Supporting Information for

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], Wei-Yun Lai[1], Zachary R. Miller[1], Armin Scheben[5], Michelle C. Stitzer[1], M. Cinta Romay[1], 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

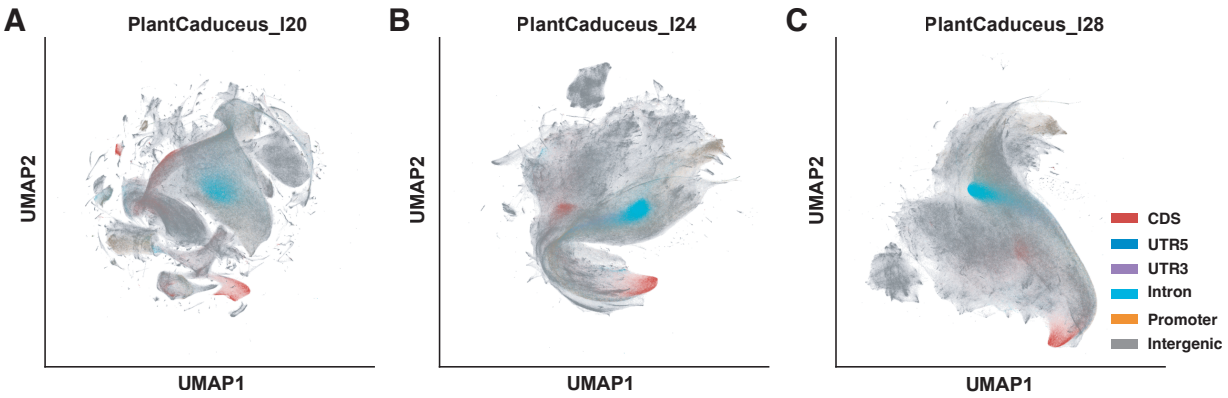
3 Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY USA 14853

4 USDA-ARS; Ithaca, NY, USA 14853

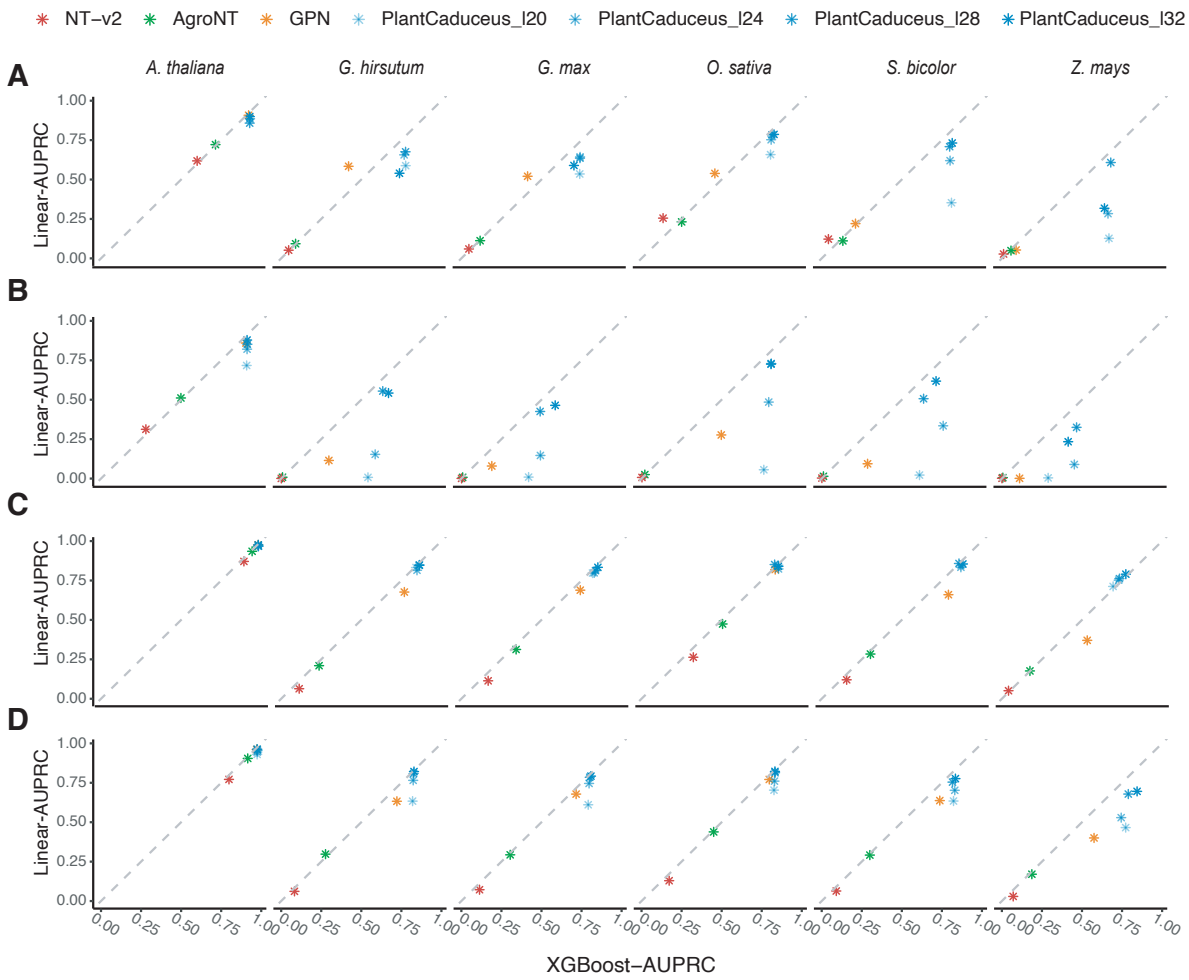
5 Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY USA 11724

