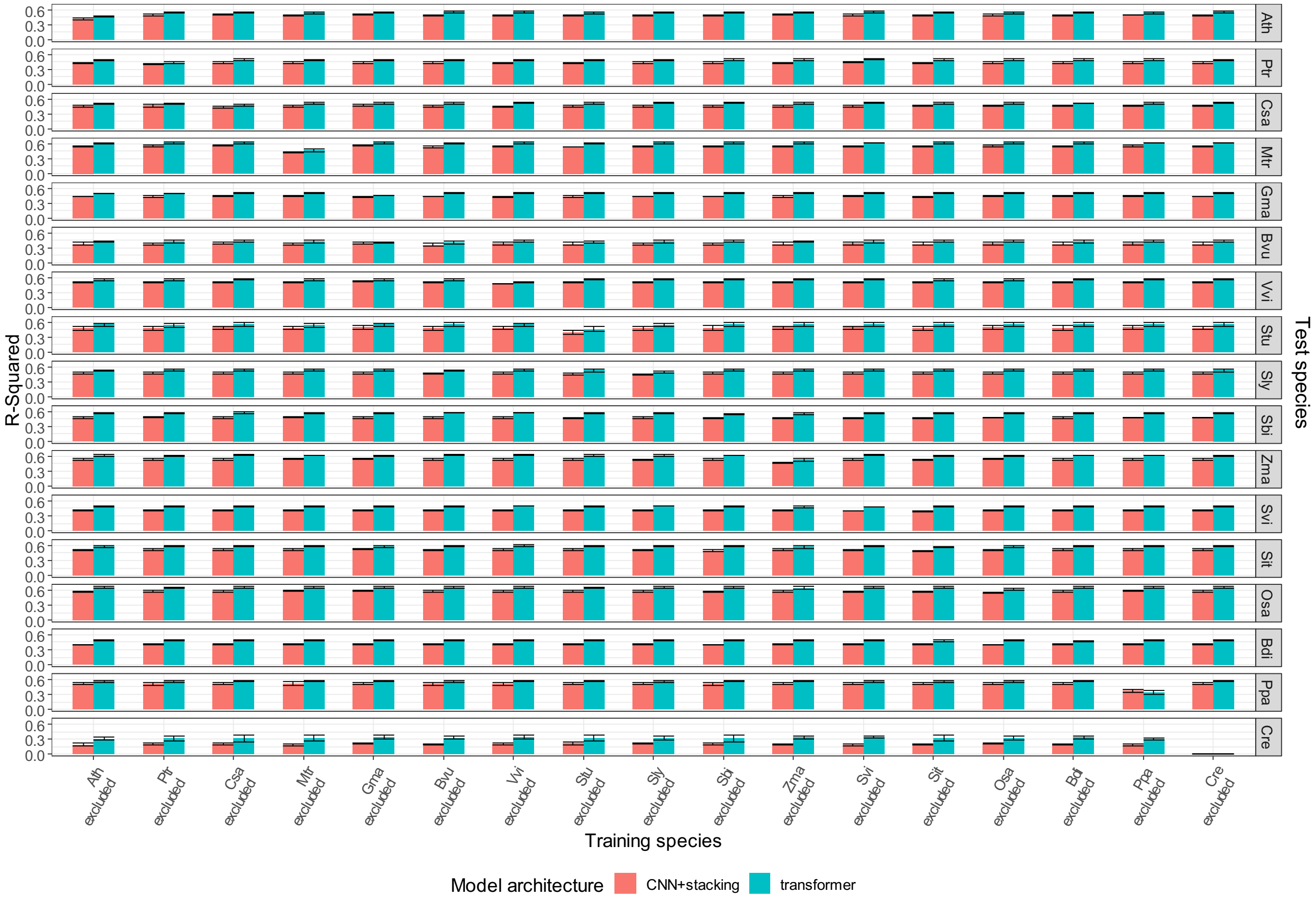


Fig. 2. Generalizability of PhytoExpr to unseen gene families in unseen plant species. Design C models (i.e., predicting both mRNA abundance and plant species) were trained on the training gene families in 16 of the 17 species, and tested on the test gene families from the held-out species. Error bars represent mean ± SD from family-aware fivefold cross-validation.

