Cross-species modeling of plant genomes at single nucleotide resolution using a pre-trained DNA language model

Jingjing Zhai[1*+], Aaron Gokaslan[2*], Yair Schiff[2], Ana Berthel[1], Zong-Yan Liu[3], Zachary R. Miller[1], Armin Scheben[5], Michelle C. Stitzer[1], M. Cinta Romay[1,3], Edward S. Buckler [1,3,4+], Volodymyr Kuleshov[2+]

1 Institute for Genomic Diversity, Cornell University, Ithaca, NY USA 14853

2 Department of Computer Science, Cornell University, Ithaca, NY, USA 14853

3 Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY USA 14853

4 USDA-ARS; Ithaca, NY, USA 14853

5 Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY USA 11724

* These authors contributed equally to this work

+ To whom correspondence may be addressed. Email: vk379@cornell.edu, jz963@cornell.edu, or ed.buckler@usda.gov

Abstract

Understanding the function and fitness effects of diverse plant genomes requires transferable models. Language models (LMs) pre-trained on large-scale biological sequences can learn evolutionary conservation, thus expected to offer better cross-species prediction through fine-tuning on limited labeled data compared to supervised deep learning models. We introduce PlantCaduceus, a plant DNA LM based on the Caduceus and Mamba architectures, pre-trained on a carefully curated dataset consisting of 16 diverse Angiosperm genomes. Fine-tuning PlantCaduceus on limited labeled Arabidopsis data for four tasks involving transcription and translation modeling demonstrated high transferability to maize that diverged 160 million years ago, outperforming the best baseline model by 1.45-fold to 7.23-fold. PlantCaduceus also enables genome-wide deleterious mutation identification without multiple sequence alignment (MSA). PlantCaduceus demonstrated a threefold enrichment of rare alleles in prioritized deleterious mutations compared to MSA-based methods and matched state-of-the-art protein LMs.

PlantCaduceus is a versatile pre-trained DNA LM expected to accelerate plant genomics and crop breeding applications.

Introduction

Plant genomes will increasingly be sequenced in the coming decades [1]. Understanding these genomes in terms of both transcription and translation is crucial for advancing plant genomics and crop breeding, including mapping causal genes for agronomic traits, improving crop fitness, and optimizing yield. Unlike biomedical applications that often focus on a few key species, plant genomics must consider hundreds of crop species, underscoring the importance of developing cross-species models that can learn general patterns.

Supervised deep learning (DL) sequence models have acquired great success in understanding DNA sequence function such as transcription initiation [2], alternative splicing [3], gene expression [4] and functional mutations. However, supervised DL models typically require large-scale labeled data, such as ENCODE-scale datasets [5,6], to achieve robust performance. Such extensive labeled data is often scarce in plant genomics. Moreover, training supervised models on model species, such as Arabidopsis, presents challenges when transferring to other plant species. However, the success of self-supervised language models (LMs) offers a promising alternative. In this paradigm, a base model is pre-trained on vast amounts of unlabeled data to learn general patterns. Pre-trained models are then fine-tuned on limited labeled data, enabling better performance on downstream tasks and enhancing generalizability across species relative to existing methods. For example, protein LMs, pre-trained on diverse protein sequences spanning the evolutionary tree, have shown successful applications in predicting atomic-level protein structure [7] and disease-causing variants [8] as well as in engineering protein design [9]. These models provide valuable tools for understanding protein function and facilitating innovative solutions in biotechnology and medicine [10].

Unlike protein LMs that are limited to coding regions, DNA LMs enable a comprehensive understanding of the entire genome, offering deeper insights into gene regulation and evolution. Protein LMs have shown success in identifying pathogenic missense mutations in human genetics [8,11], but increasing evidence shows that mutations in noncoding regions, including both intergenic and intronic regions, contribute significantly to both agronomic traits [12] and human diseases [13,14].

Additionally, training multi-species DNA LMs can capture evolutionary conservation at the DNA level, enhancing our understanding of genetic variation across different species.

However, DNA LMs face significant challenges compared to protein LMs. Firstly, eukaryotes, especially plants [15], contain varied percentages of repetitive sequences, complicating the pre-training task. Given that LMs are pre-trained to either predict the next token or tokens masked arbitrarily in a sequence, repetitive sequences that are easier to recover but do not necessarily improve downstream applications can reduce overall model quality [16]. Additionally, noncoding regions are less conserved than coding regions, leading to potential biases if entire genomes are included in pre-training. Lastly, modeling double-stranded DNA requires consideration of reverse complementary base pairing [17] and a bi-directional model that accounts for both upstream and downstream sequences.

To tackle these challenges, we introduce PlantCaduceus, a DNA language model pre-trained on a carefully curated dataset consisting of 16 angiosperm genomes. PlantCaduceus employs single-nucleotide tokenization, enabling precise modeling at base-pair-resolution across diverse plant genomes. By down-sampling noncoding regions and down-weighting repetitive sequences, we generated an unbiased genomic dataset for pre-training. In contrast, other publicly available DNA LMs, such as AgroNT [18] and Nucleotide Transformer [19], use the entire genomes for pre-training, potentially introducing biases toward certain genomes and repetitive sequences. Unlike the unidirectional HyenaDNA [20] or Evo [21], or the convolutional neural network-based GPN [16], PlantCaduceus offers bi-directional context, providing a more comprehensive understanding of DNA interactions. Furthermore, to handle double-stranded DNA, we used the Caduceus architecture [22], which builds on the Mamba [23] architecture and supports reverse complement equivariance. By evaluating the pre-trained PlantCaduceus model on five cross-species tasks, including transcriptional junctions, translational junctions, and evolutionary conservation prediction, we found that our model demonstrated the best performance compared to baseline models for all five tasks. Notably, downstream classifiers fine-tuned on PlantCaduceus with limited labeled data in Arabidopsis maintained the best performance on other crop species such as maize, improving the PRAUC from 1.45-fold to 7.23-fold compared to the best baseline model, indicating that PlantCaduceus effectively captures broad evolutionary conservation. Additionally,

the deleterious mutations identified with the zero-shot strategy of PlantCaduceus show a three-fold enrichment of rare alleles compared to the most commonly used evolutionary-based methods such as phyloP [24] and phastCons [24]. For missense mutations, PlantCaduceus matches the performance of the state-of-the-art protein LMs, suggesting that PlantCaduceus could be effectively used for genome-wide deleterious mutation identification. These results suggest that PlantCaduceus could serve as a foundational model to accelerate plant genomics and crop breeding.