* These authors contributed equally to this work

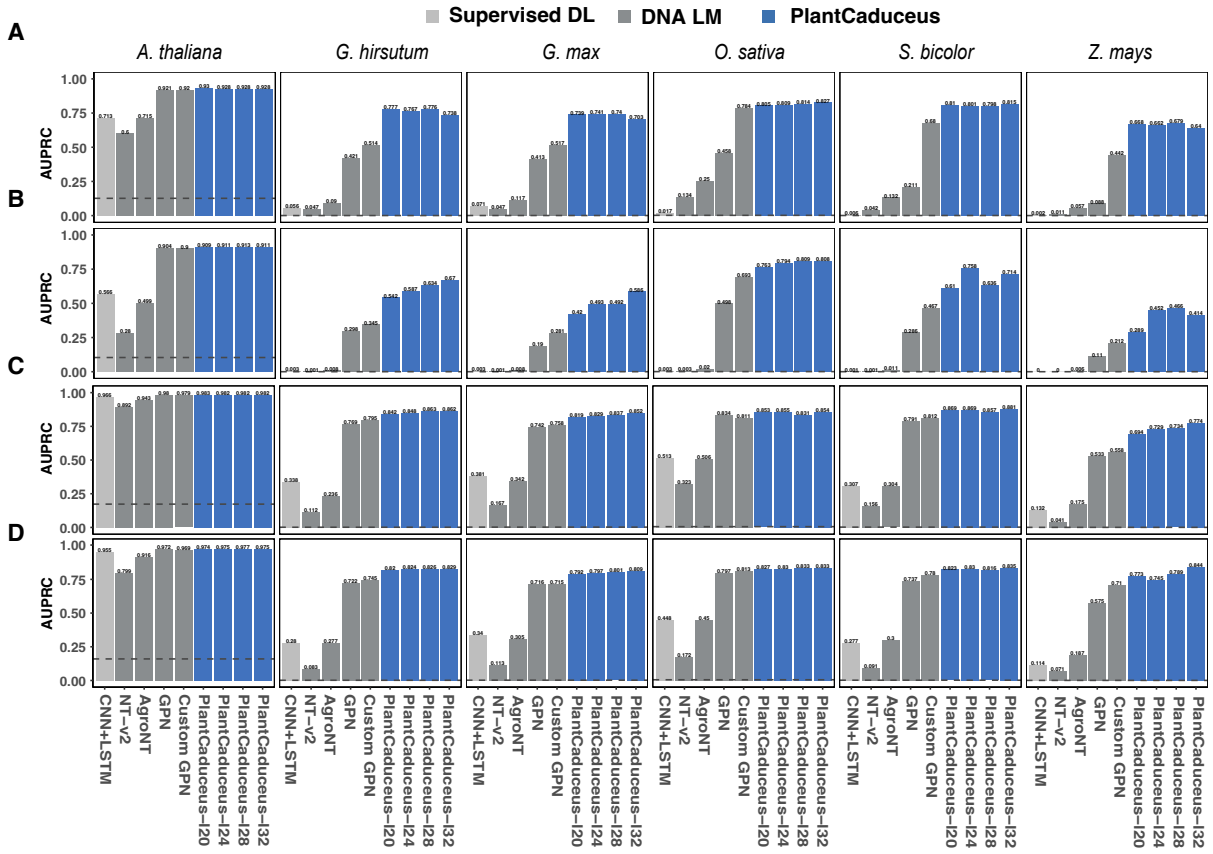
+ To whom correspondence may be addressed. Email: jz963@cornell.edu, ed.buckler@usda.gov and vk379@cornell.edu



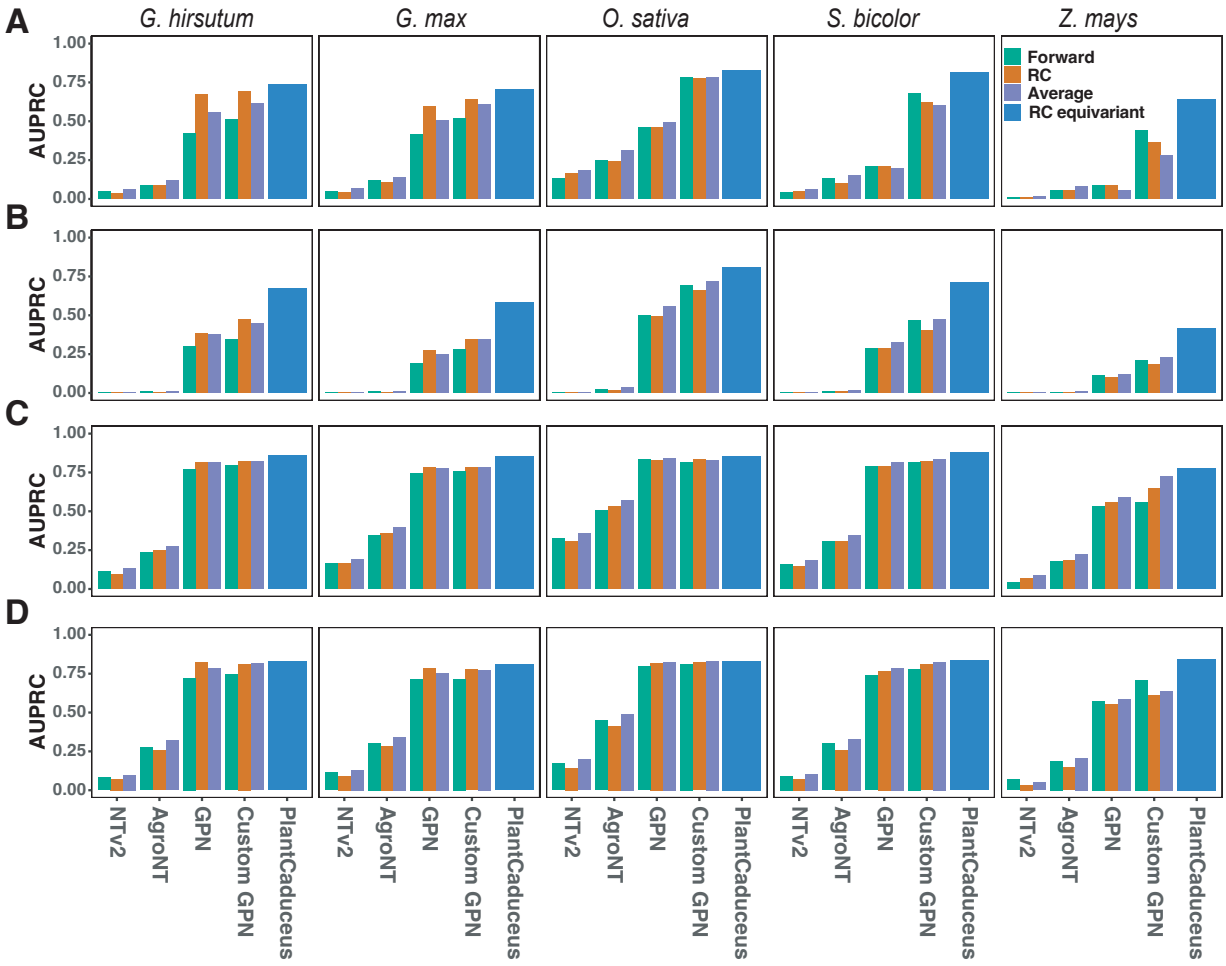
Supplemental Fig. 1. Unsupervised clustering of genomic windows. UMAP visualization of embeddings from PlantCaduceus_l20 **(A)**, PlantCaduceus_l24 **(B)**, and PlantCaduceus_l28 **(C)** averaged over non-overlapping 100-bp windows along the sorghum genome. PlantCaduceus_l20, PlantCaduceus_l24, and PlantCaduceus_l28 represent PlantCaduceus models with 20 layers, 24 layers, and 28 layers, respectively.



