A sequence-based classifier distinguishes phenotype-associated genes from other gene models in plants

Nikee Shrestha, Zhongjie Ji, Xiuru Dai, et al.

Genome Res. published online July 14, 2026
Access the most recent version at doi:10.1101/gr.281802.125

A sequence-based classifier distinguishes phenotype-associated genes from other gene models in plants

Nikee Shrestha[1,2], Zhongjie Ji[1], Xiuru Dai[3], Pinghua Li[3], and James C. Schnable[*,1,2]

[1]Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE, 68503, USA

[2]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, 68583, USA

[3]State Key Laboratory of Crop Biology, College of Agronomic Sciences, Shandong Agricultural University, Tai'an, Shandong, 271018, China

*schnable@unl.edu

ABSTRACT

Only a small fraction of annotated plant genes possess experimentally validated associations with specific phenotypes. Phenotype associated genes have distinct structural, molecular, and evolutionary characteristics compared to non-validated gene models. Here, we develop a simple classifier that uses sequence and evolutionary features which can be generated for any species with an annotated reference genome assembly, to accurately distinguish phenotype associated genes from both the overall population of annotated gene models and a specific set of genes identified as being tolerant of premature stop mutations. A model trained solely on genes from maize (*Zea mays*) identify and prioritize rice (*Oryza sativa*) and *Arabidopsis* (*Arabidopsis thaliana*) genes that are highly enriched in genes with experimentally validated links to phenotypes in both of these evolutionarily distant species. Gene models predicted to have a higher probability of being linked to phenotypes display patterns consistent with known biological properties of phenotype associated genes. Notably, the sets of genes predicted to have

a high probability of being linked to phenotype variation do not consist exclusively of well-characterized gene families but included many uncharacterized gene families carrying domains of unknown function. The quantitative scores generated by this model offer a valuable resource for prioritizing and exploring the vast number of uncharacterized gene models in plants, reducing the risk of failure in future reverse genetics efforts and potentially accelerating gene discovery and functional annotation in crops.

INTRODUCTION

The genomes of typical multicellular eukaryotes contain tens of thousands of gene model annotations. However, few of these genes have been the subject of specific genetic or molecular investigations. Indeed, in *Arabidopsis* less than 10% of genes have been linked to specific phenotypes (Lloyd and Meinke, 2012); likewise, in rice approximately 10% of gene models are associated with phenotypes (Huang et al., 2022), and in maize this is less than 1% (Schnable and Freeling, 2011). The situation in other plants is even worse (Oellrich et al., 2015). This is not unique to plants; even in the human genome many annotated genes have not been the subject of detailed investigation by even a single study (Sinha et al., 2000), including hundreds of genes known to be essential for cell survival (Wang et al., 2015). Research continues to focus on the same well-studied genes (Kustatscher et al., 2022a,b), thereby hindering our understanding of the core functionalities of eukaryotes and the mechanisms responsible for lineage-specific processes and phenotypes.

The process of functionally characterizing less-studied gene models or generating new hypotheses about gene-phenotype associations is time-intensive and high-risk. These efforts can take years and frequently yield inconclusive or null results, creating a very real career risk for researchers. These limitations constrain both reverse genetics approaches and strategies that

leverage natural genetic diversity. Marker trait association signals that fall near already known genes tend to be viewed as confirmatory. In contrast, quantitative trait nucleotides located next to gene models not previously implicated in a phenotype are less pursued, largely because functional validation requires substantial time, resources, and carries a high risk of yielding null results. Similar obstacles arise when interpreting results from differential gene expression analyses, co-expression networks, and other multi-omics approaches. In contrast, once a gene has been shown to influence at least one phenotype, the risk of investigation declines, further amplifying attention toward genes with existing functional evidence while discouraging exploratory work on the vast number of uncharacterized gene models (Kwon et al., 2024). For example, the maize teosinte branched1 (tb1) gene plays an important role in reducing tillering in domesticated maize (Doebley et al., 1990, 1997). The association of tb1 with important domestication phenotypes in maize drove the study of this gene to understand its biological role in maize and in other species (Igartua et al., 2020) with at least 140 additional studies examining *tb1* since the original characterization per MaizeGDB (Woodhouse et al., 2025). In principle characterization of a gene likely also reduces the perceived risk of studying other members of the same gene family, resulting on concentration of investigation not only on specific genes but specific gene families, although the latter effect has not been rigorously evaluated in plants. Being able to predict, in advance, which genes are likely to influence phenotypes would reduce the risk of investigating currently uncharacterized or under characterized genes.

Previous studies have shown that classifiers can distinguish genes associated with different classes of phenotypes. For example, prior work has separated lethal from nonlethal genes in *Arabidopsis*, distinguished essential from nonessential genes across six model eukaryotes with cross-species applicability (Lloyd et al., 2015, Beder et al., 2021), and classified

Arabidopsis genes involved in primary versus specialized metabolism (Moore et al., 2019) or even prediction of gene-to-phenotype associations to prioritize candidate genes for traits (Lee et al., 2010). Features used in these studies typically include both sequence-derived properties and diverse omics data collected across tissues and experiments. Beyond distinguishing biologically meaningful gene sets, previous work has also shown that annotation errors can be identified computationally, including in maize using evolutionary conservation and in *Pisum sativum* using nucleotide and amino acid composition features (Upadhyay et al., 2022; Schulz et al., 2025). It has been shown that genes already associated with phenotypes differ from other gene models across a broad range of characteristics, and these differences appear to be consistent regardless of whether the associated phenotype is lethal, constitutive, conditional, or detectable only at the biochemical or molecular level (Schnable, 2020). This pattern suggests the existence of a distinct class of gene models that are either truly unrelated to organismal phenotype or, at a minimum, unlikely to be linked to phenotype using current experimental approaches. Studying this class of genes is challenging. The set of gene models lacking a known phenotype comprises a mixture of both genes, which could be linked to phenotype if studied, and those truly unrelated to determining the observable phenotypes of an organism. The lack of incentives to publish null results makes it difficult to specifically identify the subset of gene models where researchers have looked for links to phenotypes and found none. In this study, we sought to evaluate how well genes known to be associated with studied phenotypes could be distinguished from the background set of annotated gene models, whether it was possible to distinguish them based solely on features which could be derived from genomic sequence, and whether models trained in a single genetic model species are likely to perform well in other plant species.