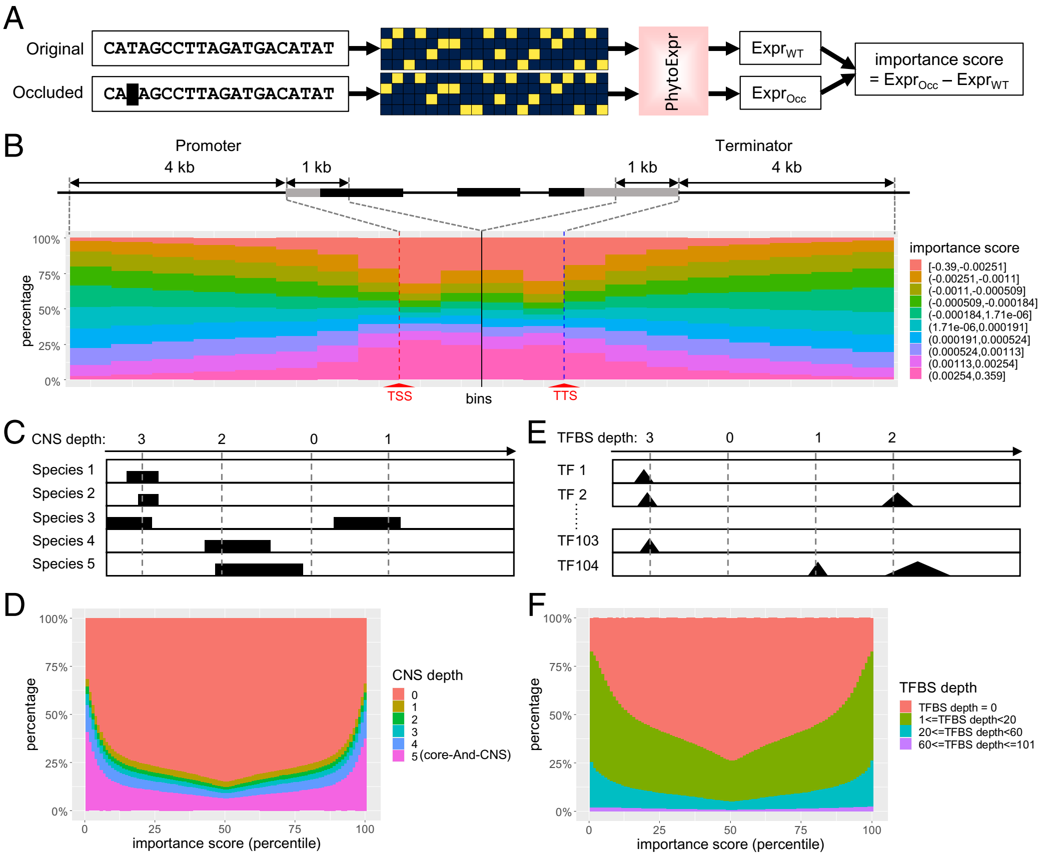


Fig. 3. Quantitative functional annotation of proximal regulatory regions at single-nucleotide resolution. (*A*) To quantify the importance of a nucleotide in the regulatory sequence, PhytoExpr with the transformer architecture was used to predict mRNA abundance from DNA sequence, as well as from the same sequence in which this nucleotide is occluded. The importance score of the nucleotide was defined as the difference between the two predicted expression values. (*B*) Distribution of importance scores along regulatory sequences of all maize genes. (*C*) The CNS depth of a nucleotide is defined as the number of species in which this nucleotide is conserved. (*D*) The relationship between importance score and CNS depth in maize. Maize pan-Andropogoneae CNS was from a previous study (16) by comparing the noncoding regions of maize and five other Andropogoneae species representing ~18 My of divergence. (*E*) The TFBS depth of a nucleotide is defined as the number of transcription factors with a binding site overlapping this nucleotide. (*F*) The relationship between importance scores and TFBS depths in maize. ChIP-Seq datasets were from a previous maize cistrome study on 104 TFs (2).

maize with those of five other Andropogoneae species (*Sorghum bicolor*, *Miscanthus sinensis*, *Saccharum spontaneum*, *Hyparrhenia diplandra*, and *Chrysopogon serrulatus*) spanning approximately 18 My of divergence (16). The conservation level of a nucleotide in a CNS was defined as the number of species in which this nucleotide was conserved (referred to as "CNS depth") (Fig. 3*C*). Notably, nucleotides that were conserved in all the five species (i.e., CNS depth = 5, also referred to as core-And-CNS) exhibited a significant enrichment of high-impact nucleotides (Fig. 3*D*). Additionally, we compared the importance scores with a previously published ChIP-Seq dataset comprising genome-wide TFBSs of 104 TFs (2). We defined "TFBS depth" of a nucleotide as the number of TFBSs covering this nucleotide (Fig. 3*E*). We observed a significant enrichment of high-impact nucleotides in TFBSs (Fig. 3*F*). However, such enrichment was not evident for nucleotides with very high TFBS depth (i.e., TFBS depth ≥ 60) (Fig. 3*F*), likely due to these "hyper-ChIPable regions" being artifacts of ChIP-Seq experiments (38). We further evaluated whether PhytoExpr effectively scores the importance of nucleotides in transposable element (TE) regions. A comprehensive annotation of maize TEs was downloaded from https://github.com/mcstitzer/maize_TEs. Although TEs were depleted in CNS and TFBS regions (*SI Appendix*, Fig. S9 *A* and *C*), a significant enrichment of high-impact nucleotides in CNS and TFBS was observed in both non-TE and TE regions (*SI Appendix*, Fig. S9 *B* and *D*). These results suggested that PhytoExpr identified important nucleotides in both non-TE and TE regions.