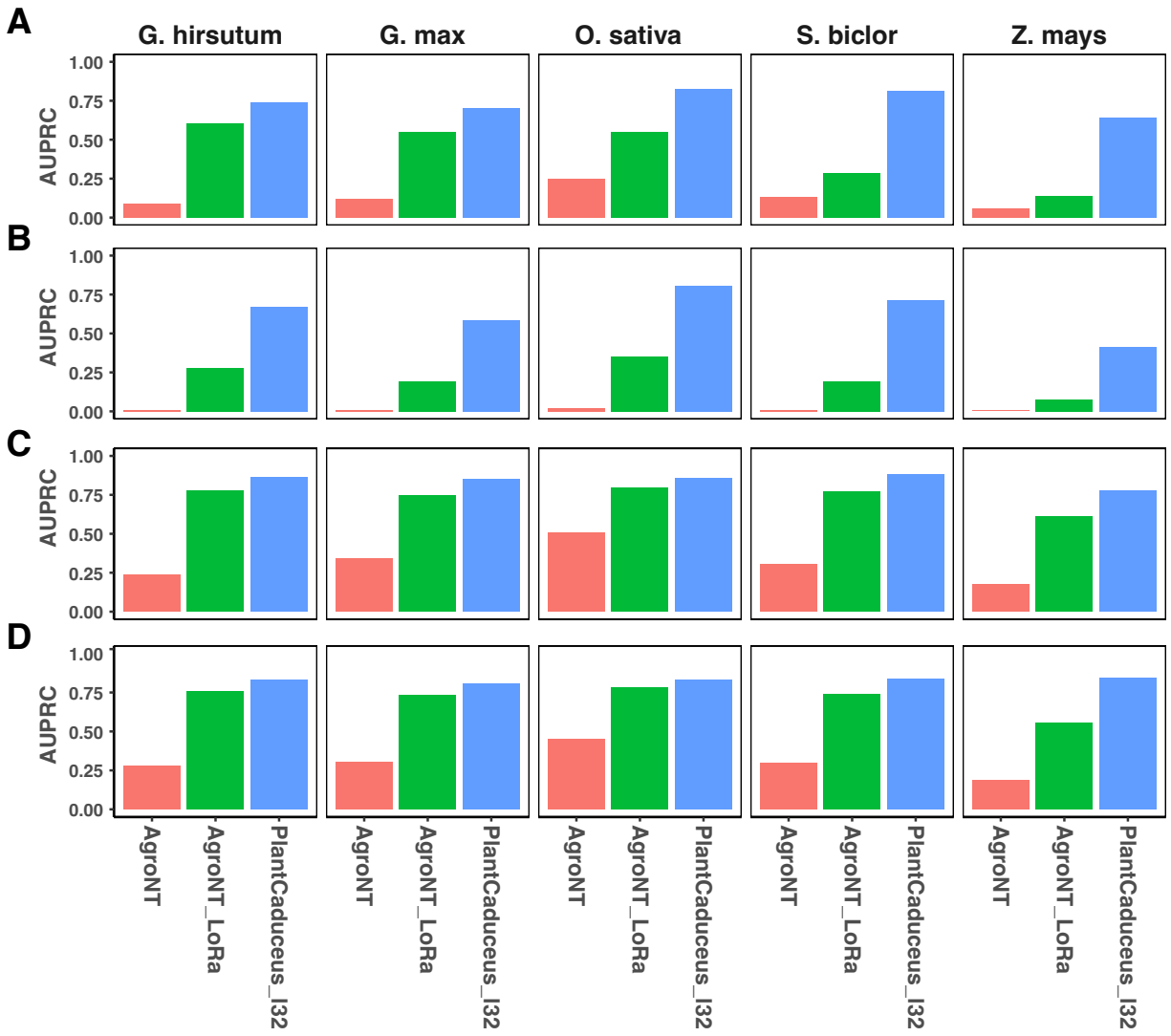
Supplemental Fig. 2. Comparative analysis of XGBoost classification versus single linear layer performance across different genomic element prediction tasks. Performance comparison between XGBoost classification and a single linear layer for predicting **(A)** TIS, **(B)** TTS, **(C)** splice donors, and **(D)** splice acceptors. The x-axis shows the Area Under the Precision-Recall Curve (AUPRC) obtained using XGBoost trained on frozen embeddings from various DNA language models, while the y-axis displays the AUPRC achieved using a single linear layer trained on the same frozen embeddings. Each point represents the performance of a different DNA language model.



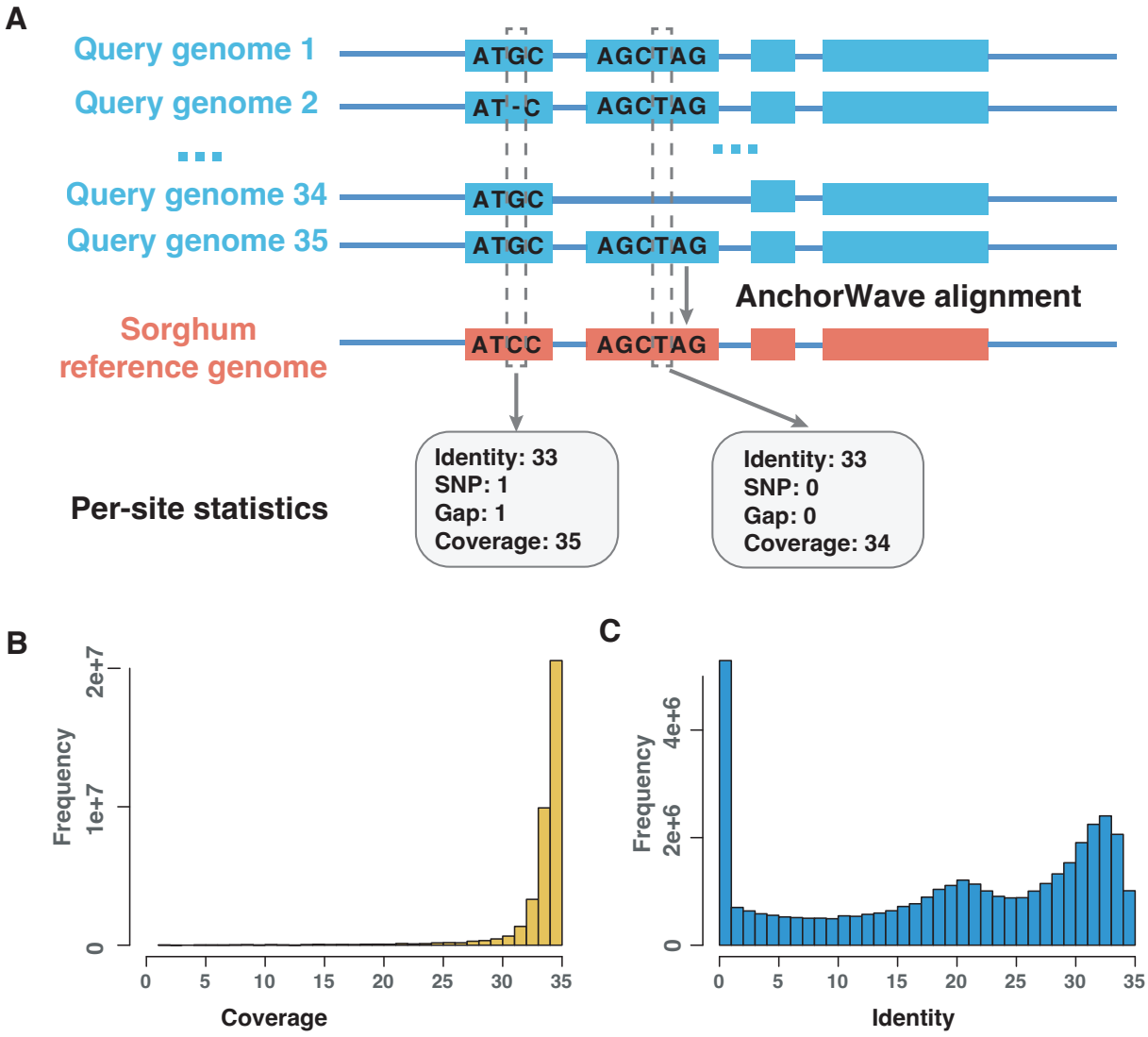
Supplemental Fig. 3. Performance evaluation of TIS, TTS, splice donor and acceptor prediction tasks. Bar plots displaying the AUPRC scores of different models across various species: *A. thaliana* (with-species), *G. hirsutum*, *G.max*, *O.sativa*, *S.bicolor* and *Z. mays* for four tasks: TIS **(A)**, TTS **(B)**, donor **(C)**, and acceptor **(D)**. The blue bars represent our four PlantCaduceus models with varying layers. The gray bars denote three DNA language models: NT-v2, AgroNT, GPN, and Custom GPN (pre-trained with the same data as PlantCaduceus). The light gray bars represent a traditional supervised model, which is a hybrid of CNN and LSTM. The gray dashed line in each panel represents the baseline for each dataset, which corresponds to the negative sample ratio.