Results

We identified 272 maize gene models linked to phenotypic outcomes based on evidence from loss-of-function alleles and named this set as phenotype-associated genes (PAGs) (Supplemental_Dataset_S1). Genes where loss of function alleles have been studied and were not linked to observed phenotypic consequences were poorly documented and it was not possible to identify a large enough set to employ as a negative set for model training. As an alternative approach to identifying genes in which loss-of-function alleles were tolerated, we identified a set of gene models where alleles creating premature stop codons were widespread using whole-genome resequencing data from 1,515 maize genotypes that include open pollinated varieties and wild relatives. This resulted in a second set of 713 gene models with variants causing a premature stop codon, i.e., premature stop enriched genes (PSEs). There was no overlap between the set of PAGs and the set of PSEs ($E = 5$ gene models, p-value = 0.017; Fisher's exact test). In many cases, the alleles that produced premature stop codons were relatively rare, defined as premature stop enriched-rare (PSE-R) genes, but in 162 cases, the frequency of the allele that resulted in a premature stop codon was greater than 0.25 that excludes across the re-sequenced population used to discover these variants (Figure 1A). We focused on these 162 genes that carry common premature stop codon variants in downstream analyses, as rarer premature stop codons may be more likely to represent alleles which would impact fitness when homozygous but are still represented in our population datasets via data from open pollinated varieties or wild relatives. These gene models were supported by transcriptional data and, frequently, by proteomic and comparative genomic evidence, and thus do not simply represent artifacts produced by ab initio gene annotation software. Below, we refer to this set of gene models as premature stop enriched-common (PSE-C) genes. Both PAGs and PSE-Cs were depleted in the category of gene models lacking evidence of expression in any tissue. Of the 1,215 gene models

meeting these criteria, 3 were PAGs (8 expected), and zero were PSE-Cs (5 expected).We compared PAGs and PSE-Cs with common premature stop codon alleles across 2,604 features including amino acid and nucleotide composition (as well as dinucleotide frequency, etc), structural characterized of gene models (e.g. sequence length, number of exons), predicted protein functional characteristics (e.g. hydrophobicity), gene expression and protein abundance data, and features derived from chromatin, co-expression, sequence variation, or transcription factor binding datasets (See Methods, Supplemental_Dataset_S2). Both PAGs and PSE-Cs were less common among non-expressed genes than expressed genes (3 non-expressed PAGs, 8 expected; 0 non-expressed PSE-Cs, 5 expected). PAGs and PSE-Cs exhibited statistically significant differences for about 44% of the time (p-value < 0.05, two-tailed Mann-Whitney *U* test, Supplemental_Dataset_S2), with 16% of these differences exceeded a more stringent Bonferroni corrected p-value (p-value < 0.05/2,604 tested features). PAGs and PSE-Cs exhibited significant differences in gene length, average protein abundance, the number of tissues expressing a gene model (Supplemental Figure S1), and syntenic conservation in sorghum (p-value<0.001, $2.2 * 10^{-16}$; Fisher's exact test). These results are consistent with previously reported differences between genes associated with mutant phenotypes and the overall set of annotated gene models (Schnable, 2020) and also between lethal and non-lethal phenotype associated *Arabidopsis* genes (Lloyd et al., 2015). No significant difference was observed between the number of exons in PAGs and PSE-Cs, another trait previously reported to differ between genes associated with mutant phenotypes and other annotated gene models (p-value=0.53, two-tailed Mann-Whitney *U* test, Figure 1C). PAGs had significantly longer coding sequences than PSE-Cs, inconsistent with the possibility that the enrichment of premature stop codons resulting from a larger number of potential sites where a mutation could introduce a stop

codon. Many other features where PAGs and PSE-Cs significantly differed had not previously been linked to the likelihood of being associated with organismal phenotypes. These include the length of the 3' untranslated region (UTR) (PAGs > PSE-Cs p-value < 0.001 [$8.13 * 10^{-15}$], two-tailed Mann-Whitney *U* test, Cohen's d=0.64,Figure 1D), number of isoforms (PAGs > PSE-Cs p-value= 0.003, two-tailed Mann-Whitney *U* test, Cohen's d=0.23, Figure 1E), and GC content in the nucleotide sequence of the gene model (PAGs > PSE-Cs p-value < 0.001 [$6.70 * 10^{-5}$], two-tailed Mann-Whitney *U* test, Cohen's d=0.53, Figure 1F). Maize genes exhibit a bimodal distribution of GC content (Alexandrove et al., 2009). The distributions for both PAGs and PSE-C genes both reflected the same bimodal distribution observed for maize genes as a whole, although PSE-Cs included a higher ratio of the lower GC content mode relative to n PAGs (Supplemental Figure S2). PAGs appeared to be more distinct from the overall population of annotated gene models than PSE-Cs. Significant differences between PAGs and non-PAG, non-PSE-C genes were detected for 54% of quantitative features (mean Cohen's d = 0.05; median Cohen's d = 0.01), with 19% remaining significant after Bonferroni correction (p-value < 0.05/2,604 tested features). By contrast, significant differences between PSE-Cs and non-PAG, non-PSE-C genes were detected for only 28% of quantitative features (mean Cohen's d = 0.008; median Cohen's d = 0.002), with only 3% remaining significant after Bonferroni correction (p-value < 0.05/2,604 tested features) (Supplemental_Dataset_S2). Collectively, these results demonstrate that PSE-Cs are more similar to the rest of the gene models than PAGs, and that PAGs are significantly different from both PSE-Cs and the rest of the average gene models for both previously tested and untested features.