Results

PlantCaduceus: a pre-trained DNA language model with 16 Angiosperm genomes

Caduceus [22] is a DNA LM architecture that builds upon the recently introduced Mamba [23] architecture, a selective state space sequence model that has demonstrated competitive performance to transformers in various NLP tasks, with more efficient scaling for longer range sequences. Unlike Mamba, Caduceus is specifically designed for DNA sequences, taking into account the bi-directional nature of DNA and introducing reverse complement (RC) equivariance. Here, we trained PlantCaduceus using the Caduceus architecture (**Figure 1A**) on 16 Angiosperm genomes (**Supplemental Table S1**), spanning 160 million years of evolutionary history (**METHODS**). PlantCaduceus takes 512 base pair (bp) windows of input sequences, tokenizing them into single nucleotides, and is pre-trained using a masked language modeling objective (**Figure 1A and METHODS**). To address the substantial variation in genome sizes and the high proportion of repetitive sequences in these genomes, we emphasized non-repetitive sequences by down-weighting repetitive sequences during pre-training (**METHODS**). To scale Caduceus, we trained a series of PlantCaduceus models with parameter sizes ranging from 20 million to 225 million (**Table 1**). The training and validation losses for each model are detailed in **Supplemental Table S2**. After pre-training, we conducted a preliminary assessment to verify the model's learning capabilities. Taking the sorghum genome as an example, we employed Uniform Manifold Approximation and Projection (UMAP) [25] to visualize the embeddings generated by the four pre-trained PlantCaduceus models. By segmenting the genome into 512 bp windows, we observed distinct clustering in the UMAP visualization, corresponding to different genomic regions (**Figure 1B**). Due to the high proportion of repetitive intergenic sequences in the sorghum genome, the

embedding spaces appeared dispersed in the UMAP visualization (**Figure 1C**; **Supplemental Figure 1**). Even without any supervision, PlantCaduceus was able to differentiate between coding and noncoding regions with high clarity.

Table 1. PlantCaduceus model parameters

Models	# of layers	Hidden size	# of parameters (million)
PlantCaduceus_l32	32	1024	225
PlantCaduceus_l28	28	768	112
PlantCaduceus_l24	24	512	40
PlantCaduceus_l20	20	384	20

Improving the accuracy and cross-species transferability of modeling transcription and translation through fine-tuning of PlantCaduceus

Accurate annotation of junction sites, including translation initiation site (TIS), translation termination site (TTS), and splice donor and acceptor sites, is crucial in understanding the genome. To evaluate PlantCaduceus's performance on these tasks, we generated training and validation datasets using *Arabidopsis thaliana* TAIR10, a relatively well-annotated genome released over two decades ago[26]. Specifically, annotated TIS, TTS, and splice donor and acceptor sites were considered as true sites, while the corresponding false sites were randomly selected from the entire genome to match corresponding motifs, such as ATG for TIS, stop codons for TTS, GT for splice donor, and AG for splice acceptor (**Figure 2A**). We used sites from chromosome 5 as the validation set the rest of the genome for training (**Figure 2B** and **METHODS**). Previous LMs focus on evaluation within the same species [19,20,27–29], however, we wanted to investigate whether a model fine-tuned with limited labeled data in Arabidopsis could be used for prediction in other species, given that the DNA LM model is pre-trained on multiple species. Therefore, we generated three extremely imbalanced cross-species testing datasets for rice, sorghum, and maize using BUSCO-supported genes as true sites (**Figure 2B** and **Supplemental Table 3**). We benchmarked the performance of PlantCaduceus against three DNA LMs: GPN [16], AgroNT [18], and Nucleotide Transformer [19], as well as a supervised hybrid model comprising a convolutional neural network (CNN) and a long short-term memory (LSTM) network [30], hereafter referred to as CNN+LSTM. For the Nucleotide Transformer, we used the multi-species version 2 that was pre-trained on 850

genomes, hereafter referred to as NT-v2. For DNA LMs, we kept model weights frozen and trained XGBoost models using embeddings extracted from the last hidden state of each DNA LM (**Figure 2C**). The CNN+LSTM model was trained from scratch in a supervised manner.

First, focusing on within species evaluation on Arabidopsis chromosome 5, PlantCaduceus (32 layers) showed consistently superior performance across the four tasks of predicting TIS (**Figure 2D**), TTS (**Figure 2E**), splice donor site (**Figure 2F**), and splice acceptor site (**Figure 2G**). Other DNA LMs like GPN and AgroNT also performed well, particularly in predicting splice donor and acceptor sites. Additionally, for splice donor and acceptor site prediction, even the supervised CNN+LSTM model achieved near perfect PRAUC values, indicating that within-species prediction is a relatively straightforward task. We then tested the fine-tuned models' performance on test sets in rice, sorghum and maize to assess the cross-species generalization ability of these models. Surprisingly, we found that, for all four tasks, all models except PlantCaduceus showed a meaningful drop in PRAUC when transferred to rice (14.9% to 99.5%), sorghum (19.3% to 99.8%) and maize (41.3% to 100%) (**Figure 2D-2G**). Particularly for the CNN+LSTM model, which performed well on TIS and TTS in Arabidopsis validation and nearly perfectly on splice donor and acceptor sites, the PRAUC dropped significantly when transferred to rice (46.9% to 99.5%), sorghum (68.2% to 99.8%), and maize (86.2% to 100%). This is expected as the supervised model had never seen sequences from these species, making cross-species generalization challenging (**Figure 2D-2G**). GPN maintained decent cross-species predictions but still showed significant drops in PRAUC for TIS and TTS tasks, ranging from 44.9% to 90.4% (**Figure 2D-2G**). As expected, the non-plant DNA NT-v2 model did not perform well on these tasks, due to the significant divergence between plant and animal genomes. But surprisingly, AgroNT also did not perform as well as expected. In contrast, PlantCaduceus consistently maintained high PRAUC values, with only slight decreases in rice, sorghum, and maize, demonstrating its strong cross-species generalization ability. The minor variations among rice, sorghum and maize could be attributed to different numbers of false sites (**Supplemental Table 3**) in each test set.

It is worth noting that the rice and sorghum genomes are included in the pre-training data of both PlantCaduceus and AgroNT, but not for GPN. AgroNT also includes the maize genome. To understand why PlantCaduceus achieved superior performance on these cross-species tasks, we

conducted an ablation test by re-training a custom GPN model using the same datasets as PlantCaduceus and scaled it to 130 million parameters, on the same order of magnitude as PlantCaduceus (**METHODS**). We observed that, including more genomes in the pre-training and scaling model size significantly improved GPN's cross-species predictability (**Supplemental Figure 2**), especially for TIS and TTS tasks. This indicates that when more genomes are included during pre-training, the embeddings learned by the LM are more general across species. However, PlantCaduceus still exhibited the best performance, indicating that its architecture is superior to that of GPN. Moreover, even with a parameter size of 20 million—6.5 times smaller than the custom 130 million GPN and 3.25x times smaller than the original GPN—PlantCaduceus still outperformed all models in predicting TIS, TTS, splice donor, and splice acceptor sites. These results demonstrate that PlantCaduceus not only captures broader evolutionary conservation features but also is more parameter-efficient than other DNA LMs.