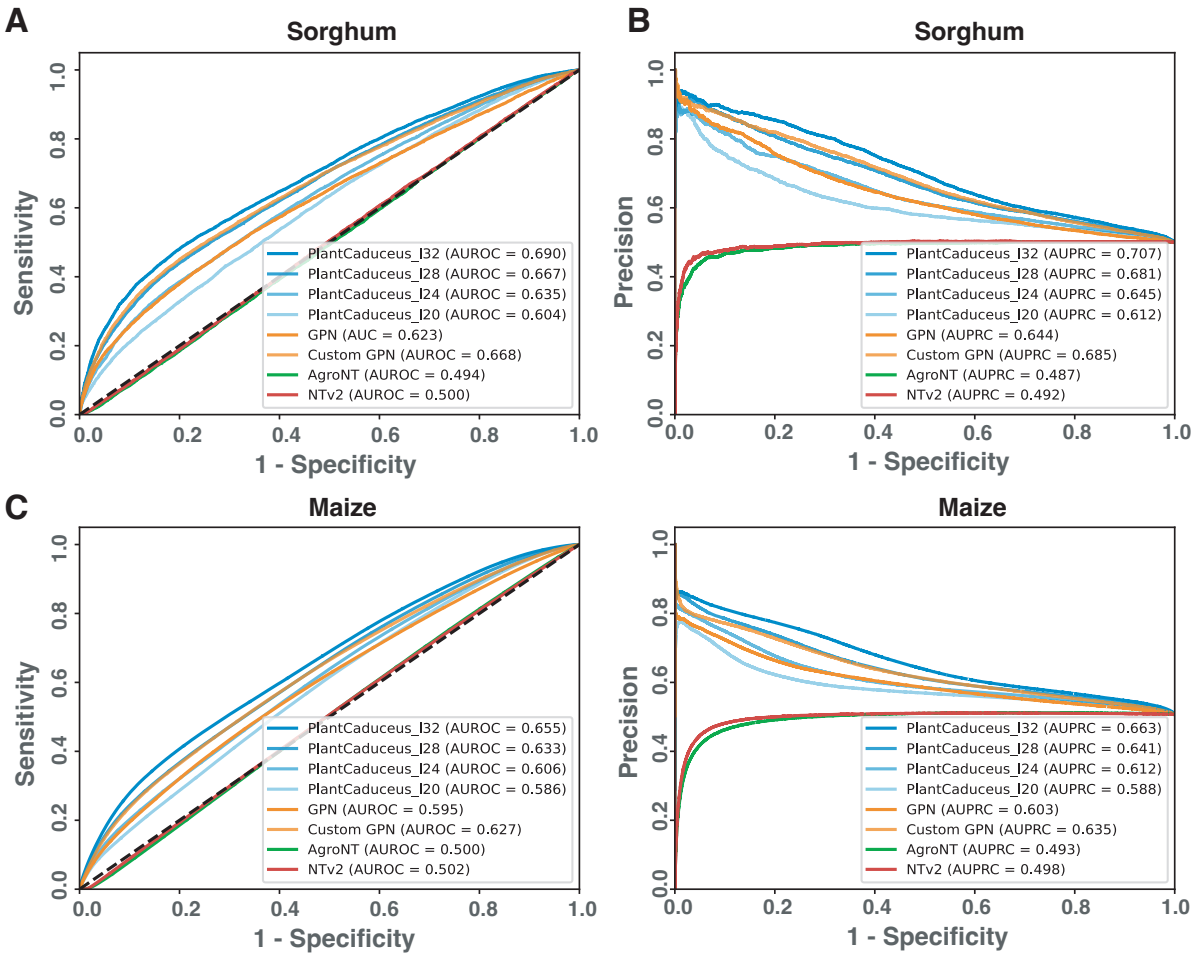
Supplemental Fig. 4. Evaluation of reverse complement (RC) sequence handling in DNA language models across gene annotation tasks. Performance comparison of non-RC equivariant DNA language models using different sequence orientations for predicting **(A)** TIS, **(B)** TTS, **(C)** splice donors, and **(D)** splice acceptors. The bar plots show AUPRC across six plant species: *A. thaliana* (with-species), *G. hirsutum*, *G.max*, *O.sativa*, *S.bicolor* and *Z. mays*. AUPRC values were obtained using XGBoost trained on frozen embeddings under four conditions: forward sequences (green), reverse complement sequences (orange), averaged predictions from forward and reverse complement sequences (purple), and RC-equivariant models (blue) as a reference.