The maize HapMap3 database encompasses 83 million variant sites (67% of them with MAFs < 5%) identified by whole genome resequencing of 1,218 diverse maize varieties worldwide (39). Among these SNPs, 17.3 million SNPs (63% of them with MAFs < 5%) were within the regulatory regions investigated in this study. PhytoExpr was deployed to evaluate the impact of each of the 17.3 million SNPs on gene expression (referred to as the SNP effect score) (*SI Appendix*, Fig. S10*A*). SNPs with large positive or negative effects on gene expression were enriched with rare variants (MAF < 0.01), implying that large-effect variants may undergo purifying selection (*SI Appendix*, Fig. S10*B*). We also observed a strong correlation between SNP effect score and the distance to TSS or TTS (*SI Appendix*, Fig. S10*C*). We further postulated that large-effect SNPs predicted by PhytoExpr might exhibit stronger associations with the variation in gene expression levels in natural populations, despite PhytoExpr had not been prior trained on population genetics datasets. In a previous comprehensive analysis of eQTL based on the transcriptome profiles from 7 tissues across 282 diverse maize inbreds, common *cis*-regulatory variants (MAF > 0.05) significantly associated with variations in gene expression levels were identified (15). We defined the rank of a SNP based on its position in a *cis*-eQTL peak on the Manhattan plot: The peak SNP was assigned with a rank of 0, while unassociated SNPs were marked as "not significant" (*SI Appendix*, Fig. S10*D*). Indeed, we found that predicted large-effect SNPs (with either positive or negative impact on gene expression) were notably enriched with SNPs holding high ranks in *cis*-eQTL peaks (*SI Appendix*, Fig. S10*E*), affirming the effectiveness of PhytoExpr in prioritizing causal *cis*-regulatory variants.

PhytoExpr Facilitates Rational Design of *cis*-Acting Variants and De Novo Creation of Species-Specific Regulatory Sequences. We further designed two strategies for the rational design of *cis*-acting variants, and de novo creation of species-specific regulatory sequences. Fine-tuning of gene expression by *cis*-engineering has been proven an effective strategy for crop improvement. While a single nucleotide mutation is feasible using base editors, often such manipulations do not really translate into significant alterations of mRNA abundance. As a result, our aim was to design for each target gene a single small deletion that can be easily created by the CRISPR/Cas9 system with two gRNA cassettes, or a small replacement that can be achieved by prime editing.

To accomplish this, we introduced an algorithm termed importance score-guided semi-brute force that made semi-random small deletions on the input sequence. The impacts of these deletions on mRNA abundance were then evaluated using PhytoExpr. In order to minimize the search space, in silico deletions were restricted to regions with high average importance scores (Fig. 4*A*). We applied this strategy to design up-regulated promoters for two maize genes, as well as down-regulated promoters for two other maize genes (Fig. 4*B*). The transcriptional activity of the designed sequences was confirmed by the dual-luciferase transcriptional activity assay conducted in maize protoplasts (Fig. 4*C*).

The other strategy involved using PhytoExpr as the selection operator within the genetic algorithm to evolve species-specific regulatory sequences in a desired direction (i.e., either higher or lower expression level) (Fig. 4*D* and *SI Appendix*, Fig. S11). We first tried this idea in *Arabidopsis* as an example. We started with 1,000 entirely random 10 kb DNA sequences (denoted as Generation 0) and aimed to evolve a 2 kb target region (2 kb to 4 kb). In each generation, sequences that demonstrated both a stronger predicted transcriptional strength and higher similarity to *Arabidopsis* had more chance to propagate themselves (Fig. 4*E*). We specified a target species during in silico evolution to facilitate the emergence of *cis*-regulatory elements specific to this target species. After 500 generations of evolution, a 2 kb sequence was randomly chosen from Generation 0 and Generation 500, respectively, chemically synthesized, fused to a minimal CaMV 35S promoter followed by eGFP, and transiently expressed in *Arabidopsis* seedlings. The procedure was replicated three times. The three sequences from Generation 500 conferred strong eGFP fluorescence signal comparable to that from the full-length CaMV 35S promoter, whereas sequences from Generation 0 did not produce detectable eGFP fluorescence (Fig. 4*F* and Dataset S4). We also tested this strategy in maize. In total 20 *cis*-regulatory sequences (270 bp in length) were evolved, synthesized, and inserted between an eGFP coding sequence and a NOS terminator (*SI Appendix*, Fig. S12*A*). This allowed us to quantify the strength of each evolved sequence by conducting a small-scale self-transcribing active regulatory region sequencing (STARR-Seq) in maize protoplasts. The 35S enhancer and 13 known strong endogenous maize enhancers from a previous study (40) were used as positive controls in the STARR-Seq assay. The STARR-Seq was conducted three times, with high reproducibility among biological replicates (*SI Appendix*, Fig. S12 *B–D*). All the 20 evolved sequences exhibited detectable activity, with some of them as strong as natural enhancers (*SI Appendix*, Fig. S12*E*). As these evolved sequences are short (270 bp) and easy to handle in vector construction, they are potentially useful in plant synthetic biology.

We further employed this strategy to enhance the transcriptional strength of three maize endogenous promoters. For each gene, we focused on evolving a 50 bp region (150 to 200 bp upstream of TSS) to ensure compatibility with current prime editing tools (Fig. 4 *G* and *H*). The targeted DNA region passed through 60 generations of evolution, and dual-luciferase transcriptional activity assays performed in maize protoplasts confirmed that the evolved promoters were indeed significantly stronger than the original wild-type promoters (Fig. 4*I*).

Discussion

Based on the performance of PhytoExpr on gene families and species not included in its training data, it is highly confident that PhytoExpr has learned general syntax rules governing gene expression. Gradient- and perturbation-based methods are commonly used to interpret the behavior of deep learning models. When applied to regulatory DNA sequences, these methods can assign an importance score to each nucleotide. However, it is widely accepted that high-order interactions among nucleotides are crucial in gene expression control, including factors such as the spacing, distance, and orientation of regulatory elements, as well as their additive, cooperative, competitive, and synergistic interactions. Extracting such complex syntax from PhytoExpr poses challenges owing to the vast number of combinations of perturbations. More efficient computational tools are needed to interpret PhytoExpr. Nevertheless, a previous study in yeast revealed that most regulatory interactions were weak (41), suggesting that single-nucleotide occlusion might suffice to unravel most, if not all, information learned by PhytoExpr.