Despite the significant differences between PAGs and PSE-Cs for many individual features, these two populations of genes exhibited only modest separation on the first principal

component calculated from the full set of gene features (Figure 1B). This indicates that there may be multiple distinct patterns of genes whose functions are likely (or unlikely) to be associated with phenotypes. This result also contrasts with the previously reported clear separation between phenotype-associated genes and other annotated gene models when using a smaller, curated feature set (Schnable, 2020). We trained a random forest model, which tends to be more robust to nonlinear relationships, to distinguish PAGs and PSE-Cs using all 2,675 gene features (2,604 continuous or numeric features + 71 binary features). Model performance was evaluated using receiver operating characteristic (ROC) curves and the area under the curve (AUC) using a holdout dataset of 48 gene models (PAGs=29, PSE-Cs=19). ROC curves visually summarize the trade-off between true and false positive rates across different thresholds, while the ROC-AUC provides a single numeric metric which quantifies performance across all thresholds. The model exhibited strong predictive performance achieving an ROC-AUC of 0.87 on the hold-out test gene sets (Figure 2A). The prediction probabilities assigned to PAGs (mean=0.89) and PSE-Cs (mean=0.17) were each significantly different from those of uncharacterized annotated gene models (mean=0.50, p-value<0.001 for both comparisons, two-tailed Mann-Whitney *U* test, Supplemental_Dataset_S3). One third of uncharacterized maize gene models were assigned lower scores than any of the genes in our dataset already linked to phenotypes, implying that a substantial fraction of annotated gene models representing sequences unlikely to produce publishable or easy to study phenotypes (Figure2B). Above this threshold, many additional annotated genes were assigned scores lower than the highest scoring PSE-C. Both of these results are consistent with a substantial fraction of annotated gene models being unlikely to produce publishable or easy to study phenotypes when disrupted. The ability of a

random forest model to achieve strong predictive performance on hold-out test sets of PSE-Cs and PAGs supports the substantial differentiation of these populations of gene models.

We next sought to determine how many, and which types of features were necessary to distinguish these populations of genes. The set of features describing maize gene models can be separated into genomic-sequence-derived and non-genomic sequence derived features. Genomic-sequence-derived features can be directly extracted or calculated, given only an assembled genome sequence and a structural gene annotation file. By contrast, non-genomic sequence derived features, including gene expression data, protein abundance, chromatin accessibility profiling, and assays of transcription factor binding, require additional experiments and are typically challenging to generate similarly across experiments conducted in multiple species. Among the 100 features with the highest Gini importance in the trained PAG/PSE-C classification model, 12 were derived from non genomic sequence or expression data. Gini importance quantifies how much each feature contributes to reducing classification error across the decision trees in the model. Only one of the top 10 most important features was not derived from genomic sequence, namely the maximum Fragments Per Kilobase of transcript per Million mapped reads (FPKM) observed for a gene across all tissue samples (Supplemental_Dataset_S4). Syntenic conservation of a gene model in sorghum appeared to be the most important feature (Supplemental_Dataset_S4). The top 100 features also included gene structural attributes, particularly 5′ and 3′ untranslated region lengths, as well as numerous protein descriptors reflecting amino acid composition, residue ordering, physicochemical properties, and local sequence architecture. Other prominent features included *k*-mer frequencies, frequency of occurrence of each amino acids (codon usage), the number of times a codon appears in a gene divided by the number of expected occurrences under equal codon usage (relative synonymous

codon usage), and GC content at the third codon position consistent with differences between PAGs and PSE-Cs (Figure 1F). A random forest model trained solely on genomic and comparative genomic derived features achieved comparable performance, with an ROC curve area of 0.86 (Figure 2A, B). Model performance remained consistent with an ROC curve area of 0.87 when the feature set was restricted to only 100 genomic sequence derived features with the highest Gini importance (Figure 2A, B). Consistent with previous models, synteny remained the most important feature. The probability of a gene in the top 10% (threshold=0.88, n=3,904) of predicted scores generated by the top hundred-feature genomic sequence derived data model being linked to a phenotype was estimated to be approximately 18.3 times higher than that of a gene with a score in the bottom 50% (threshold=0.88, n=19,516) based on evaluation of the hold-out test data. These results suggest that it may be feasible to build high-performing models based solely on feature sets that can be generated consistently across different species.

The model trained solely on genome sequence and comparative genome derived features was able to separate the prediction probability distribution of premature stop enriched-common (PSE-C) maize gene models from phenotype associated genes in maize, with uncharacterized gene models having overlapping distribution with both extreme classes (Figure 3A). We next tested how well our model translates to other plant species. The genetic characterization of several plant species, notably rice and *Arabidopsis,* also includes information on the phenotypic consequences of disrupting comparable or greater numbers of genes compared to the maize dataset used in this study. However, in most plant species, with sequenced genomes, few or no genes have been directly linked to mutant phenotypes. We assessed the transferability of a model trained on maize data to other species using a model trained on 99 of the top 100 maize features, excluding one feature that could not be easily generated for rice gene models (see Methods), to

predict the probability that annotated rice genes would be associated with phenotypic consequences. From the set of approximately 4,500 cloned rice genes in the funRiceGenes database (Huang et al., 2022), we manually curated a non-exhaustive set of 100 which met the same phenotype criteria applied to maize. The 96 rice PAGs which could be mapped to MSU v7 gene models included 72 which were conserved at syntenic locations in maize and 24 which lacked maize syntenic orthologs. However, in only seven cases was the maize syntenic ortholog of a rice PAG also classified as a PAG in maize. The 96 rice PAGs were assigned significantly higher probabilities of being associated with a phenotype than with the background set of annotated gene models in the rice genome (Figure 3B, p-value < 0.001 [$5.27 * 10^{-19}$], two-tailed Mann-Whitney *U* test, Supplemental_Dataset_S5). The probability of a rice gene in the top 10% (threshold=0.63, n=5,564) of scores being linked to a phenotype was eleven times higher than that of a gene in the bottom 50% (threshold=0.45, n=27,015). The divergence of the lineages leading to maize and rice is estimated to have occurred approximately 50 million years ago, roughly one-third the time since the estimated age of the most recent common ancestor of maize and *Arabidopsis* (~160 million years ago) (Kumar et al., 2017). The *Arabidopsis* genome is also less than 10% the size of the maize genome, with substantially lower transposon content. A model trained solely on maize data using the same 99 features assigned significantly higher probabilities of being associated with a phenotype to a set of *Arabidopsis* PAGs than to the background set of annotated gene models in the *Arabidopsis* genome (Figure 3C, p-value < 0.001 [$1.12 * 10^{-176}$], two-tailed Mann-Whitney *U* test, Supplemental_Dataset_S5). The prediction probability was not different between essential, morphological, cellular and biochemical *Arabidopsis* PAGs and conditional *Arabidopsis* PAGs (p-value=0.77, two-tailed Mann-Whitney *U* test, Supplemental Figure S3). The probability of an *Arabidopsis* gene model

in the top 10% (threshold=0.52, n=2,806) of scores being linked to a phenotype was four times higher than a gene in the bottom 50% (threshold=0.41, n=13,227). These results confirm the notion that a model trained solely on maize data can still achieve both statistically significant and operationally meaningful performance in prioritizing genes for phenotypic characterization when applied to a distantly related angiosperm with a radically different genome architecture.