Cross-species evolutionary constraint prediction through fine-tuning PlantCaduceus

Genome-wide association studies (GWAS) have identified thousands of variants associated with complex traits. However, identifying causal variants is complicated by linkage disequilibrium (LD), as significant SNPs identified by GWAS are usually in LD with causal variants. In contrast, evolutionary constraint can directly indicate fitness effects by detecting candidate causal mutations through conservation of DNA across species. Given that PlantCaduceus is pre-trained on 16 Angiosperm genomes, we hypothesize that it can be fine-tuned to predict evolutionary constraint using DNA sequences alone. Maize and sorghum, both members of the Andropogoneae clade, descend from a common ancestor approximately 18 million years ago [31]. To generate evolutionary constraints in the sorghum genome, we aligned 34 genomes from the Andropogoneae clade, with rice as an outgroup (**Supplemental Table 4**), to the *Sorghum bicolor* reference genome (**Supplemental Figure 3**). We focused on the 277 million sites with nearly complete coverage and defined those sites with an identity threshold of 15 as conserved versus neutral with an identity threshold of 15 (**Figure 3A**). We used chromosomes 1-9 to train an XGBoost model and evaluated it on sorghum chromosome 10. We benchmarked this task against GPN, AgroNT, NT-v2, and the supervised CNN+LSTM model. On the validation set, PlantCaduceus achieved the best performance, with an AUC of 0.896 (**Figure 3B**) and a PR-AUC of 0.876 (**Figure 3C**). In comparison, the best AUC and PR-AUC for other DNA LMs were 0.778 and 0.790, respectively.

As expected, the supervised CNN+LSTM model performed the worst, with an AUC of 0.638, as it had only seen sequences from sorghum (**Figures 3B-3C**). This demonstrates that PlantCaduceus enables predicting evolutionary constraint without multiple sequence alignment.

To further explore the cross-species predictive power of the model fine-tuned on sorghum evolutionary constraint data, we generated an analogous testing dataset for maize (**METHODS**). Remarkably, when our PlantCaduceus model, originally fine-tuned on sorghum, was applied to the maize dataset, it demonstrated strong cross-species prediction performance, achieving an AUC of 0.829 (**Figure 4D**) and a PR-AUC of 0.797 (**Figure 4E**). In contrast, all other models consistently showed poor performance on maize (**Figure 4D-4E**). As above, we also evaluated the performance of our custom GPN model which was trained on the same dataset as PlantCaduceus. While the custom GPN model showed improved performance with an AUC of 0.8326 and a PR-AUC of 0.8148, PlantCaduceus, with only 20 million parameters, outperformed both the original GPN and the custom GPN models (**Supplemental Figure 6**). These results highlight the robustness and effectiveness of our DNA LM for cross-species predictions of evolutionary constraints using only sequence data as input. The transferability of our model across different species within the Andropogoneae clade suggests that it captures fundamental evolutionary patterns and can be readily adapted to predict evolutionary constraint in related species with limited additional training data.

Zero-shot variant effect prediction identifies deleterious mutations in different species

The training objective of PlantCaduceus is to predict masked nucleotides based on sequence context; if a pre-trained multi-species DNA LM can accurately predict masked tokens, it suggests that similar sequence patterns, conserved across different species, were frequently observed during pre-training. We hypothesize that the predicted likelihood of the reference allele versus the alternate allele can identify deleterious mutations, as mutations in conserved regions across species are likely deleterious. To test this hypothesis, we used a zero-shot strategy to estimate each mutation's effect (**Figure 4A**). For each SNP, we calculated the log-likelihood difference between the reference and alternate alleles, where a more negative value indicates higher conservation. Deleterious mutations tend to have lower frequencies within a population due to selective constraints [32], we therefore used minor allele frequency (MAF) to quantify the deleteriousness of

mutations predicted by different methods. Despite the potential for low MAF in neutral/beneficial alleles, we believe this approach provides useful signals for assessing deleterious mutations [32].

We benchmarked PlantCaduceus against two evolutionary-informed methods, phyloP [24] and phastCons [24], as well as GPN [16]. Both phyloP and phastCons assess evolutionary constraint using multiple sequence alignments and phylogenetic models (**METHODS**), assigning higher scores to conserved regions. For GPN, we used the same zero-shot strategy (**Figure 4A**) as PlantCaduceus. We first analyzed 4.6 million SNPs in the sorghum TERRA population [33] and observed that most of the SNPs showed a neutral zero-shot score, while there was still a heavy tail with negative zero-shot scores (**Figure 4B**). We categorized SNPs into four percentiles based on zero-shot scores: the top 50%, top 10%, top 1%, and top 0.1% most deleterious mutations (**Figure 4B**) and observed that all models showed a decreasing average MAF of SNPs in higher percentiles for missense, nonsynonymous, and noncoding SNPs (**Figure 4C**). Notably, the putative deleterious mutations identified by PlantCaduceus exhibited the lowest average MAF across all percentiles, outperforming GPN and significantly surpassing phyloP and phastCons. Given the success of protein LMs in predicting deleterious missense mutations [8,11], we also incorporated ESM [7] as a benchmark. For missense mutations, we found that PlantCaduceus matches the performance of the state-of-the-art protein language model ESM. At the top 50%, 10%, and 1% percentiles, PlantCaduceus even slightly outperforms ESM.

To assess the model's transferability to unseen genomes, we excluded the maize genome during pre-training, given that maize and sorghum are evolutionarily close species. We then analyzed 9.4 million SNPs in the maize Hapmap 3.2.1 population from 1,224 lines [34] and observed the putative deleterious mutations identified by PlantCaduceus consistently showed the lowest average MAF at different percentiles (**Figure 4D**), demonstrating cross-species generalizability for this task.

However, since GPN is only pre-trained with genomes from eight Brassicales species and specifically designed for mutation effect prediction in Arabidopsis, we further validated PlantCaduceus by analyzing over 10 million mutations from the Arabidopsis 1001 Genomes Project [35]. Being pre-trained with a broader range of evolutionarily distant genomes, PlantCaduceus effectively captured deleterious mutations in Arabidopsis and slightly

outperformed GPN (**Supplemental Figure 5**). For missense mutations, PlantCaduceus matched the performance of the state-of-the-art protein language model ESM and was nearly competitive with GPN for noncoding mutations. These results highlight PlantCaduceus's ability to identify genome-wide deleterious mutations, and demonstrate its broad applicability across diverse species, even those not included in pre-training.

Discussion

Functional annotation of plant genomes is crucial for plant genomics and crop breeding but remains limited by the lack of functional genomic data and accurate predictive models. Here, we introduced PlantCaduceus, a multi-species plant DNA LM pretrained on a well-curated set of 16 evolutionarily distant Angiosperm genomes, enabling cross-species prediction of functional annotations with limited data. PlantCaduceus leverages Mamba [23] and Caduceus [22] architectures to support bi-directional, reverse complement equivariant sequence modeling. We demonstrated the superior cross-species performance of PlantCaduceus on five tasks involving transcription, translation, and evolutionary constraint modeling. These results highlight the potential of PlantCaduceus to serve as a foundational model for comprehensively understanding plant genomes.