PhytoExpr takes only proximal *cis*-regulatory sequences as its input without considering distal *cis*-elements or trans-acting factors, and it explained over 50% of the variation in gene expression, highlighting a more important role of proximal regulatory regions than distal ones. It is conceivable that incorporating distal regulatory sequences and trans-acting factors into the model would significantly expand the number of features, potentially leading to model overfitting, especially if the number of observations (i.e., the number of genes) remains constant. To mitigate this, using data from more plant species will be the key in the future for PhytoExpr to explore larger variation space to avoid overfitting. *Cis*-engineering on proximal regulatory regions seems adequate to reach any desired expression level, as exemplified by several recent promoter editing studies (21, 23, 25, 26) and our in silico evolution analysis. This stands in contrast to the scenario in human genomics, where variations in distal regulatory elements are playing important roles in human disease genetics and should not be ignored.

In addition to refining model inputs, several other aspects can be improved in future research. For instance, transcriptome datasets utilized in this study relied on short-read Illumina platforms, which might be biased by PCR amplification during library preparation and multimapped reads. Gene expression levels could also be affected by imprecise gene model annotations, especially in untranslated regions (UTRs). Introducing long-read RNA sequencing could alleviate some of these challenges. Two model architectures were extensively optimized for predicting mRNA abundance: an ensemble learning architecture and a transformer architecture. While the transformer architecture exhibited superior performance, it is plausible that more suitable architectures may exist. Another potential avenue involves predicting spatiotemporal expression patterns. Training of PhytoExpr with median expression levels may limit its ability to learn *cis*-elements conferring tissue-specificity or environment responsiveness. As cell- or tissue-specific gene expression and environmental responsiveness of genes are important contributors of agronomic traits of crops, updating PhytoExpr with