Cross-validation and testing on a randomly selected set of hold-out test data are both widely used approaches to estimate the performance of machine learning models. However, the accuracy of machine learning models can sometimes decline when evaluated on truly out of sample datasets relative to estimates based on held out subsets of the same dataset used to train the initial model (Guan et al., 2025; Lei et al., 2025). We generated a new, curated set of maize PAGs from literature published after our original PAG dataset was finalized (Supplemental Data Set S7). The estimated probabilities of being being linked to phenotype of these new out-of-sample PAGs (n=27) were significantly higher than those of a typical uncharacterized gene (p-value=0.002, two-tailed Mann-Whitney *U* test, Figure 3D). The average estimated probability assigned to the out of sample PAG genes was modestly lower than the probabilities assigned to PAGs from the original hold-out test dataset (μ=0.68 vs μ=0.79). However, the estimated probability that a gene in the top 10% of the prediction scores is associated with a phenotype was still approximately 13 times as high as the probability for a gene in the bottom 50% of scores. This ratio was estimated to be 18.3 times using the original hold-out test genes. These results demonstrate that the prediction model retains strong predictive performance even when applied to truly out-of-sample datasets.

The gene models predicted to be linked to phenotypes are supported by other data types associated with purifying selection or the consequences of mutations. Gene models in the bottom

quartile (threshold=0.27, n=9,759) of probability of being associated with a phenotype exhibited approximately 4.4 times higher odds of containing a premature stop codon compared to those in the top quartile (threshold=0.76, n=9,759, p-value < $2.2 * 10^{-16}$; Fisher's exact test) (Figure 4A, B). Variants introducing premature stop codons in bottom-quartile gene models were observed at higher minor allele frequencies than variants introducing premature stop codons in top-quartile genes (p-value=0.01 two-tailed Mann-Whitney *U* test). Based on alignments to orthologous genes in sorghum, bottom-quartile genes exhibited higher rates of sequence evolution at both the DNA sequence and protein sequence level with elevated rates of synonymous substitutions (median Ks, 0.53 vs 0.28, p-value < 0.001), elevated non-synonymous substitutions (median Ka, 0.19 vs 0.03, p-value < 0.001) and elevated K_a/K_s (median K_a/K_s, 0.14 vs 0.20, p-value < 0.001), consistent with relaxed purifying selection or a mixture of purifying and positive selection. Bottom-quartile gene models showed higher CG and CHG methylation levels across annotated gene regions than top-quartile gene models. In contrast, CHH methylation across annotated gene regions was higher in top-quartile gene models than in bottom-quartile gene models, consistent with the bottom-quartile set not being primarily composed of transposon fragments. Variant frequency per kilobase of coding sequence (CDS) was higher in top-quartile gene models than bottom quartile gene models, but this effect was driven by a higher prevalence of variants predicted to have small (low) effects on protein function, typically synonymous variants, and the proportion of variants predicted to have moderate (p-value < 0.001, two-tailed Mann-Whitney *U* test) or strong (high) effects on protein function (p-value < 0.001, two-tailed Mann-Whitney *U* test) were significantly higher in bottom-quartile gene models (Figure 4C, D). The apparent inconsistency between higher Ka and Ks ratios both being observed in bottom quartile genes and a higher overall variant frequency being observed in top quartile genes may have resulted from

the subset of genes which lacked syntenic orthologs in sorghum. These genes were still included in estimates of variant frequency but it was not possible to estimate substitution rates for these genes. A similar pattern may also arise because bottom-quartile genes were more likely to show presence-absence variation. When a gene is present in some maize lines but absent in others, SNP and InDel calls within that gene tend to have higher levels of missing data and are therefore more likely to be removed during quality filtering. We employed PlantCaduceus, a DNA language model that has learned evolutionary conservation patterns in angiosperm genomes (Zhai et al., 2025) to estimate the effects of DNA substitutions without the need to align against genes in related species. Genetic variants in the gene body or CDS of top-quartile gene models exhibited significantly higher predicted likelihoods of being deleterious (based on average absolute zero-shot scores from PlantCaduceus) than bottom-quartile gene models (p-value < $1 * 10^{-200}$, two-tailed Mann-Whitney *U* test, Figure 4E). By contrast, variants present in UTRs or introns exhibited a small but significant bias in the opposite direction, with variants in these regions of bottom-quartile genes assigned marginally higher absolute zero-shot scores by PlantCaduceus than those in top-quartile genes (p-value < $1 * 10^{-20}$, two-tailed Mann-Whitney *U* test, Figure 4E).

To assess whether the model was prioritizing members of well-annotated gene families, we examined Gene Ontology (GO) term enrichment analysis among genes from top quartile and bottom quartile of the distribution, constraining GO enrichment analysis to the population of only those genes assigned at least one GO term. An average top quartile gene models were assigned 17 GO term per gene model and bottom quartile gene models were assigned 6 GO terms per gene model (minimum GO term assigned to a gene model: 1, maximum GO term assigned to a gene model: 297, Supplementary Figure S4). GO enrichment analysis showed that

top-quartile genes were enriched in 424 GO terms (out of 7,277 GO terms tests) and bottom-quartile genes were enriched in 37 GO terms (Supplemental_Dataset_S6). This result raised the concern that the model's strong performance might partly reflect memorization of the characteristics of well-annotated gene families. Ninety-six percent of top-quartile genes and forty-seven percent of bottom-quartile genes were assigned at least one GO term. However, 157 genes in the top decile of predicted probabilities of being phenotype-associated encode proteins with domains of unknown function. Manual curation of the top 15 uncharacterized gene models showed that four genes had only domains of unknown function. These results indicate that the model captured biologically important patterns to distinguish gene models highly likely to be associated with phenotype, and that genes predicted to be associated with phenotypes extend beyond those of well-characterized gene families.