PlantCaduceus has the potential to accurately annotate newly sequenced Angiosperm plant genomes. Unlike supervised deep learning models that easily overfit on limited labeled data, PlantCaduceus demonstrates robust cross-species performance in modeling transcription, translation, and evolutionary constraints. This indicates that through self-supervised pre-training on large-scale genomic datasets, PlantCaduceus has captured DNA grammar and evolutionary conservation. The cross-species prediction ability of PlantCaduceus can significantly accelerate plant genomics research, aiding initiatives such as the 1000 Plant Genomes Project [1] by providing accurate annotations and insights across diverse plant species.

PlantCaduceus also offers a more effective approach to estimating deleterious mutations without relying on multiple sequence alignments (MSAs). Deleterious mutations are considered the genetic basis of heterosis, where hybrids yield more due to the suppression of deleterious recessives from

one parent by dominant alleles from the other [36]. Traditionally, deleterious mutations have been identified by generating MSAs [32,37,38] and using evolutionary methods such as phyloP [24] and phastCons [24]. However, the prevalence of transposable elements and polyploidy in plant genomes complicates the MSA generation [39,40]. PlantCaduceus overcomes these challenges by using a masked language modeling strategy to learn the conservation from large scale genomic datasets of diverse species. Promisingly, the deleterious mutations prioritized by PlantCaduceus with the zero-shot strategy showed three-fold rare allele enrichment compared to phyloP and phastCons, and our approach is also competitive with state-of-the-art protein LM for missense mutations. These results suggest that PlantCaduceus can be utilized to identify genome-wide deleterious mutations across diverse crop species, enhancing crop breeding by optimizing parental line selection and thus promoting hybrid vigor [36].

In future work, we plan to incorporate additional plant genomes from diverse lineages, such as gymnosperms, to capture broader evolutionary conservation. Additionally, we plan to pre-train PlantCaduceus with longer context windows, enabling it to capture long-range DNA interactions and better handle tasks benefiting from long-range cis-effects, such as, allele-specific expression, chromatin state prediction, and chromatin interaction mapping. Furthermore, it would also be interesting to explore how to better tokenize repetitive sequences in plant genomes. We envision that these approaches will allow us to push the boundaries of what PlantCaduceus can achieve, establishing it as an even more powerful and versatile foundation model for advancing genomic research and facilitating crop improvement.

Methods

Pre-training dataset

The pre-training dataset comprises 16 genomes from two distinct clades: eight genomes from the family Poaceae and eight genomes from the order Brassicales (**Supplemental Table S1**). The Poaceae species displayed substantial variation in genome size and repetitive sequence content, with the hexaploid wheat genome exhibiting a size of 15 Gbp. For each Poaceae genome, except for Tripsacum, we obtained the genome and corresponding genome annotation and repeat-masked annotation from the Joint Genome Institute (JGI). For the Tripsacum genome, the genome FASTA

and annotation files were downloaded from MaizeGDB (https://maizegdb.org/genome/assembly/Td-FL_9056069_6-DRAFT-PanAnd-1.0), and the EDTA tool [41] was used to identify repetitive sequences within the genome. Based on the repeat-masked annotation, each genome was softmasked with bedtools [42] and subsequently divided into genomic windows of 512 bp with a step size of 256 bp. Each window was assigned to a unique class based on the genome annotation, and all coding sequence regions were selected for pre-training. The remaining genomic regions were then down-sampled to ensure an equal number of CDS regions and noncoding regions. It is important to note that for the hexaploid wheat genome, only subgenome A was utilized to avoid species bias. The Brassicales genomes datasets were acquired from a Hugging Face repository (https://huggingface.co/datasets/songlab/genomes-brassicales-balanced-v1). The validation and testing datasets were randomly selected and constituted 5% of the total dataset.

Caduceus model architecture and pre-training

We use the recently proposed Caduceus architecture [22], which is tailored to three important aspects of DNA sequence modeling. Caduceus is built off of the Mamba architecture [23], a recently proposed structured state space model which scales to long sequences more efficiently than attention-based methods while maintaining accuracy. To account for upstream and downstream gene interactions, Caduceus employs weight sharing to enable memory-efficient bi-directionality. Finally, Caduceus is designed to consider the reverse complement (RC) symmetry of DNA sequences. This is accomplished by encoding RC equivariance as an inductive bias: the Caduceus language model commutes with the RC operation. Combining these three design decisions, Caduceus has shown promising results when applied to human genome modeling [22].

The implementation of RC equivariance in Caduceus entails doubling the number of channels for intermediate representations. At a high level, half the channels are used to encode information about a sequence and the other half are used to encode information about its RC. For downstream tasks in which we fine-tuned a classifier on top of learned embeddings, the labels were invariant to the RC operation, since both DNA strands carry the same label. To account for this, we therefore split embeddings of the Caduceus model along the channel dimension and averaged. This ensures

that both a sequence and its RC will have the same final embedding, i.e., we render the embeddings invariant to the RC operation as well.

For the pre-training of PlantCaduceus, each model was trained for 480,000 steps using a Decoupled AdamW optimizer [43] with the global batch size of 2,048. The learning rate is 2E-4 with a cosine decay scheduler, and 6% of the training duration was dedicated to warm up. The learning rate decayed to 4E-6 by the end of training. The default BERT [44] masking recipe was used with a masking probability of 0.15. For each masked token: (i) there is an 80% probability it will be replaced by a special token ([MASK]), (ii) a 10% probability it will be replaced by a random token, and (iii) a 10% probability it will remain unchanged. Unless otherwise specified, all models were trained using a sequence length of 512 base pairs. A weight decay of 1E-5 was applied throughout the training process.

TIS, TTS, splice donor and acceptor training, validation and testing dataset generation

To generate high-quality training datasets for translation initiation sites (TIS), translation termination sites (TTS), splice donor sites, and splice acceptor sites, we used the well-annotated model plant genome of Arabidopsis with Araport 11 annotation [45]. To accurately reflect the inherent imbalance in junction sites prediction, all annotated junction sites were considered as positive observations, while a randomly selected subset of sites (5%) that matched specific appropriate motifs (e.g., ATG for TIS, UAA, UAG, and UGA for TTS, GT for donor splice sites, and AG for acceptor splice sites) were used as negative observations. For each task, the pre-trained model weights were frozen, and XGBoost models (n_estimators=1000, max_depth=6, learning_rate=0.1) were trained using embeddings extracted from the last hidden state of the pre-trained model. To ensure robust model training and validation, chromosome 5 was used for validation (**Figure 2A-2B**), and the rest of the Arabidopsis genome was used for training.