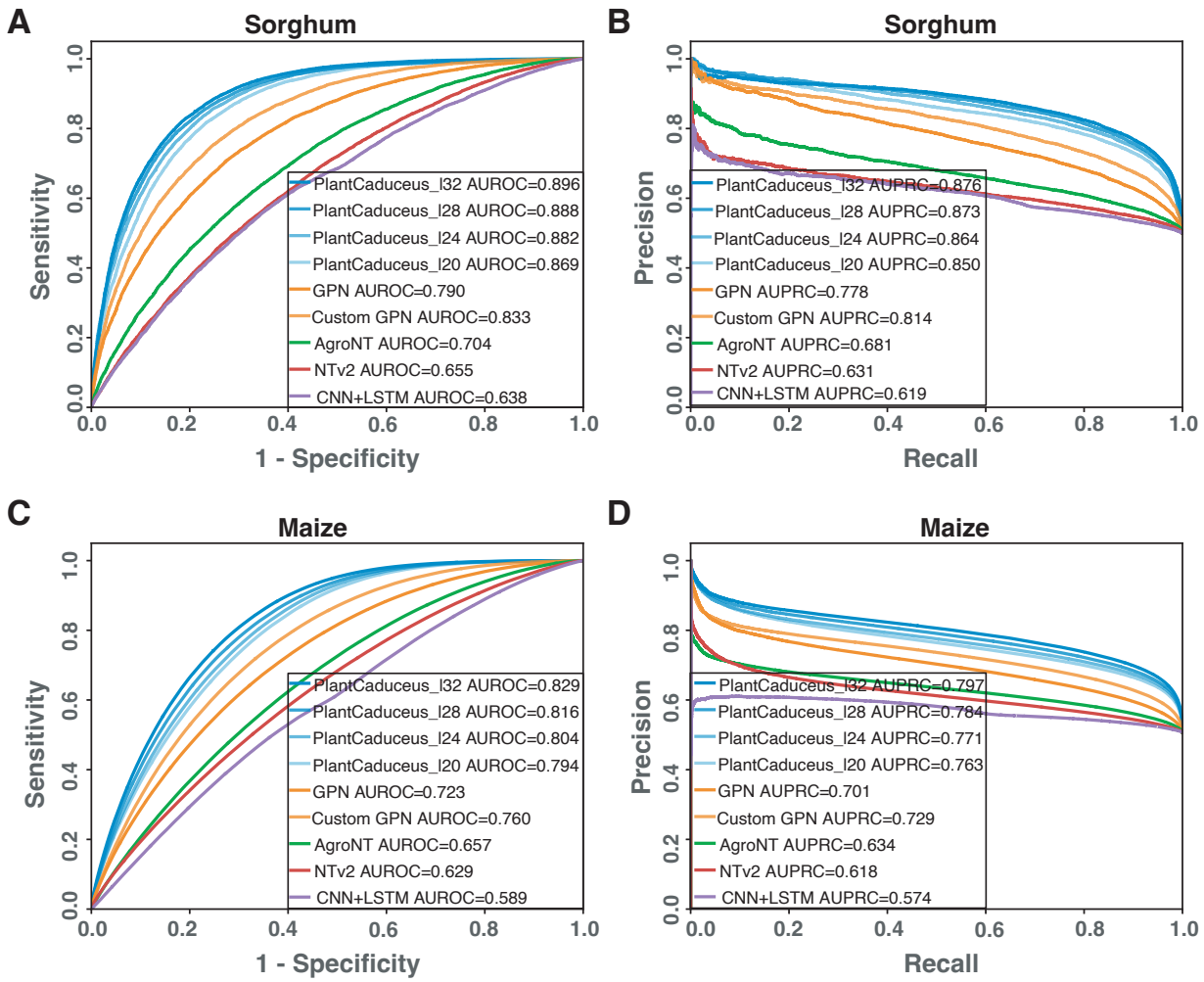
Supplemental Fig. 5. Impact of LoRA fine-tuning on AgroNT's performance in gene annotation tasks. Comparison of prediction performance for **(A)** TIS, **(B)** TTS, **(C)** splice donors, and **(D)** splice acceptors. Bar plots show AUPRC under different training strategies: AgroNT with frozen embeddings (red), AgroNT with LoRA fine-tuning (green), and the largest PlantCaduceus model trained with frozen embedding (blue) serving as a reference benchmark.



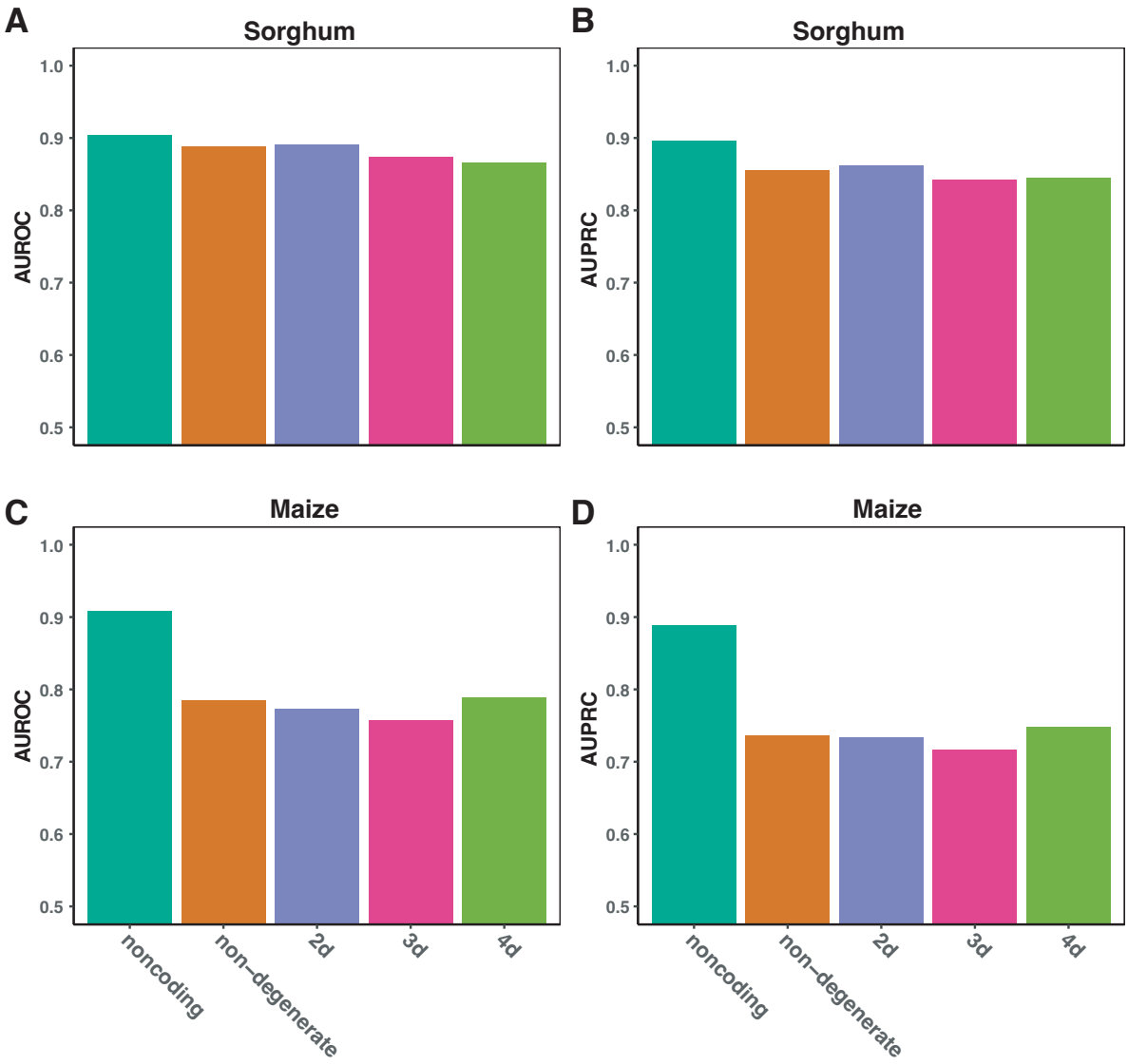
Supplemental Fig. 6. Evolutionary constraint estimation. (A) Illustration of the evolutionary conservation estimation process. **(B)** Distribution of coverage in coding regions. **(C)** Distribution of identity in coding regions.