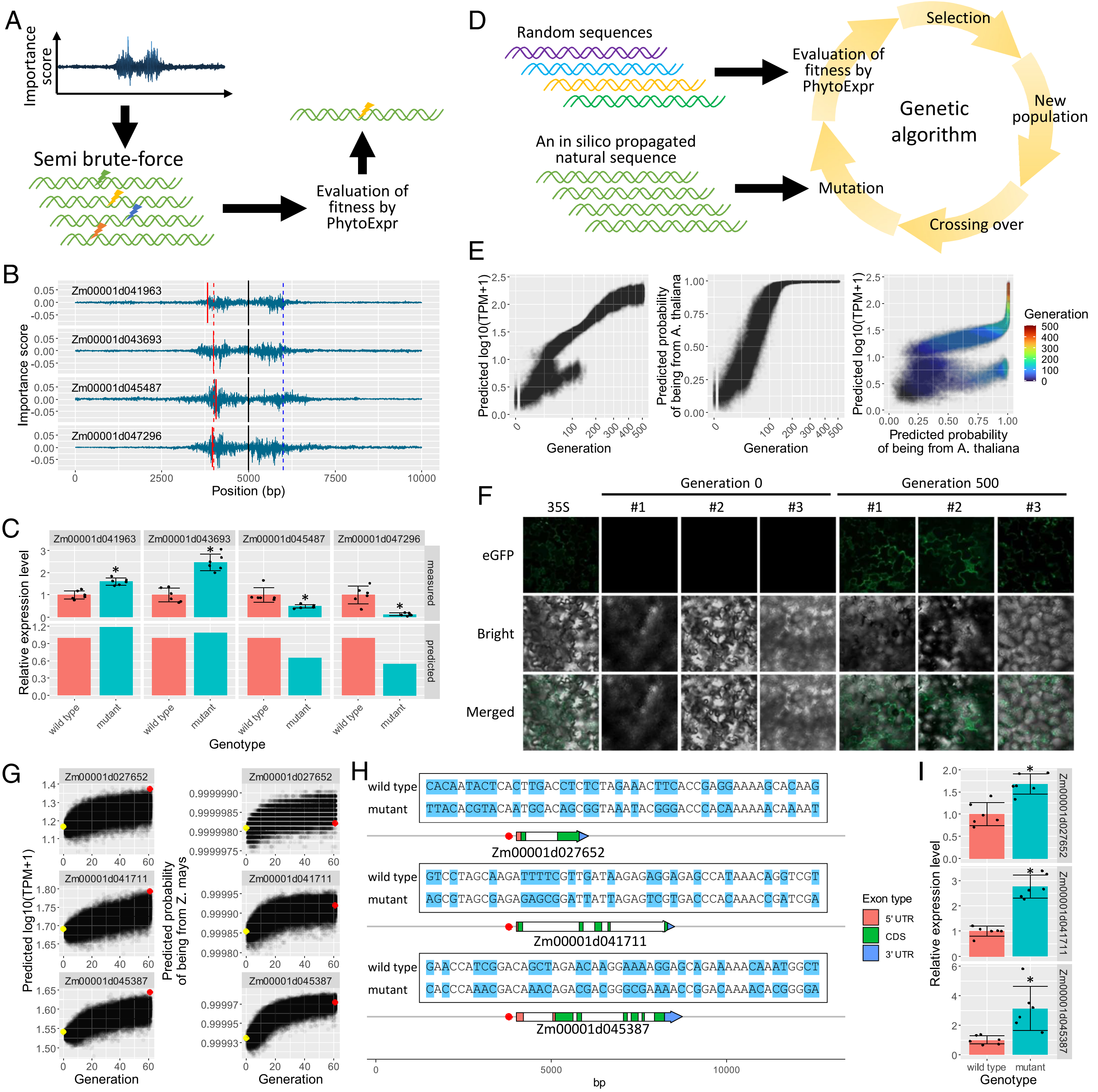


Fig. 4. In silico evolution of natural *cis*-regulatory sequences and random sequences. (*A*) Importance score-guided semi-brute-force approach. (*B*) Importance scores of the *cis*-regulatory regions of four maize genes at single-nucleotide resolution. The position of the 25 bp deletion on each gene was indicated by a vertical solid red line. Dashed red and blue lines indicate TSS and TTS, respectively. (*C*) Predicted and measured promoter strengths for wild-type and mutated promoters. (*D*) A schematic representation of in silico evolution of sequences by genetic algorithm. (*E*) Genetic algorithm-dependent in silico evolution of random sequences. Random sequences were 10 kb long to fit the input shape of PhytoExpr, and only the 2 kb to 4 kb region was subject to evolution. Evolution was replicated three times and a representative result was shown. (*F*) Sequences from Generation 0 and Generation 500 were fused with the minimal CaMV 35S promoter, to drive the transient expression of eGFP in *Arabidopsis*. The full-length CaMV 35S promoter was used as a positive control. (*G*) Genetic algorithm-dependent in silico evolution of three natural maize promoters. Only the region 150 to 200 bp upstream of TSS was evolved. Yellow dots represent wild-type sequences, while red dots represent the sequences with the highest predicted expression level at the 60th generation. (*H*) The wild-type and mutated sequences within the 50 bp target region (marked by red dots). (*I*) The effect of mutations on gene expression measured by the dual-luciferase transcriptional activity assay in maize protoplasts. *, significant difference from the wild type ($P < 0.05$) by the *t* test.

single-cell transcriptome datasets in the future might be essential. It is also worth noting that the RNA-seq datasets used in this study may be enriched for particular tissue types in some plant species while depleted in others. And this may partly explain why models not trained on *C. reinhardtii* have little predictive power in this species, considering that it is a single-cell organism. Generation of transcriptome datasets from matched tissues for multiple species may improve the accuracy of PhytoExpr in the future.

Methods

RNA-seq Data Collection and Processing. The reference genomes of the 17 plant species used in this study were downloaded from EnsemblPlants (42) or Phytozome (43) (summarized in Dataset S1). Raw reads were downloaded from the NCBI SRA in the sra format, converted to the fastq format by fastq-dump from the SRA Toolkit (version 2.8.2), and then quality-trimmed by Sickle (version 1.33, https://github.com/najoshi/sickle) with default settings. Following quality assessment by FastQC (0.11.5, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), the clean reads were aligned to their corresponding reference genomes with HISAT2 (version 2.1.0) (44). The resulting bam files were sorted by SAMtools (version 1.9) (45). Read counts were normalized to Transcripts Per Kilobase Million (TPM) by StringTie (version 2.0) (46). TPM is used as it is proportional to the relative RNA molar concentration (47). RNA-seq samples analyzed in this study are outlined in Dataset S2.

PCA and Coexpression Analysis. Unnormalized raw read counts of genes were obtained from bam files using the featureCounts function from the Subread package (version 1.6.2) (48). These counts were then subjected to variance stabilizing transformation using the vst function from the DESeq2 package (49). Dimensionality reduction of RNA-seq samples by PCA was computed using the plotPCA function in the DESeq2 package. For each species, pairwise biweight midcorrelations (bicor) for all protein-coding genes were calculated using the bicor function from the WGCNA package (50). Bicor is a median-based measure of similarity and is more robust and less sensitive to outliers, compared to other similarity metrics such as Pearson correlation or mutual information (51).

Gene Family-Guided Splitting of Training and Test Datasets. Protein-coding genes from the 17 plant species were divided into gene families as previously described (36). Briefly, pairwise similarities among proteins were obtained by an all-by-all BLASTP search on our local server. Subsequently, an in-house python script was used to construct a graph, where nodes represented genes and edges connected homologous genes. This graph was further divided into clusters (i.e., gene families) by the Markov clustering algorithm implemented in the markov_clustering package in Python with default parameters except that inflation was set to 1.1. If a gene was not assigned to any gene family, it was considered as an orphan gene belonging to a family with only a single member. For gene family-guided fivefold cross-validation, gene families were randomly partitioned into five subsamples with equal numbers of families. In each iteration, one subsample was retained as the test data, while the remaining four subsamples were used as the training set.