Given the relatively poor annotation in other species compared to Arabidopsis, to generate reliable testing datasets in other species, we used the BUSCO tool [46] to identify 3,236 orthologous genes specific to monocotyledons in rice, sorghum and maize. This approach ensures that the selected annotated genes are highly conserved and likely to be correctly annotated, mitigating the issue of

inaccurate performance evaluations. Specifically, BUSCO was utilized to scan the annotated protein isoforms, and only complete BUSCO genes were considered as true positives. For those BUSCO genes with multiple transcripts, we selected the longest transcript to avoid sequence redundancy in the testing dataset. Subsequently, BUSCO gene/transcript-supported junction sites were used as positive examples for their respective tasks. To generate negative sites, all sites within the BUSCO genes that matched appropriate motifs (e.g., ATG for TIS, TAA, TAG, and TGA for TTS, GT for donor splice sites, and AG for acceptor splice sites) but were not part of any annotated gene models were used as true sites. Sites belonging to alternate transcripts were excluded to avoid ambiguity. Furthermore, to expand the negative observations and capture a broader range of non-junction sites, we included sites in the intergenic regions flanking the BUSCO genes that matched the appropriate junction motifs. By incorporating both genic and intergenic sites from the BUSCO gene set as negatives, we created an extremely imbalanced testing dataset to reflect the real-world scenario of junction site prediction (**Supplemental Table S3).**

Evolutionary constraint estimation

The evolutionary constraint was estimated primarily within the Andropogoneae tribe, a large clade of grasses comprising approximately 1,200 species that descended from a common ancestor approximately 18 million years ago [31]. In this analysis, 34 genomes from Andropogoneae and the rice genome were used to estimate the evolutionary constraint. Due to the substantial transposable element (TE) content in these genomes, AnchorWave, a sensitive genome-to-genome alignment tool [39], was used to align the 35 genomes to the sorghum reference genome using the parameters "-R 1 -Q 1". Following the alignments to the sorghum reference genome, we counted the number of identities, SNPs, and coverages (**Supplemental Figure 5**). Then the fine-tuned labels were generated based on per-site identity and coverage (**Figure 4A**). Conserved sites were defined as having an identity greater than 34, while neutral sites were defined as having an identity of 15 or less and coverage of at least 34. Sites with low coverage were excluded due to their potential ambiguity. Given the large size of the training dataset, only 5% of conserved sites were randomly selected for training, and an equivalent proportion of neutral sites was also randomly selected. Sites from chromosomes 1 to 9 were used for training, while those from chromosome 10 were used for validation. To generate the testing dataset in maize, the maize reference genome B73 was

used. Then, using the same approach, genome-wide evolutionary constraints were generated by aligning 35 genomes to the maize reference genome with AnchorWave, using the parameters "-R 1 -Q 2," except for Tripsacum clades. For Tripsacum and maize, which share the most recent whole genome duplication, we used "-R 1 -Q 1".

phyloP and phastCons calculation

With the same 34 genomes from Andropogoneae, we generated pairwise genome-to-genome alignments using Cactus [47], a multiple genome alignment tool that uses a progressive alignment strategy. The neutral model was calculated from fourfold degenerate coding sites across the entire genome. The resulting alignments were then analyzed using PHAST [24] to quantify evolutionary conservation with phyloP conservation scores – using the SPH scoring method (--method SPH) and CONACC mode (--mode CONACC) – and phastCons scores.

GPN, custom GPN, AgroNT and NT-v2 baselines

To comprehensively evaluate our foundation model's performance, four foundation models including GPN [16], custom GPN, AgroNT [18] and NT-v2 [19] were used as baselines for various tasks. GPN is a convolutional DNA LM pre-trained on eight genomes of Arabidopsis and seven other species from the Brassicales order. However, since GPN was pre-trained with only eight evolutionarily close species and has only 65M parameters and most of the tasks in this paper focus on evaluation in crops, we re-trained a custom GPN with 130M parameters using 50 convolutional layers and the same dataset as PlantCaduceus for a fair comparison. The other hyperparameters were kept identical to the original GPN (**Supplemental Table S5**). In contrast, AgroNT [18] is a transformer-based [48] language model with 1 billion parameters, pre-trained on 48 plant genomes. NT-v2 [19], is a non-plant multi-species transformer model pre-trained on 850 genomes excluding plant species. These models employ different tokenization strategies: GPN uses single-nucleotide tokenization, while AgroNT and NT-v2 use 6-mer tokenization. To ensure a fair comparison, we extracted the middle token embeddings for GPN and the middle k-mer token embeddings for AgroNT and NT-v2.

Supervised CNN+LSTM baseline

To establish a fair comparison between our DNA LM and existing supervised models, which are primarily trained on human data, we used the DanQ model architecture [30] as the supervised baseline. DanQ is a hybrid convolutional and recurrent neural network specifically designed for predicting the function of DNA sequences. It has demonstrated impressive performance in predicting chromatin states in plant species, making it a suitable choice for our comparative analysis [49]. For each task, the CNN+LSTM model was trained from scratch using one-hot encoded DNA sequences as input. The Adam optimizer with a learning rate of 0.01 was employed for model optimization. The batch size was set to 2,048. Early stopping with a patience of 20 steps was implemented.

Data availability

The pre-training genomes are available at: https://huggingface.co/datasets/kuleshov-group/Angiosperm_16_genomes. All fine-tuned datasets are available at Hugging Face: https://huggingface.co/datasets/kuleshov-group/cross-species-single-nucleotide-annotation

Code availability

The pre-trained models, along with documentation on how to use them, are available at Hugging Face: https://huggingface.co/collections/kuleshov-group/plantcaduceus-512bp-len-665a229ee098db706a55e44a

Acknowledgments

This work is funded by the USDA-ARS, NSF PanAnd grant (#1822330), NSF CAREER grant (#2145577) and NIH MIRA grant (#1R35GM151243-01). We thank Edgar Marroquin (Cornell University) for discussing fine-tuning tasks, Travis Wrightsman (Cornell University) for providing DanQ code, Arun S. Seetharam and Matthew B Hufford (Iowa State University) for sharing Andropogoneae assemblies, Merritt Khaipho-Burch (Cornell University) for sharing the liftover version HapMap3 VCF file, and all members of the E.S.B. laboratory (Cornell University) for helpful discussions. We would also like to thank the SCINet project, the AI Center of Excellence

of the USDA Agricultural Research Service (0201-88888-003-000D and 0201-88888-002-000D) and MosaicML for providing compute resources for pre-training and fine-tuning experiments.

References

1. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
2. Dudnyk, K., Cai, D., Shi, C., Xu, J. & Zhou, J. Sequence basis of transcription initiation in the human genome. *Science* **384**, eadj0116 (2024).
3. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).
4. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
6. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
7. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
8. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).
9. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
10. Ruffolo, J. A. & Madani, A. Designing proteins with language models. *Nat. Biotechnol.* **42**, 200–202 (2024).
11. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

12. Engelhorn, J. *et al.* Genetic variation at transcription factor binding sites largely explains phenotypic heritability in maize. *bioRxiv* 2023.08.08.551183 (2024) doi:10.1101/2023.08.08.551183.

13. Gaulton, K. J., Preissl, S. & Ren, B. Interpreting non-coding disease-associated human variants using single-cell epigenomics. *Nat. Rev. Genet.* **24**, 516–534 (2023).

14. Leeman-Neill, R. J. *et al.* Noncoding mutations cause super-enhancer retargeting resulting in protein synthesis dysregulation during B cell lymphoma progression. *Nat. Genet.* **55**, 2160–2174 (2023).