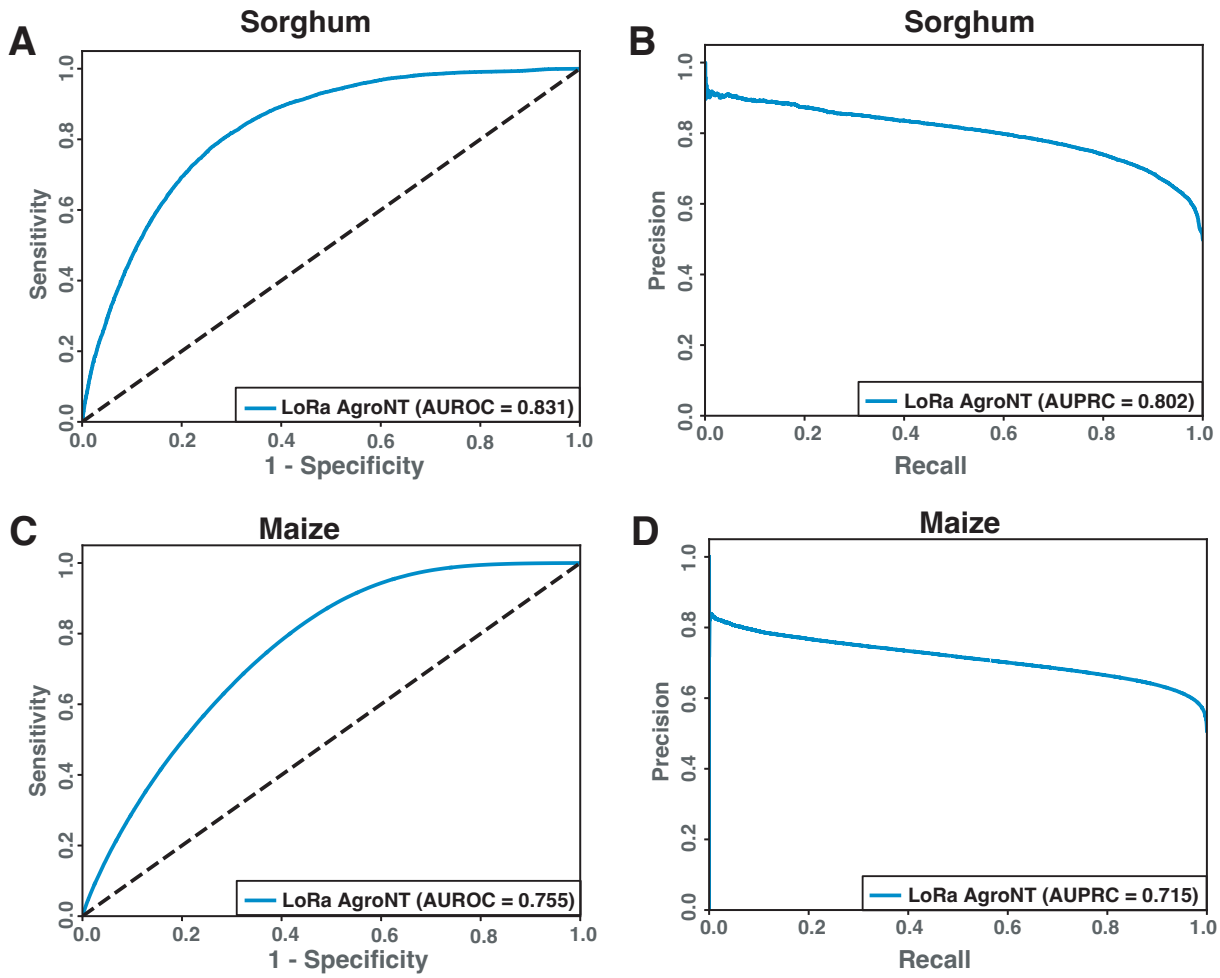
Supplemental Fig. 7. Zero-shot prediction of evolutionary constraints across plant species. Performance evaluation of DNA language models in predicting evolutionary constraints using zero-shot inference based on masked token prediction accuracy. **(A)** Receiver Operating Characteristic (ROC) and **(B)** Precision-Recall (PR) curves showing model performance in Sorghum bicolor. **(C)** ROC and **(D)** PR curves demonstrating cross-species generalization capability when models trained on *S. bicolor* data are applied to unseen *Z. mays*. Model predictions were generated using masked token prediction accuracy scores without task-specific fine-tuning.



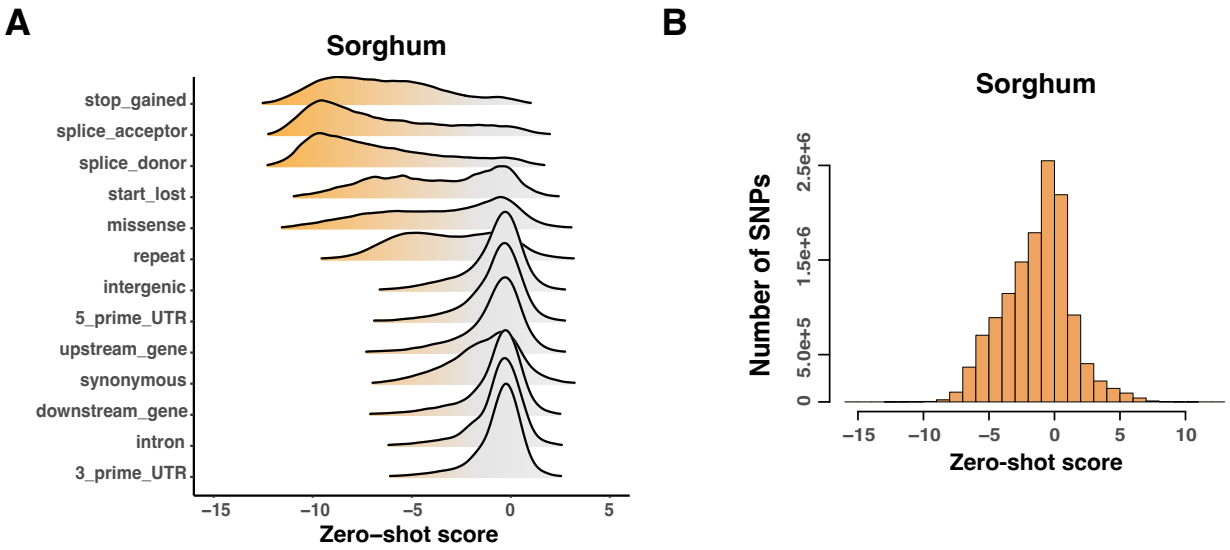
Supplemental Fig. 8. Evolutionary constraint prediction using XGBoost classification. Performance evaluation of DNA language models in predicting evolutionary constraints using XGBoost trained on frozen embeddings. **(A)** Receiver Operating Characteristic (ROC) and **(B)** Precision-Recall (PR) curves showing model performance in sorghum. **(C)** ROC and **(D)** PR curves demonstrating cross-species generalization capability when models trained on S. bicolor data are applied to unseen maize. Model predictions were generated by training an XGBoost classifier on frozen embeddings from different DNA language models.