Model Architecture and Hyperparameter Optimization. All models were constructed in Python 3 using Keras 2 with a Tensorflow backend. The two model architectures are summarized in *SI Appendix*, Fig. S3. Hyperparameter optimization are summarized in Dataset S3. As possible combinations of hyperparameters are numerous, it is infeasible to exhaustively evaluate them with limited computational resources. Instead, we used random forest models to optimize hyperparameters. For the CNN+Stacking architecture, 300 hyperparameter combinations (200 were randomly chosen and 100 were chosen based on experience) were evaluated. Then, a random forest model was trained and used to predict the accuracies for all possible combinations, and the top 300 predicted best-performing combinations were evaluated. Among them, the top 15 combinations for mRNA abundance prediction and top 15 combinations for species prediction (27 in total) were chosen to constitute the basal layer of the final ensemble learning model. A similar approach was used to optimize the transformer architecture. Hyperparameter tuning was conducted using an approach similar to the standard nested cross-validation approach. As hyperparameter tuning is too computationally expensive, it was only done in Fold 1~4. The resulting model architectures perform similarly on each of the five folds, suggesting that the model is not biased toward a certain fold. For the CNN+Stacking architecture, an Adam optimizer with the default learning rate (0.001) was used. The 27 basal CNN branches were trained individually before being assembled into the full CNN+Stacking architecture. For the transformer architecture, an Adam optimizer with a learning rate of 0.0001 was used. Both models were trained with early stopping (patience = 2) to prevent overfitting. Models were trained on a work station equipped with four NVIDIA Titan RTX graphics card each with 24 GB video memory. The number of epochs vary, due to random initiation of model weights. It typically took ~90 h (~20 epochs for each basal CNN branch) to train a CNN+Stacking model, and ~5 h (~10 epochs) to train a transformer model.

Maize Protoplast Preparation and Transient Transformation. Maize (B73) seeds were grown in soil at room temperature in the dark for about 12 d. The second leaves were cut into thin strips, and placed in 10 mL of enzyme solution [20 mM MES (2-morpholinoethanesulfonic acid), pH 5.7, 1.5% cellulase R-10, 0.5% macerozyme R-10, 0.4 M mannitol, 20 mM KCl, 10 mM CaCl$_2$, 0.1% BSA, and 5 mM β-mercaptoethanol]. The samples were vacuumed (0.8 MPa) for 30 min and then gently shaken for 3 to 4 h at 40 rpm at 25 °C in dark. For each sample, 10 mL W5 (2 mM MES, 154 mM NaCl, 125 mM CaCl$_2$, and 5 mM KCl) solution was added, and then the sample was filtered with a 200-mesh filter. The filtrate was transferred into a 50 mL centrifuge tube and centrifuged at 100 g for 2 min at 4 °C, and then the pellet of protoplasts was resuspended with 10 mL of W5 solution and centrifuged again at 100 g for 2 min at 4 °C. The pellet of protoplasts was resuspended with 5 mL of W5 buffer and kept on ice for 30 min. After removing the supernatant, the protoplasts were resuspended in the MMG buffer (4 mM MES, 0.4 M mannitol, and 15 mM MgCl$_2$) at 10^5-2×10^5 cells/mL. For transient transformation of protoplasts, 10 μL plasmid (10 μg) and 100 μL protoplasts were gently mixed in a 2 mL centrifuge tube. After adding 110 μL 40% PEG-Ca^{2+} solution (40% PEG, 0.2 M mannitol, and 0.1 M CaCl$_2$), the mixture was kept at 25 °C for 15 min in dark. Then, 440 μL W5 solution was added and centrifuged at 100 g for 2 min. The supernatant was carefully removed and the pellet was resuspended with 500 μL W5 solution. Protoplasts were then centrifuged again at 100 g for 2 min at 25 °C. Finally, cells were resuspended in 1 mL W5 solution and cultured at 25 °C for 15 h in dark.

Dual-Luciferase Transcriptional Activity Assay in Maize Protoplasts. Wild-type and mutated promoter sequences were cloned into the plant expression vector pGreenII0800-LUC. Promoter sequences and primers used are listed in the Dataset S4. Luciferase activities were measured using the Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's instructions.

Transient Gene Expression Analysis in *Arabidopsis* Seedlings. Transient gene expression in *Arabidopsis* was performed as previously described (52). *Arabidopsis* seeds were surface-sterilized with 75% ethanol for 10 min and sown on 1/4 MS agar plate (pH = 6) with 1% sucrose. After stratification at 4 °C for 24 h, the seeds were germinated in an artificial climate incubator (16 h light and 8 h dark at 22 °C). Artificially designed promoter sequences were chemically synthesized at a local vendor and cloned into pCAM-eGFP-N1 plasmid between the *Xma*I and *Nco*I restriction sites. Plasmids were transformed into *Agrobacterium tumefaciens* (GV3101) and stock onto LB agar plates with kanamycin and rifampicin for 2 d at 28 °C. A single colony from the plate was inoculated into 2 mL LB medium with antibiotics (25 μg/mL kanamycin plus 25 μg/mL rifampicin) and cultured at 28 °C for 18 to 24 h, diluted with 10 mL fresh YEB medium (5 g/L beef extract, 1 g/L yeast extract, 5 g/L peptone, 5 g/L sucrose, 0.5 g/L MgCl$_2$) to OD$_{600}$ = 0.3 and grown until OD$_{600}$ = 1.5. *A. tumefaciens* cells were harvested by centrifugation at 6,000 g for 5 min, washed with 10 mL washing solution (10 mM MgCl$_2$, and 100 μM acetosyringone) and resuspended in 1 mL washing solution. Four-day-old *Arabidopsis* seedlings were cocultivated with *A. tumefaciens* for 36 to 40 h in the dark in 20 mL cocultivation medium (1/4 MS, 1% sucrose, pH = 6.0, 100 μM AS, 0.005% Silwet L-77). Fluorescence imaging was performed using an LSM900 confocal microscope.