15. Novák, P. *et al.* Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nat Plants* **6**, 1325–1329 (2020).

16. Benegas, G., Batra, S. S. & Song, Y. S. DNA language models are powerful predictors of genome-wide variant effects. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2311219120 (2023).

17. Zhou, H., Shrikumar, A. & Kundaje, A. Towards a Better Understanding of Reverse-Complement Equivariance for Deep Learning Models in Genomics. in *Proceedings of the 16th Machine Learning in Computational Biology meeting* (eds. Knowles, D. A., Mostafavi, S. & Lee, S.-I.) vol. 165 1–33 (PMLR, 22--23 Nov 2022).

18. Mendoza-Revilla, J. *et al.* A Foundational Large Language Model for Edible Plant Genomes. *bioRxiv* 2023.10.24.563624 (2023) doi:10.1101/2023.10.24.563624.

19. Dalla-Torre, H. *et al.* The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv* 2023.01.11.523679 (2023) doi:10.1101/2023.01.11.523679.

20. Nguyen, E. *et al.* HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *arXiv [cs.LG]* (2023).

21. Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with Evo. *bioRxiv* 2024.02.27.582234 (2024) doi:10.1101/2024.02.27.582234.

22. Schiff, Y. *et al.* Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling. *arXiv [q-bio.GN]* (2024).

23. Gu, A. & Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv [cs.LG]* (2023).

24. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011).

25. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).

26. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).

27. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

28. Zhou, Z. *et al.* DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes. (2023).

29. Zhang, D. *et al.* DNAGPT: A Generalized Pretrained Tool for Multiple DNA Sequence Analysis Tasks. *bioRxiv* 2023.07.11.548628 (2023) doi:10.1101/2023.07.11.548628.

30. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107 (2016).

31. Welker, C. A. D. *et al.* Phylogenomics enables biogeographic analysis and a new subtribal classification of Andropogoneae (Poaceae—Panicoideae). *J. Syst. Evol.* **58**, 1003–1030 (2020).

32. Ramstein, G. P. & Buckler, E. S. Prediction of evolutionary constraint by genomic annotations improves functional prioritization of genomic variants in maize. *Genome Biol.* **23**, 183 (2022).

33. Lozano, R. *et al.* Comparative evolutionary genetics of deleterious load in sorghum and maize. *Nature Plants* **7**, 17–24 (2021).

34. Bukowski, R. *et al.* Construction of the third-generation Zea mays haplotype map. *Gigascience* **7**, 1–12 (2018).

35. 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at & 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).

36. Crow, J. F. 90 years ago: the beginning of hybrid maize. *Genetics* **148**, 923–928 (1998).

37. Mezmouk, S. & Ross-Ibarra, J. The pattern and distribution of deleterious mutations in maize. *G3* **4**, 163–171 (2014).

38. Lye, Z., Choi, J. Y. & Purugganan, M. D. Deleterious Mutations and the Rare Allele Burden on Rice Gene Expression. *Mol. Biol. Evol.* **39**, (2022).

39. Song, B. *et al.* AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proc. Natl. Acad. Sci. U. S. A.* **119**, (2022).

40. Song, B., Buckler, E. S. & Stitzer, M. C. New whole-genome alignment tools are needed for tapping into plant diversity. *Trends Plant Sci.* **29**, 355–369 (2024).

41. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

42. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

43. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *arXiv [cs.LG]* (2017).

44. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv [cs.CL]* (2018).

45. Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J.* **89**, 789–804 (2017).

46. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

47. Paten, B. *et al.* Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011).

48. Vaswani, A. *et al.* Attention Is All You Need. *arXiv [cs.CL]* (2017).

49. Wrightsman, T., Marand, A. P., Crisp, P. A., Springer, N. M. & Buckler, E. S. Modeling chromatin

state from sequence across angiosperms using recurrent convolutional neural networks. *Plant Genome* **15**, e20249 (2022).

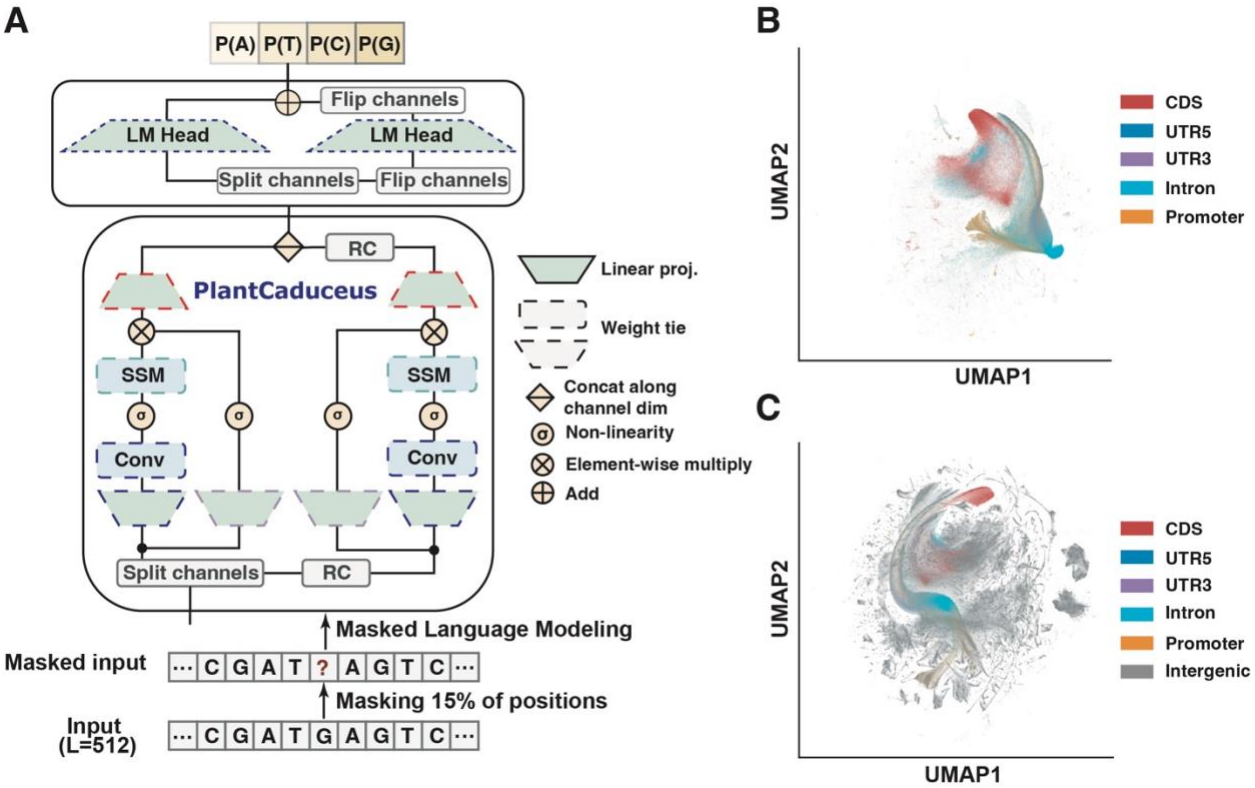

Figure 1. Overview of PlantCaduceus. (A) The input for PlantCaduceus consists of 512-bp DNA sequences with 15% of positions randomly masked. The pre-training objective is cross-entropy loss on the masked positions. The sequences are processed through the bi-directional Caduceus architecture, which is based on the Mamba sequence operator—a recently proposed structured state space model. Caduceus also contains a reverse complement equivariance inductive bias. **(B)** UMAP visualization of embeddings from PlantCaduceus (32 layers) averaged over non-overlapping 100-bp windows along the sorghum genome without intergenic regions. **(C)** The same UMAP visualization as in **(B)** but with intergenic regions.