DISCUSSION

A substantial proportion of gene models remain uncharacterized (Pena-Castillo and Hughes, 2007; Schnable and Freeling, 2011; Stoeger et al., 2018). Importantly, this lack of functional characterization is not limited to individual gene models, as some entire gene families also remain poorly understood. The time and resources required to functionally characterize an individual gene remain costly. As a result, the social structure of research and funding allocation decisions incentivize focusing on well-studied genes. Therefore, more than three-fourths of gene models across humans and plants are understudied and have not been functionally characterized (Schnable, 2020; Stoeger et al., 2018). Although the initial impulse may be to assume that all currently uncharacterized gene models are equally good candidates for characterization, here, we demonstrated that it is possible to train models on relatively simple and generalizable sets of features thatincrease the odds of successful reverse genetics characterization. We curated an

extended list of phenotype-associated genes identified in maize from the 1900s to November 27, 2021, and identified a complementary set of premature stop enriched-common gene models that can serve as a negative set for model training. All of the genes in this set had evidence of transcription (e.g. expression in RNA-seq studies), many had evidence of translation from proteomic investigations, and more than one-third were conserved at syntenic locations in the sorghum genome, indicating these loss-of-function tolerant genes did not simply represent artifacts produced by *ab initio* gene annotation software. Consistent with this, even after restricting the analysis to gene models classified as true genes by reelGeneScore (threshold > 0.5, removing more than one-quarter of maize gene models; Schulz et al., 2025), the evolutionarily informed model still distinguished PAGs with an ROC-AUC of 0.78. This is a decline relative to the ROC-AUC of 0.87 achieved when all annotated genes were included, but still much higher than the 0.5 expected of an uninformative model. Together, these findings support that the predictions generated by the model represent biologically meaningful differences rather than solely the identification of annotation artifacts. The difference between PAGs and PSE-Cs could not be explained by difference in gene family size to which these gene models belong to (p-value=0.13, two-tailed Mann-Whitney *U* test). This was previously found to be an important difference between genes associated with lethal phenotypes and genes associated with non-lethal phenotypes, where presence of even a single paralog reduces the likelihood of lethality under disruption of gene (Lloyd et al., 2015). A simple classifier model provided with evolutionary information and genomic sequence-derived features was able to classify gene models in maize into the correct classes with an ROC-AUC of 0.87.

The top quartile of genes predicted to be linked to phenotypes included more than 20% of all maize gene models encoding proteins containing domains of unknown function, more than

300 genes in the top quartile lacked GO annotations, and more than 200 genes lacked any functional description at all, including automated domain annotations. This suggests that our model is not simply learning the properties of well-characterized gene families as uncharacterized genes and members of uncharacterized gene families were not excluded from those predicted to be associated with phenotypes. Syntenic conservation in a closely related species (specifically sorghum for the maize model) consistently ranked as the most important across all trained models. However, when nonsyntenic genes were excluded from both the training and testing set (removing the information content of synteny), a retrained top-100-feature model was able to classify maize gene models with an ROC-AUC of 0.85 compared with 0.87 for the original model, indicating that although synteny is an important predictor, its inclusion is not essential to achieve substantial prediction performance.

Phenotype-associated genes were defined based on reports in the literature describing visible phenotypes in relatively easy-to-characterize tissues. This definition necessarily biased our training dataset against genes whose phenotypes are most apparent in tissues or organs that are more difficult to observe, such as roots. Our initial maize PAG set was also enriched for genes associated with easily characterized kernel phenotypes. Of the 27 newly validated gene models, nearly half were kernel mutant phenotypes. However, kernel-related and non-kernel-related genes showed no significant difference in mean predicted probability of being phenotype-associated (p-value=0.39, two-tailed Mann-Whitney *U* test), and in fact the newly characterized genes with non-kernel related phenotypes were assigned higher mean probabilities by the model (mean=0.63 vs mean = 0.72) suggesting the high representation of these genes did not bias the model towards genes with phenotypes observed specifically in kernels. Our criteria for the training dataset also specifically excluded genes linked to phenotypes only under specific

environmental conditions (e.g. stress specific phenotypes). However, *Arabidopsis* genes associated with conditional phenotypes were not assigned significantly lower predicted probabilities of being phenotype-associated than *Arabidopsis* genes associated with constitutive phenotypes (Supplemental Figure S3).

We were able to assess overfitting or species-specific prediction by testing models trained in maize on known phenotype-associated genes from other species. Models trained in maize generalized well across species, assigning substantially higher predicted phenotype probabilities to genes with known mutant phenotypes in both rice and *Arabidopsis*, two distantly related species with large differences in genome size and gene content. Using a top-decile threshold yields very low false positive rates but misses many true phenotype-associated genes, indicating a substantial false negative rate. In the absence of confirmed true negative gene set, predicted probabilities should be interpreted as relative prioritization scores rather than absolute probabilities of phenotype association. It should also be noted that the probability threshold corresponding to any given percentile of scores varied across the species we evaluated. Application of the model may be more comparable when applying similar percentile based threshold than comparable absolute scores. Selection of specific threshold would necessarily be dependent on a user's relative tolerance for false positives and false negatives. As additional genes continue to be cloned and functionally characterized, it will be informative to reassess model performance and determine how well the current patterns hold in larger and more diverse validation sets.

One potential application of our classifier is predicting the likelihood that disrupting a given gene will be associated with phenotypic changes. Association studies with genetic markers narrow down the genomic region associated with a given trait of interest. There are cases where

the causal gene is the nearest one to the associated genetic marker (Sahay et al., 2023), especially when there is a high-density genetic marker across the genome and strong linkage disequilibrium decay (Grzybowski et al., 2023). However, for many plants, markers may be at low density, or there may be strong linkage disequilibrium, or in some cases, a combination of both. Having a complementary tool that provides a quantitative score for each gene model indicating the probability that a gene model is linked to a phenotype could guide researchers in prioritizing candidate genes in such association studies. Such a framework could help reduce uncertainty when large genomic intervals contain dozens or even hundreds of annotated gene models. In addition, this tool could provide complementary evidence to traditional QTL mapping or GWAS approaches, ultimately accelerating functional validation. The vast majority of gene models, even in the most studied plants such as maize, rice, and *Arabidopsis*, are yet to be characterized, let alone many orphan crops, which are equally, if not more important (Shrestha et al., 2023). Editing, validating, and characterizing gene models takes a long time and is costly (Stoeger et al., 2018). Quantitative scores for gene models that could be used as confidence scores for prioritizing targets would therefore be invaluable. Such scores could help researchers focus their limited resources on the most promising candidates, thereby accelerating basic discovery and crop improvement. Identifying true genes based on high quantitative scores could mark the starting point for optimizing their functional roles across diverse crop species.