Supplemental Fig. 9. Evolutionary constraint prediction performance under different conservation context. Performance evaluation of PlantCaduceus_l32 model using XGBoost classification across different conservation contexts and species. **(A, C)** AUROC and **(B, D)** AUPRC for sorghum **(A, B)** and maize **(C, D)**. Five different conservation contexts are compared: noncoding (noncoding sites, green), non-degenerate (non-degenerate sites, brown), 2d (two-fold degenerate sites, purple), 3d (three-fold degenerate sites, pink), and 4d (four-fold degenerate sites, light green). All predictions were generated using XGBoost trained on frozen embeddings from the PlantCaduceus_l32 model.



Supplemental Fig. 10. LoRA fine-tuning evaluation of AgroNT for evolutionary conservation prediction. Performance assessment of LoRA fine-tuned AgroNT model in predicting evolutionary conservation across two species. **(A, C)** Receiver Operating Characteristic (ROC) curves for sorghum and maize, respectively, showing sensitivity versus 1-specificity. **(B, D)** Precision-Recall (PR) curves for sorghum and maize, respectively. Blue solid lines represent model performance, while black dashed lines in ROC curves indicate random prediction baseline.



Supplemental Fig. 11. Deleterious mutations identification in Sorghum. (A) The zero-shot score distribution of different types of variants generated by in silico mutagenesis in sorghum. **(B)** The zero-shot score distribution of 4.6M SNPs in the sorghum TERRA population.

A **Sorghum**

	Top 0.1% deleterious SNPs	Other SNPs
CDS	3,700	443,259
Not CDS	840	4,097,716

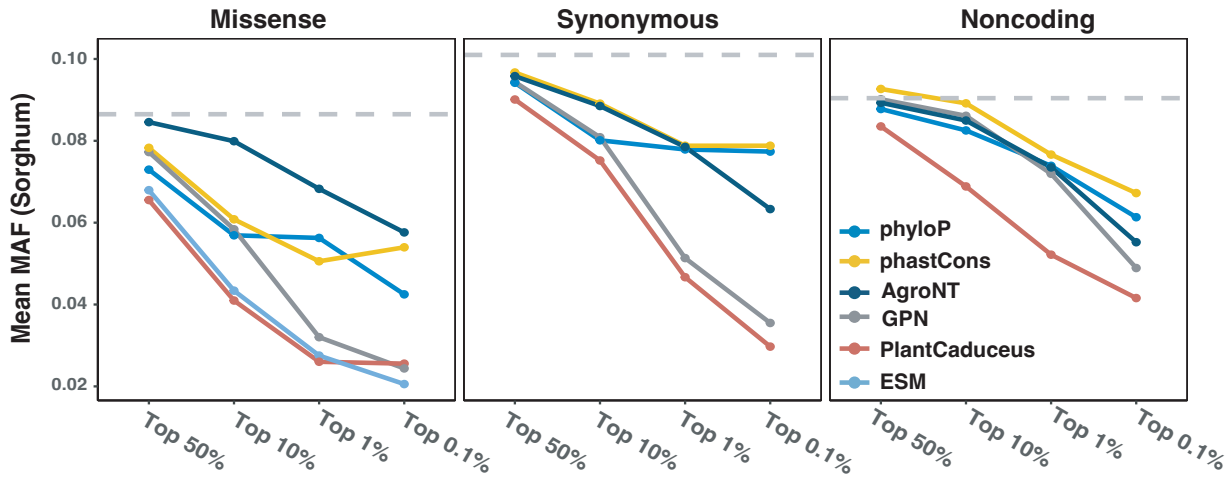
Odds ratio = 40.70; p-value < 2.2e-16

B **Maize**

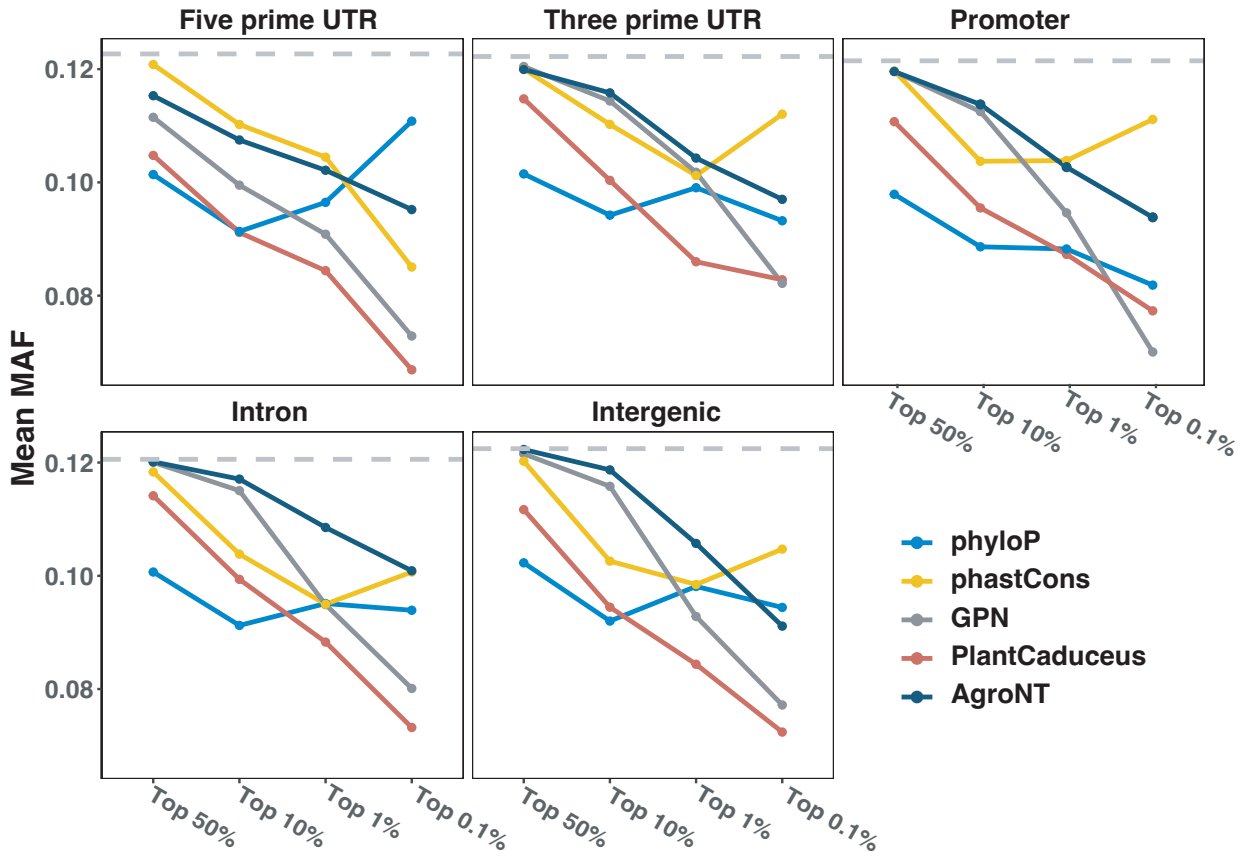
	Top 0.1% deleterious SNPs	Other SNPs
CDS	7,136	658,111
Not CDS	2,222	8,706,658