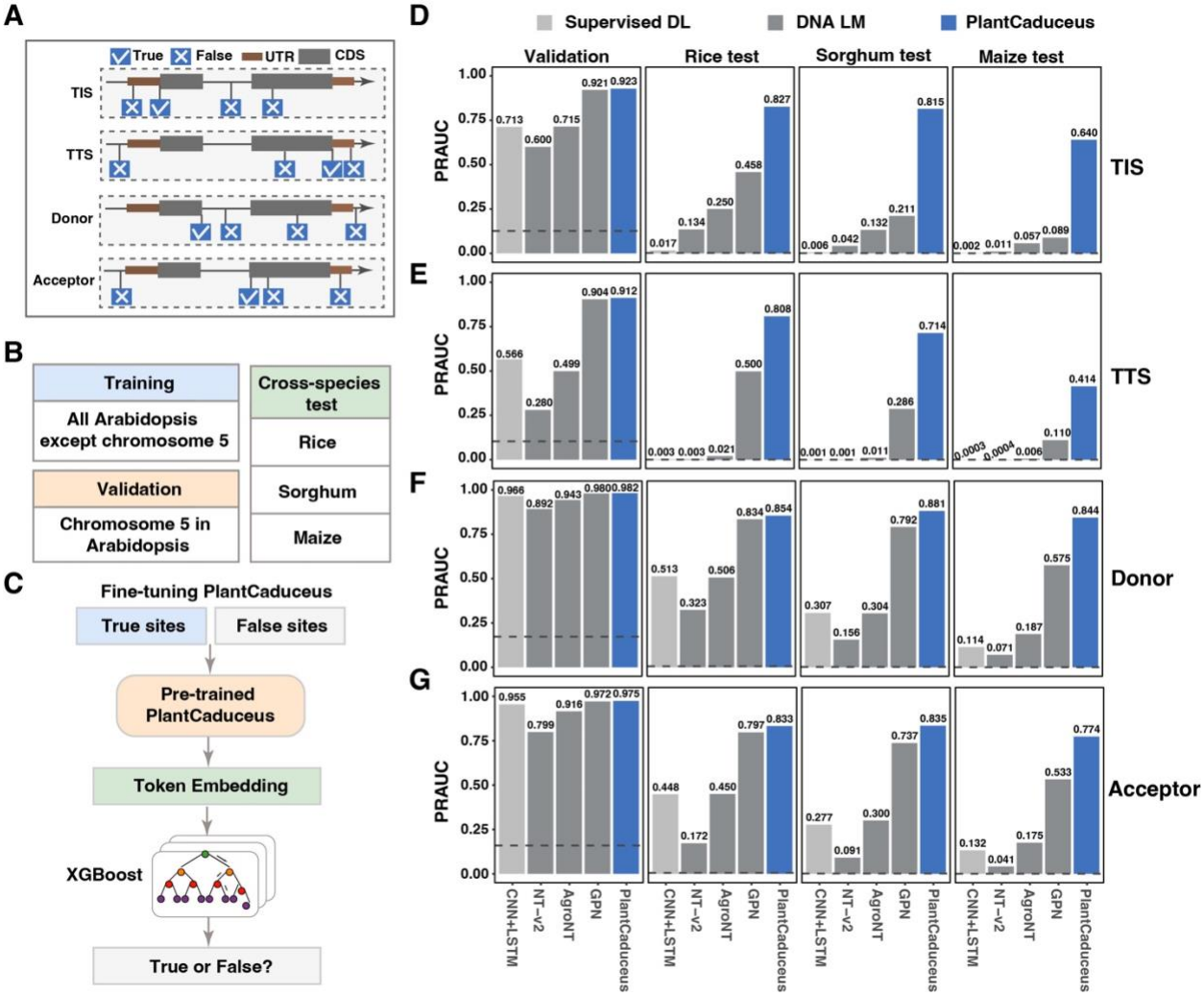

Figure 2. Modeling translation and transcription through fine-tuning PlantCaduceus. (A) TIS, TTS, splice donor and splice acceptor dataset generation. Annotation supported sites are considered as true, randomly selected sites from the genome but matching corresponding motifs are considered as false sites. **(B)** Sites from Arabidopsis are used as training and validation, the chromosome 5 in Arabidopsis is considered as validation, the rest of the Arabidopsis is used for training. Three cross-species tests are generated for rice, sorghum and maize. **(C)** Fine-tuning strategy for PlantCaduceus, the weights of pre-trained PlantCaduceus are kept frozen during pre-training, and then the last hidden state of PlantCaduceus were used as features of XGBoost model. **(D-G)** bar plots display the PRAUC scores for validation, rice test, sorghum test, and maize test datasets for four tasks: TIS **(D)**, TTS **(E)**, splice donor **(F)**, and splice acceptor **(G)**. Blue bars represent our PlantCaduceus model with 32 layers. Gray bars denote three DNA language models: NT-v2, AgroNT, and GPN. Light gray bars represent a traditional supervised model, a hybrid of CNN and LSTM. The gray dashed line in each panel indicates the baseline for each dataset, corresponding to the negative sample ratio.