METHODS

Definitions of Phenotype-Associated and Premature Stop Enriched gene sets

The initial set of 272 maize genes whose loss of functional alleles have been experimentally shown to produce a detectable change as an easily observable phenotype, along with the organ affected by each gene, was defined through our manual curation of a larger set of

603 maize locus records annotated as phenotype-associated and retrieved from MaizeGDB (Sen et al., 2023) on November 27, 2021. This was done to exclude loci linked to purely molecular phenotypes (e.g., variation in gene expression) or when the evidence for a genotype-phenotype link came from studies using techniques with higher false-positive rates (e.g., association studies). From the set of approximately 4,500 cloned rice genes in the funRiceGenes database, we manually curated a non-exhaustive set of 100 which met the same phenotype criteria applied to maize (Huang et al., 2022). In four cases, a gene identified in the analysis did not correspond to a gene in the sequence-derived feature set calculated from the MSU v7 rice genome annotations, leaving 96 phenotype-associated rice genes. The set of phenotype-associated genes in *Arabidopsis* was retrieved from previously published list (Lloyd and Meinke, 2012). Excluding gene models in the Lloyd and Meinke list with phenotypes that were classified as 'NC; non-confirmed' resulted in a set of 2,018 phenotype-associated genes for *Arabidopsis*. Out-of-sample maize phenotype-associated genes were identified using a set of 2,582 citations (through April 14, 2025) describing maize loci that had not been associated with a phenotype in our original November 27, 2021, dataset. These citations were derived from 682 unique papers, and screening these publications identified an additional 27 loci that met the criteria for publishing a new genotype-to-phenotype association supported by loss-of-function data.

The maize premature stop enriched gene set was defined using a published set of genotype calls derived from whole-genome resequencing of 1,515 maize varieties (Grzybowski et al., 2023, Supplemental_Dataset_S8). The functional consequence of each of the approximately 46 million biallelic genetic variants (SNPs and InDels) identified within this study were annotated using SNPeff v5.1d (De Baets et al., 2012) and the published gene model annotations of the B73_RefGen_V5 reference genome (Hufford et al., 2021). The primary

transcript of each annotated gene model containing a genetic variant that produces a premature stop codon and is present at an alternate allele frequency greater than 0.25 across the population was considered premature stop enriched-common, resulting in a final dataset of 162 premature stop enriched-common maize genes.

Definition of the feature dataset describing maize gene models

A dataset of 14,418 different features describing 39,756 annotated gene models in the B73_RefGen_V5 reference genome was downloaded from the Maize Feature Store in MaizeGDB (Sen et al., 2023, https://mfs.maizegdb.org/featurebase). Gene models that were not assigned to any of the ten maize chromosomes (i.e., on unanchored scaffolds) were dropped from the dataset, reducing the total number of gene models to 39,034. Redundant features introduced by the presence of both raw and normalized values in the dataset, features with missing values for more than 1 gene model, annotation of gene model in sorghum version 3.1 annotation and in Tzi8, chromosome assignment, and four subcellular localization features while retaining only one subcellular localization feature, Subcel1 were removed from the dataset. In addition, features describing tri-peptide amino acid composition, which would otherwise have constituted approximately 55% of the total feature set, were also dropped from the dataset. Gene expression and protein abundance values of "none" were converted to "0" prior to analysis. After excluding these features, 2,639 remained, including features describing gene structure, codon usage, nucleotide and protein composition, and non-genomic sequence derived data regarding gene expression atlas, protein expression, chromatin-related features, expression localization, and transcription factor binding sites associated with each gene model in maize. A single non-MaizeGDB-sourced feature, syntenic conservation in sorghum, scored using the dataset published in (Dias et al., 2026), was added, resulting in a final dataset of 2,640 unique features.

Synteny based on sorghum gene models included a filtering step that removed members of tandem duplicate arrays that did not possess reciprocal best BLAST hit relationships between maize and sorghum. The feature generation process and definition of features have been described in detail by Sen et al. (2023). The final feature set of 2,640 features included 31 binary features, 5 categorical features, and 2,604 continuous and discrete features. A total of 350 features were non genomic sequence derived, while the remaining 2,290 features could be directly computed from the maize genome assembly and annotation files.

For use in machine learning applications, categorical features were one-hot encoded, increasing the effective number of variables to 2,675 (Supplemental_Dataset_S9), and continuous and discrete features were min-max normalized across all gene models using the MinMaxScaler function in Scikit-learn v1.3.0 implemented in Python v3.11.4 (Pedregosa et al., 2011).

Random forest classifier model for prediction

Prior to developing or conducting any analysis, approximately 10% of the gene models were set aside as a separate testing dataset. The data were split into training and testing datasets using a previously proposed family-guided strategy (Washburn et al., 2019) and gene family assignments retrieved from PLAZA (Van Bel et al., 2022) to minimize the risk of data leakage. The testing data set included 29 phenotype-associated genes and 19 premature stop enriched-common genes. For each model trained in this study, the remaining 386 labeled gene models were used to optimize the hyperparameters and train the final models. The number of trees in a model, number of features for splitting each tree, maximum depth of the tree, minimum number of samples required to split a node, minimum number of samples required to be at a leaf node, and whether bootstrap samples were used when building trees were optimized via three-fold

cross-validation of the training dataset using the RandomizedSearchCV function in Scikit-learn v1.3.0 implemented in Python (Pedregosa et al., 2011). The optimal set of hyperparameters identified through this process for each model was then used to train the respective final model using the whole training data. ROC curves were plotted based on the model's performance on the hold-out test dataset, and the AUC was used to evaluate the model's performance.

Feature importance scores for both models using 2,675 and 2,290 features were calculated using the built-in feature_importance_ function of a trained random forest model from the Scikit-learn library implemented in Python. This method estimates the importance of each feature based on the mean decrease in impurity across all trees in the ensemble. We were only able to generate values for 99 of the 100 maize features of greatest estimated importance in rice and *Arabidopsis*. One feature, 'gene age' was not included in cross-species prediction. A new model trained on maize data using only those 99 genomic features and the same procedure described above, was employed for cross-species predictions.