STARR-Seq Assay in Maize Protoplasts. The STARR-Seq vector and experimental procedure was from a previous study (5). The 20 evolved sequences, 35S enhancer, and the 13 maize strong endogenous enhancers from a previous study (40) (listed in Dataset S4) were synthesized and amplified using the primer pool-F and pool-R, and then ligated with the linearized STARR-Seq vector (digested by MfeI and StuI) using the In-Fusion HD Cloning Kit (Clontech). The ligated product was used to transform homemade competent *E. coli* cells. Positive *E. coli* clones were harvested from the plate and cultured in a 100 mL flask, followed by plasmid isolation. The isolation and transfection of maize protoplasts were performed as described in the previous section. Transformed maize protoplasts were divided equally into two samples. Total RNA was extracted from one sample using the TransZol Up Plus RNA Kit (TransGen Biotech). The recovered mRNA was reverse transcribed using TransScript One-Step gDNA Removal and cDNA Synthesis SuperMix (TransGen Biotech) with the primer RT, and the cDNA was amplified by the primers NGS-F and NGS-R. Total DNA was extracted from the other sample by the CTAB method, and amplified by the primers NGS-F and NGS-R. The two libraries were sequenced on the Illumina HiSeq X Ten platform. The transcriptional strength of each sequence was calculated by counting the number of uniquely mapped reads from the cDNA library normalized by the number of uniquely mapped reads from the DNA library for this sequence.

Data, Materials, and Software Availability. Code was deposited at https://doi.org/10.6084/m9.figshare.24417076.v1 (53). Sequences and mRNA abundance data: https://doi.org/10.6084/m9.figshare.24417085.v1 (54). The CNN+stacking model with weights: https://doi.org/10.6084/m9.figshare.24417079.v1 (55). The transformer model with weights: https://doi.org/10.6084/m9.figshare.24417082.v1 (56). The importance scores of transcriptional regulatory nucleotides in maize: https://doi.org/10.6084/m9.figshare.25467022.v1 (57). Previously published data were used for this work (2, 15, 16). All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. This work was supported by the National Key Research and Development Program of China (2020YFE0202300), National Natural Science Foundation of China (32071464), Chinese Universities Scientific Fund (2023RC002), and the 2115 Talent Development Program of China Agricultural University. We thank Dr. Xinhao Ouyang from Xiamen University and Dr. Yaxi Liu from Sichuan Agricultural University for their assistance in the STARR-Seq assay.

Author affiliations: [a]State Key Laboratory of Maize Bio-breeding, National Maize Improvement Center, Frontiers Science Center for Molecular Design Breeding, Department of Plant Genetics and Breeding, China Agricultural University, Beijing 100193, People's Republic of China; [b]Legumes Research Program, Research and Innovation Division, Tanzania Agricultural Research Institute, Ilonga, Kilosa, Morogoro 67410, Tanzania; [c]Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus 8000, Denmark; [d]National Key Laboratory of Wheat Improvement, Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agriculture Sciences in Weifang, Weifang, Shandong 261325, People's Republic of China; [e]Key Laboratory of Maize Biology and Genetic Breeding in Arid Area of Northwest Region of the Ministry of Agriculture, College of Agronomy, Northwest A&F University, Yangling, Shaanxi 712100, People's Republic of China; [f]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853; [g]Agricultural Research Service, United States Department of Agriculture, Ithaca, NY 14853; [h]Center for Crop Functional Genomics and Molecular Breeding, China Agricultural University, Beijing 100193, People's Republic of China; and [i]Sanya Institute of China Agricultural University, Sanya 572025, People's Republic of China

Author contributions: E.S.B. and H.W. designed research; H.X., S.T., M.S., M.X., and Y.Q. performed research; G.P.R., B.S., E.S.B., and H.W. contributed new reagents/analytic tools; T.L., H.X., S.T., R.B., and H.W. analyzed data; and T.L., R.B., G.P.R., and H.W. wrote the paper.