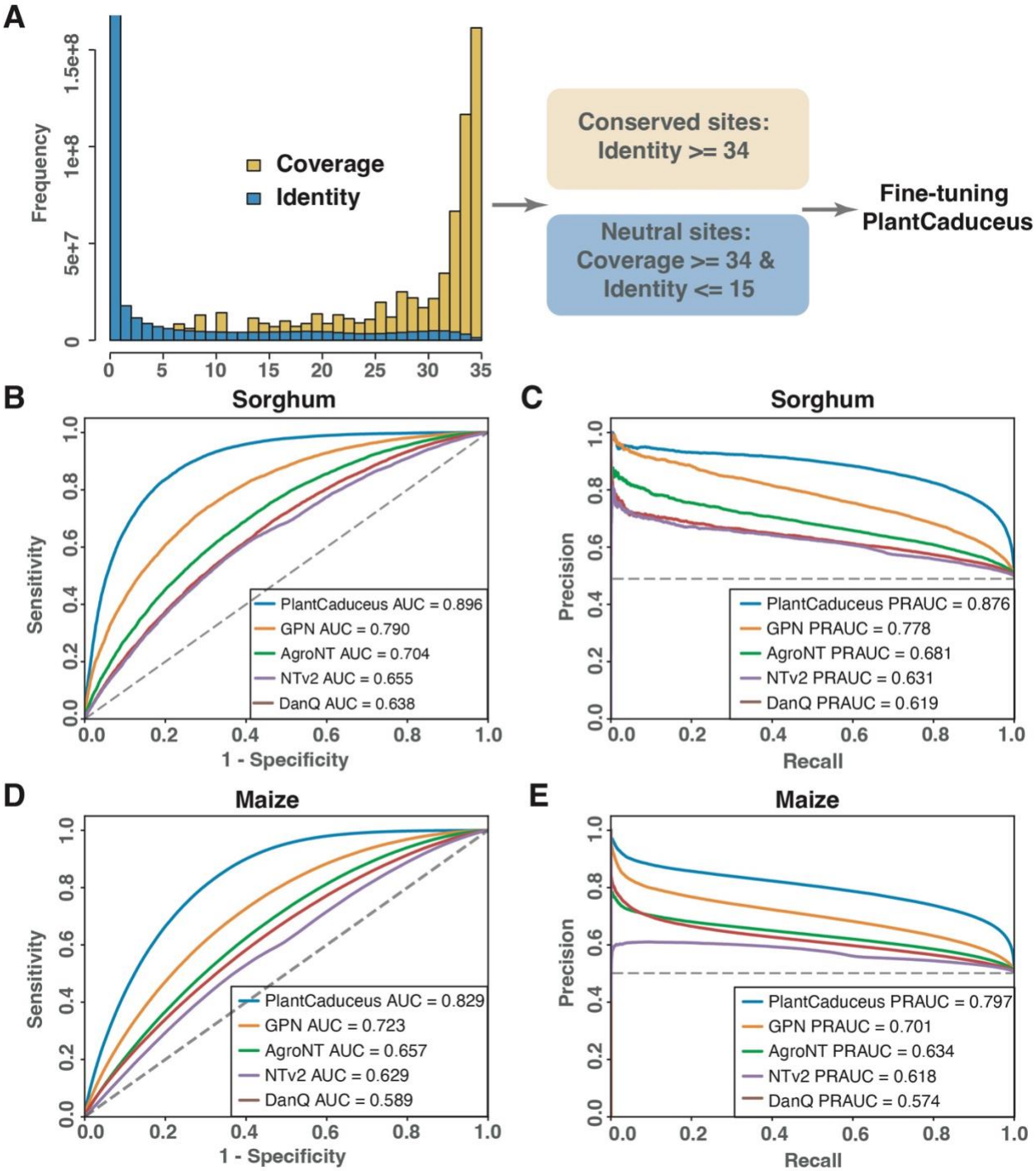

Figure 3. Evolutionary constraint prediction. (A) Illustration of the evolutionary conservation data curation. **(B)** Receiver operating characteristic (ROC) and **(C)** precision-recall (PR) curves of different models in sorghum. **(D)** ROC and **(E)** PR curves of transferring different models trained in sorghum to unseen maize data.

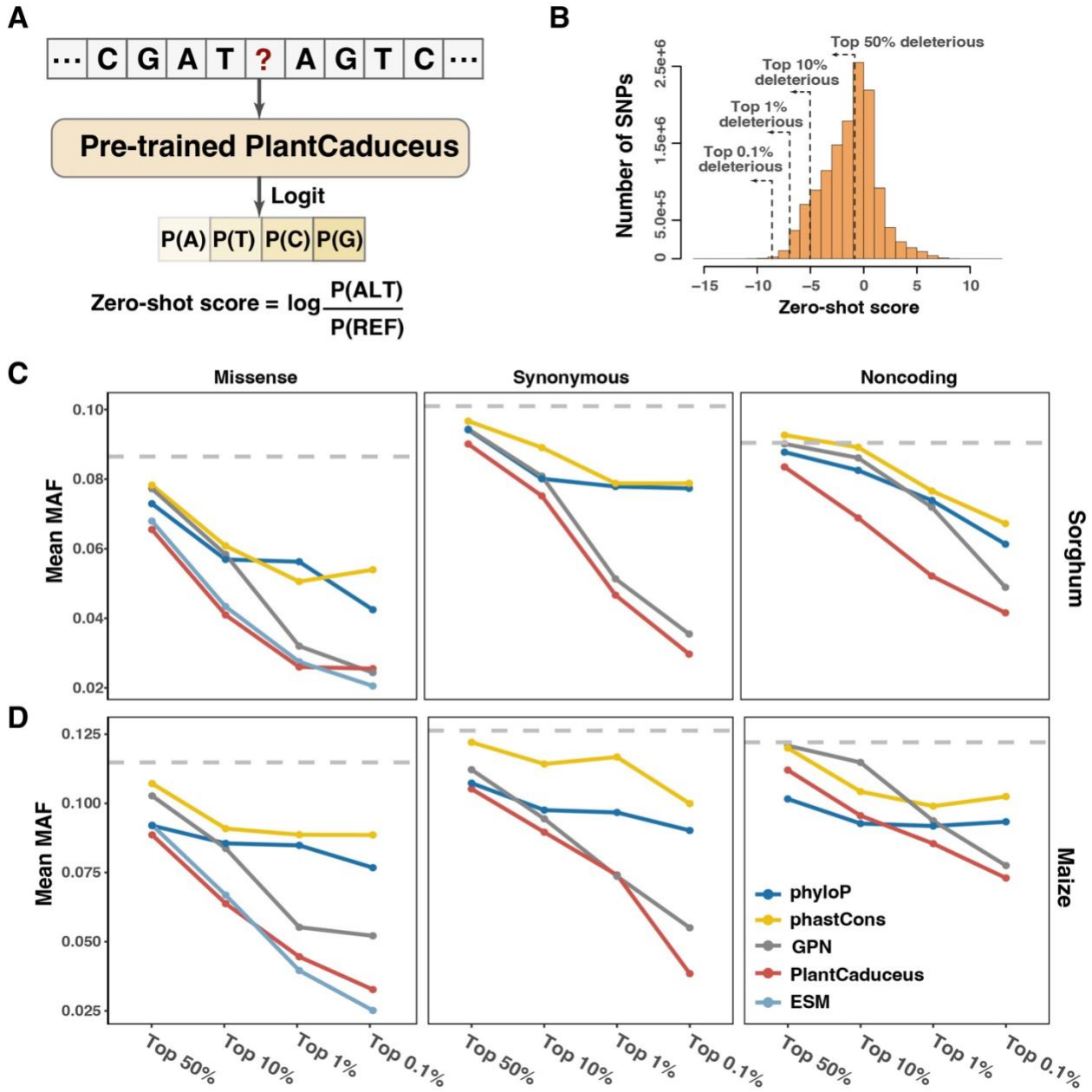

Figure 4. Deleterious mutations identification. (A) The zero-shot strategy of PlantCaduceus for identifying deleterious mutations. **(B)** The distribution of zero-shot scores in the Sorghum TERRA population. **(C)** The mean minor allele frequency (MAF) of putative deleterious mutations prioritized by different models in sorghum. **(D)** The MAF of putative deleterious mutations prioritized by different models in maize.