Validation of model predictions

Variant density was calculated by dividing the number of variants with minor allele frequency greater than 0.05 from the same resequencing-based dataset used to identify premature stop enriched genes occurring in the CDS of the primary transcript by the CDS length of the primary transcript of that same gene. Additional density values were generated by classifying genetic variants present within the primary transcript of a given gene into low effect, moderate effect, modifier effect, and high effect categories via SnpEff (De Baets et al., 2012). CHH methylation per gene model, rates of synonymous (ks) and non-synonymous substitutions (ka) based on alignment to orthologous genes in sorghum and K_a/K_s data was downloaded from MaizeGDB webpage (Woodhouse et al. 2025, https://mfs.maizegdb.org/Other,

https://mfs.maizegdb.org/DNAmethylation). The functional impact of genetic markers present between the annotated start and stop sites of each gene of interest was estimated using zero-shot scores provided by PlantCaduceus (Zhai et al., 2025). The absolute value of the log-likelihood difference between reference and alternate alleles was used when the reference or the alternate allele was assigned low relative probabilities, both of which indicate the presence of functional differences segregating in the population. ReelGene scores was downloaded for version 4 maize gene models (Schulz et al., 2025) which were then mapped to version 5 maize gene models. Gene models from version 5 genome annotation that mapped to multiple version 4 gene IDs were excluded resulting in removal of approximately 25% of gene models in version 5 which were likely due to gene annotation error in version 4 but were resolved in version 5. InterProScan V5.75 was used to identify gene models whose translated CDS coded for one or more domains of unknown function (Jones et al., 2014). Tests for GO enrichment or purification were performed using the Python library Goatools v1.4.12 (Klopfenstein et al., 2018). The Gene Ontology annotation file, "Zm00001eb.1.fulldata.txt.gz", for B73 reference genome version 5 was downloaded from maizeGDB (Hufford et al., 2021). To exclude bias introduced by the presence of greater proportions of some gene sets representing entirely uncharacterized proteins, the background set of all analyses was defined as the set of genes associated with one or more GO terms unless otherwise specified. GO term enrichment analysis was calculated as the fold change between the proportion of gene models annotated with a given GO term in the study set and its proportion in the background population set of the gene models.

Curation of rice and *Arabidopsis* gene model features

The rice genome was retrieved from the Rice Genome Annotation Project (Hamilton et al., 2025; Kawahara et al., 2013), and the *Arabidopsis* genome was retrieved from The

Arabidopsis Information Resource (TAIR) (Lamesch et al., 2012). For both species, only data from primary transcripts were used, and genes annotated as transposable elements (TEs) were excluded. Length features including 3' UTR length, 5' UTR length, total exon length, and average exon length were directly calculated from gene annotations in the GFF3 file. *k*-mer ratios were derived using a custom R script that integrated the FASTA sequence with GFF3 annotations. Codon usage frequency, relative synonymous codon usage (RSCU), and GC content at the third codon position (GC3s) were also computed from representative CDS.

Multiple protein sequence descriptors were extracted from one representative peptide sequence per gene using the protr R package (Xiao et al., 2015; R Core Team., 2021). After removing stop codons and filtering out sequences shorter than three amino acids, the following descriptor types were computed: triad composition, Geary, Moran, and Moreau-Broto autocorrelations, dipeptide composition, pseudo-amino acid composition (PAAC), quasi-sequence order (QSO), and composition-transition-distribution (CTD: CTDC, CTDD, CTDT). Custom functions with built-in error handling were applied to extract specific subsets of features.

All computations were implemented in parallel to accelerate processing. Syntenic rice genes were identified based on orthology with sorghum genes and directly obtained from the supplementary materials of a previously published study (Dias et al., 2026). Syntenic *Arabidopsis* genes were inferred via comparison with *Brassica rapa* genes using the latest reference genome and the SynOrths pipeline (Cheng et al., 2012). A single feature 'gene age' was not included, as no reliable public resources were found, and they could not be readily derived from sequence or annotation data using straightforward computational methods.

Code availability

All the code used to train, test and evaluate the model, along with the models trained in this study, is available in supplemental code and also in GitHub repository (https://github.com/NikeeShrestha/GenePrediction_Paper).

Acknowledgements

The authors would like to thank scientists from MaizeGDB, especially Dr. Maggie Woodhouse and Dr. Eddy Cannon, for providing information, access to the data used in this study, and help with curating the new phenotype-associated gene models. The authors would also like to thank Jingjing Zhai for providing access to zero-shot scores for maize variants published in MaizeGDB. This research was supported in part by the U.S. Department of Agriculture, National Institute of Food and Agriculture, under award numbers 2020-67021-31528 and 2020-68013-30934, and the AI Research Institutes program supported by NSF and USDA-NIFA under the AI Institute for Resilient Agriculture, Award No. 2021-67021-35329. Zhongjie Ji was additionally supported by a Heuermann Postdoctoral Fellow award from the University of Nebraska–Lincoln.

Author contributions statement

J.C.S. conceived of the project. N.S. processed and analyzed the data. Z.J. generated data for validation of the model on rice and *Arabidopsis*. J.C.S. provided supervision, guidance, and feedback on the design of analyses and approaches. N.S. drafted the paper with input from J.C.S. All authors contributed to the revision of the final manuscript.

Conflict of Interest:

James C. Schnable has equity interests in Data2Bio, LLC; and Dryland Genetics LLC and has performed paid work for Alphabet. He is a member of the scientific advisory board Aflo Sensors. The authors declare no other competing interests.

References

Alexandrov, N. N., Brover, V. V., Freidin, S., Troukhan, M. E., Tatarinova, T. V., Zhang, H., ... & Feldmann, K. A. (2009). Insights into corn genes derived from large-scale cDNA sequencing. Plant molecular biology, 69(1): 179-194.

Beder, T., Aromolaran, O., Dönitz, J., Tapanelli, S., Adedeji, E.O., Adebiyi, E., Bucher, G. and Koenig, R. (2021). Identifying essential genes across eukaryotes by machine learning. NAR genomics and bioinformatics, 3(4):p.lqab110.

Cheng, F., Wu, J., Fang, L., and Wang, X. (2012). Syntenic gene analysis between brassica rapa and other brassicaceae species. Frontiers in plant science, 3:198.

De Baets, G., Van Durme, J., Reumers, J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., Schymkowitz, J., and Rousseau, F. (2012). Snpeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. Nucleic acids research, 40(D1):D935–D939.

Dias, H. M., Sagawa, G. Y., Torres-Rodriguez, V. J., Mural, R. V., Schnable, J. C., and Van Sluys, M.-A. (2026). A sorghum-anchored pan-grass syntenic gene set in grasses. NAR Genomics and Bioinformatics, 8(2):Iqag030.