1. A. P. Marand, A. L. Eveland, K. Kaufmann, N. M. Springer, Cis-regulatory elements in plant development, adaptation, and evolution. *Annu. Rev. Plant Biol.* **74**, 111–137 (2023).
2. X. Tu *et al.*, Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nat. Commun.* **11**, 5089 (2020).
3. S. D. Savadel *et al.*, The native cistrome and sequence motif families of the maize ear. *PLoS Genet.* **17**, e1009689 (2021).
4. J. E. Moore *et al.*; ENCODE Project Consortium, Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
5. W. Tian, X. Huang, X. Ouyang, Genome-wide prediction of activating regulatory elements in rice by combining STARR-seq with FACS. *Plant Biotechnol. J.* **20**, 2284–2297 (2022).
6. J. Sun *et al.*, Global quantitative mapping of enhancers in rice by STARR-seq. *Genomics Proteomics Bioinformatics* **17**, 140–153 (2019).
7. T. Jores *et al.*, Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat. Plants* **7**, 842–855 (2021).
8. T. Jores *et al.*, Identification of plant enhancers and their constituent elements by STARR-seq in tobacco leaves. *Plant Cell* **32**, 2120–2131 (2020).
9. A. Boija *et al.*, Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell* **175**, 1842–1855.e16 (2018).
10. D. M. Suter *et al.*, Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).
11. J. Crocker *et al.*, Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* **160**, 191–203 (2015).
12. J. Hausser, A. Mayo, L. Keren, U. Alon, Central dogma rates and the trade-off between precision and economy in gene expression. *Nat. Commun.* **10**, 68 (2019).
13. Y. Liang, H. J. Liu, J. Yan, F. Tian, Natural variation in crops: Realized understanding, continuing promise. *Annu. Rev. Plant Biol.* **72**, 357–385 (2021).
14. X. Sun *et al.*, The role of transposon inverted repeats in balancing drought tolerance and yield-related traits in maize. *Nat. Biotechnol.* **41**, 120–127 (2023).
15. K. A. G. Kremling *et al.*, Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* **555**, 520–523 (2018).
16. B. Song *et al.*, Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. *Genome Res.* **31**, 1245–1257 (2021).
17. X. Zhou *et al.*, Systematic annotation of conservation states provides insights into regulatory regions in rice. *J. Genet. Genomics* **49**, 1127–1137 (2022).
18. G. M. Cooper *et al.*, Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
19. E. Rodgers-Melnick *et al.*, Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3823–3828 (2015).
20. Y. Wu *et al.*, Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. *Cell* **186**, 2313–2328.e15 (2023).
21. D. Rodriguez-Leal, Z. H. Lemmon, J. Man, M. E. Bartlett, Z. B. Lippman, Engineering quantitative trait variation for crop improvement by genome editing. *Cell* **171**, 470–480.e8 (2017).
22. H. Zhang *et al.*, Genome editing of upstream open reading frames enables translational control in plants. *Nat. Biotechnol.* **36**, 894–898 (2018).
23. L. Liu *et al.*, Enhancing grain-yield-related traits by CRISPR-Cas9 promoter editing of maize CLE genes. *Nat. Plants* **7**, 287–294 (2021).
24. Y. Lu *et al.*, A donor-DNA-free CRISPR/Cas-based approach to gene knock-up in rice. *Nat. Plants* **7**, 1445–1452 (2021).
25. X. Wang *et al.*, Dissecting cis-regulatory control of quantitative trait variation in a plant stem cell circuit. *Nat. Plants* **7**, 419–427 (2021).
26. X. Song *et al.*, Targeting a gene regulatory element enhances rice grain yield by decoupling panicle number and size. *Nat. Biotechnol.* **40**, 1403–1411 (2022).
27. A. H. Yona, E. J. Alm, J. Gore, Random sequences rapidly evolve into de novo promoters. *Nat. Commun.* **9**, 1530 (2018).
28. E. D. Vaishnav *et al.*, The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603**, 455–463 (2022).
29. C. G. de Boer *et al.*, Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* **38**, 56–65 (2020).
30. Y. Wang *et al.*, Synthetic promoter design in Escherichia coli based on a deep generative network. *Nucleic Acids Res.* **48**, 6403–6412 (2020).
31. J. Zrimec *et al.*, Controlling gene expression with deep generative design of regulatory DNA. *Nat. Commun.* **13**, 5099 (2022).
32. J. Zhou *et al.*, Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
33. D. R. Kelley *et al.*, Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
34. D. R. Kelley, Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* **16**, e1008050 (2020).
35. Z. Avsec *et al.*, Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
36. J. D. Washburn *et al.*, Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5542–5549 (2019).
37. J. Zrimec *et al.*, Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* **11**, 6141 (2020).
38. K. Wreczycka *et al.*, HOT or not: Examining the basis of high-occupancy target regions. *Nucleic Acids Res.* **47**, 5735–5745 (2019).
39. R. Bukowski *et al.*, Construction of the third-generation Zea mays haplotype map. *Gigascience* **7**, 1–12 (2018).
40. W. A. Ricci *et al.*, Widespread long-range cis-regulatory elements in the maize genome. *Nat. Plants* **5**, 1237–1249 (2019).
41. C. G. de Boer *et al.*, Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* **38**, 56–65 (2020).
42. A. D. Yates *et al.*, Ensembl genomes 2022: An expanding genome resource for non-vertebrates. *Nucleic Acids Res.* **50**, D996–D1003 (2021).
43. D. M. Goodstein *et al.*, Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2011).
44. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
45. P. Danecek *et al.*, Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
46. M. Pertea *et al.*, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
47. J. H. Bullard, E. Purdom, K. D. Hansen, S. Dudoit, Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
48. Y. Liao, G. K. Smyth, W. Shi, featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
49. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
50. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
51. P. Langfelder, S. Horvath, Fast R functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* **46**, i11 (2012).
52. J.-F. Li, E. Park, A. G. von Arnim, A. Nebenführ, The FAST technique: A simplified Agrobacterium-based transformation method for transient gene expression analysis in seedlings of Arabidopsis and other plant species. *Plant Methods* **5**, 6 (2009).
53. T. Li *et al.*, The code developed in this study. Figshare. https://doi.org/10.6084/m9.figshare.24417076.v1. Deposited 21 October 2023.
54. T. Li *et al.*, Sequences and mRNA abundance data. Figshare. https://doi.org/10.6084/m9.figshare.24417085.v1. Deposited 21 October 2023.
55. T. Li *et al.*, The CNN+stacking model with weights. Figshare. https://doi.org/10.6084/m9.figshare.24417079.v1. Deposited 21 October 2023.
56. T. Li *et al.*, The transformer model with weights. Figshare. https://doi.org/10.6084/m9.figshare.24417082.v1. Deposited 21 October 2023.
57. T. Li *et al.*, The importance scores of transcriptional regulatory nucleotides in maize. Figshare. https://doi.org/10.6084/m9.figshare.25467022.v1. Deposited 24 March 2024.