Odds ratio = 42.42; p-value < 2.2e-16

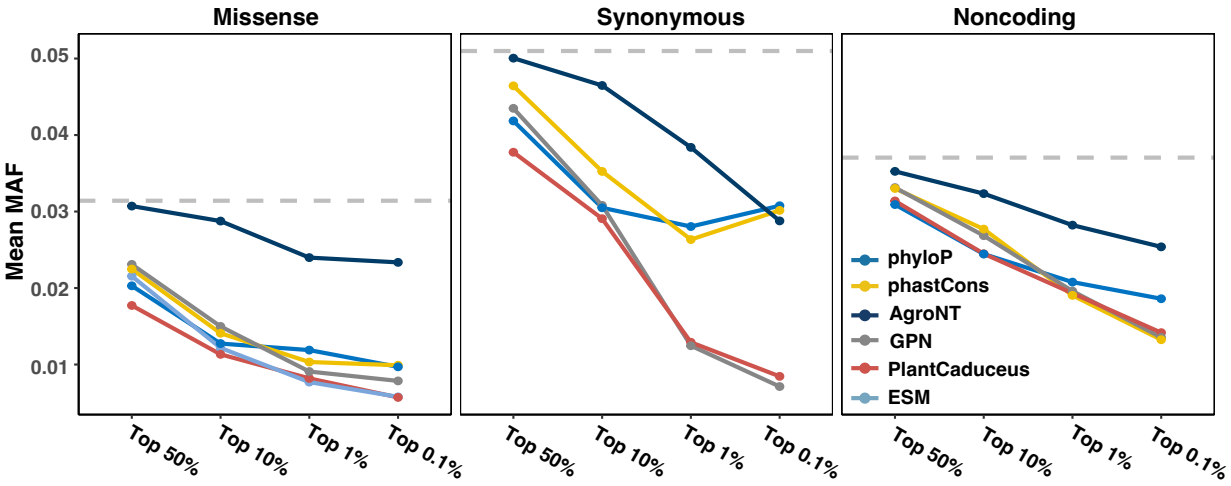
Supplemental Fig. 12. CDS enrichment for predicted deleterious mutations in sorghum and maize. Contingency table and odds ratio showing the enrichment of putative deleterious SNPs in CDS regions for sorghum **(A)** and maize **(B)**.



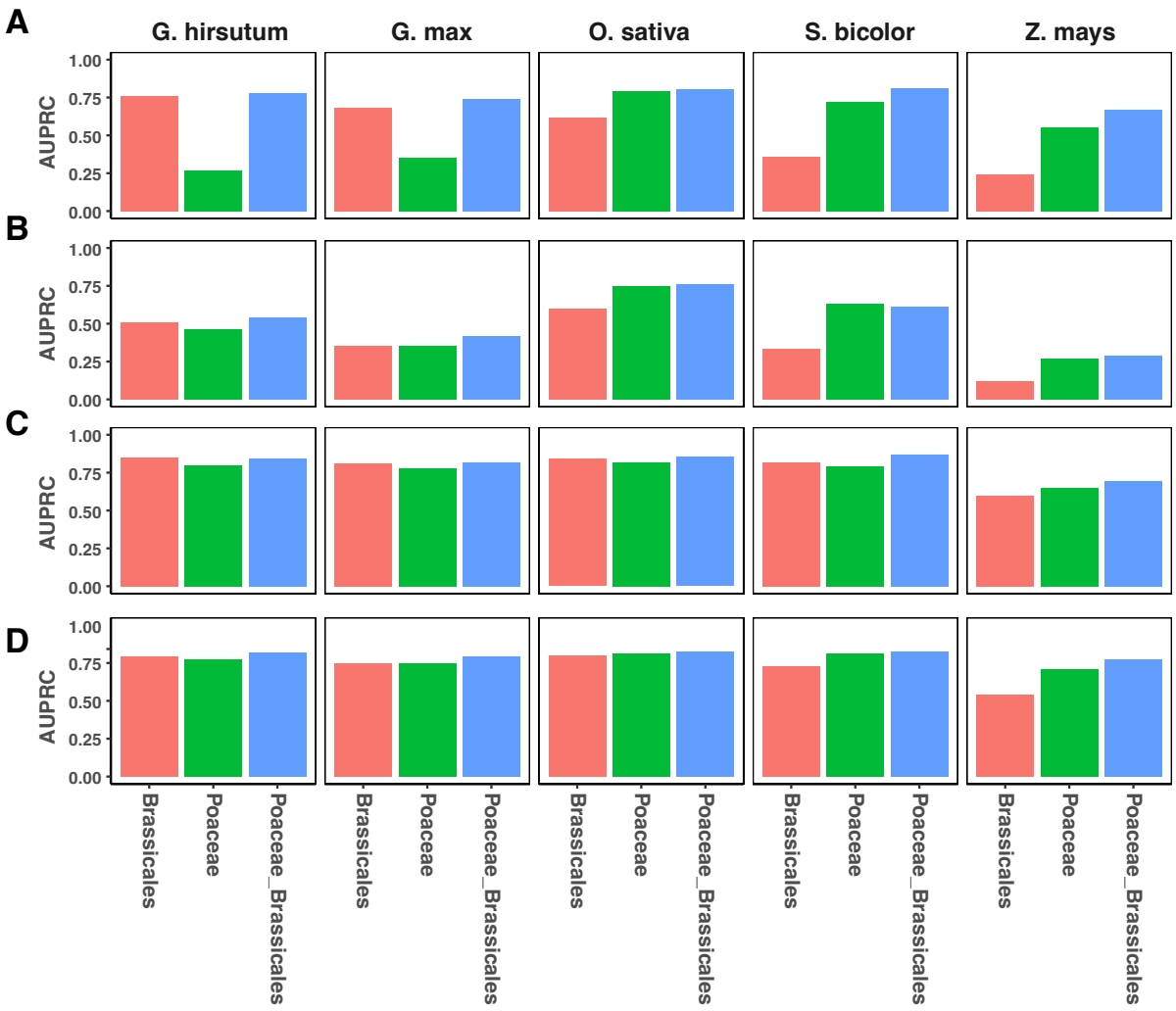
Supplemental Fig 13. The MAF of putative deleterious mutations prioritized by different models in sorghum.



Supplemental Fig 14. The MAF of putative deleterious mutations prioritized by different models in noncoding regions in maize. *Note*: phyloP doesn't have resolution between the top 1% and top 0.1% of sites. Therefore, for the top 0.1% category of phyloP, we randomly selected SNPs from the top 1% set to calculate the mean MAF.



Supplemental Fig. 15. The MAF of putative deleterious mutations prioritized by different models in Arabidopsis.



Supplemental Fig. 16. Impact of species diversity in pre-training data on gene annotation performance. Comparison of prediction performance across five plant species (*G. hirsutum, G. max, O. sativa, S. bicolor,* and *Z. mays*) for four gene annotation tasks: **(A)** Translation Initiation Sites, **(B)** Translation Termination Sites, **(C)** splice donors, and **(D)** splice acceptors. All models use the PlantCaduceus_l20 architecture but differ in their pre-training data: Brassicales (red, trained only on 8 Brassicales species), Poaceae (green, trained only on 8 species from Poaceae family), and Poaceae_Brassicales (blue, trained the full set of 16 species). The y-axis shows AUPRC scores for each model. All predictions are from XGBoost models trained on top of frozen embeddings.