Doebley, J., Stec, A., Wendel, J., and Edwards, M. (1990). Genetic and morphological analysis of a maize-teosinte f2 population: implications for the origin of maize. Proceedings of the National Academy of Sciences, 87(24):9888–9892.

Doebley, J., Stec, A., and Hubbard, L. (1997). The evolution of apical dominance in maize. Nature, 386(6624):485–488.

Grzybowski, M. W., Mural, R. V., Xu, G., Turkus, J., Yang, J., and Schnable, J. C. (2023). A common resequencing-based genetic marker data set for global maize diversity. The Plant Journal, 113(6):1109–1121.

Guan, J., Wang, J., Niu, W., Peng, Z., Wang, S., Liu, Z., Zhou, G., and Ren, B. (2025). Towards recognizing food types for unseen subjects. ACM Transactions on Computing for Healthcare, 6(1):1–21.

Hamilton, J. P., Li, C., and Buell, C. R. (2025). The rice genome annotation project: an updated database for mining the rice genome. Nucleic Acids Research, 53(D1):D1614–D1622.

Huang, F., Jiang, Y., Chen, T., Li, H., Fu, M., Wang, Y., Xu, Y., Li, Y., Zhou, Z., Jia, L., et al. (2022). New data and new features of the funricegenes (functionally characterized rice genes) database: 2021 update. Rice, 15(1):23.

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Ou, S., Liu, J., Ricci, W. A., Guo, T., Olson, A., Qiu, Y., et al. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science, 373(6555):655–662.

Igartua, E., Contreras-Moreira, B., and Casas, A. M. (2020). Tb1: from domestication gene to tool for many trades. Journal of Experimental Botany, 71(16):4621–4624.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). Interproscan 5: genome-scale protein function classification. Bioinformatics, 30(9):1236–1240.

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S., et al. (2013). Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data. Rice, 6(1):4.

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., et al. (2018). Goatools: A python library for gene ontology analyses. Scientific reports, 8(1):10872.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). Timetree: a resource for timelines, time, trees, and divergence times. Molecular biology and evolution, 34(7):1812–1819.

Kustatscher, G., Collins, T., Gingras, A.-C., Guo, T., Hermjakob, H., Ideker, T., Lilley, K. S., Lundberg, E., Marcotte, E. M., Ralser, M., et al. (2022a). An open invitation to the understudied proteins initiative. Nature Biotechnology, 40(6):815–817.

Kustatscher, G., Collins, T., Gingras, A.-C., Guo, T., Hermjakob, H., Ideker, T., Lilley, K. S., Lundberg, E., Marcotte, E. M., Ralser, M., et al. (2022b). Understudied proteins: opportunities and challenges for functional proteomics. Nature Methods, 19(7):774–779.

Kwon, J. J., Pan, J., Gonzalez, G., Hahn, W. C., and Zitnik, M. (2024). On knowing a gene: A distributional hypothesis of gene function. Cell Systems, 15(6):488–496.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., et al. (2012). The arabidopsis information resource (tair): improved gene annotation and new tools. Nucleic acids research, 40(D1):D1202–D1210.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M. and Rhee, S.Y. (2010). Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. Nature biotechnology, 28(2):149-156.

Lei, X., Shao, H., Tang, Z., Xu, S., and Zhong, D. (2025). Cross-domain remaining useful life predictionunder unseen condition via mixed data and domain generalization. Measurement, 244:116451.

Lloyd, J. P., Seddon, A. E., Moghe, G. D., Simenc, M. C., and Shiu, S. H. (2015). Characteristics of plant essential genes allow for within-and between-species prediction of lethal mutant phenotypes. The Plant Cell, 27(8), 2133-2147.

Lloyd, J. and Meinke, D. (2012). A comprehensive dataset of genes with a loss-of-function mutant phenotype in arabidopsis. Plant physiology, 158(3):1115–1129.

Moore, B.M., Wang, P., Fan, P., Leong, B., Schenck, C.A., Lloyd, J.P., Lehti-Shiu, M.D., Last, R.L., Pichersky, E. and Shiu, S.H. (2019). Robust predictions of specialized metabolism genes through machine learning. Proceedings of the National Academy of Sciences, 116(6):2344-2353.

Oellrich, A., Walls, R.L., Cannon, E.K., Cannon, S.B., Cooper, L., Gardiner, J., Gkoutos, G.V., Harper, L., He, M., Hoehndorf, R. and Jaiswal, P. (2015). An ontology approach to comparative phenomics in plants. Plant methods, 11(1):10.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12:2825-2830.

Pena-Castillo, L. and Hughes, T. R. (2007). Why are there still over 1000 uncharacterized yeast genes? Genetics, 176(1):7–14.

R Core Team, 2021. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.

Sahay, S., Grzybowski, M., Schnable, J. C., and Głowacka, K. (2023). Genetic control of photoprotection and photosystem ii operating efficiency in plants. New Phytologist, 239(3):1068–1082.

Schnable, J. C. (2020). Genes and gene models, an important distinction. New Phytologist, 228(1):50–55.

Schnable, J. C. and Freeling, M. (2011). Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. PloS one, 6(3):e17855.

Schulz, A.J., Zhai, J., AuBuchon-Elder, T., Andorf, C.M., El-Walid, M.Z., Ferebee, T.H., Gilmore, E.H., Hufford, M.B., Johnson, L.C., Kellogg, E.A., La, T., Long, E., McMorrow, S.J., Miller, Z.R., Portwood, J.L., II, Romay, M.C., Seetharam, A.S., Stitzer, M.C., Woodhouse, M.R., Wrightsman, T., Buckler, E.S., Monier, B. and Hsu, S-K. (2025), Fishing for a reelGene: evaluating gene models with evolution and machine learning. Plant J, 123:e70483.

Sen, S., Woodhouse, M. R., Portwood, J. L., and Andorf, C. M. (2023). Maize feature store: A centralized resource to manage and analyze curated maize multi-omics features for machine learning applications. Database, 2023:baad078.

Shrestha, N., Hu, H., Shrestha, K., and Doust, A. N. (2023). Pearl millet response to drought: A review. Frontiers in Plant Science, 14:1059574.

Sinha, S., Eisenhaber, B., Jensen, L.J., Kalbuaji, B. and Eisenhaber, F. (2018). Darkness in the human gene and protein function space: widely modest or absent illumination by the life science literature and the trend for fewer protein function discoveries since 2000. Proteomics, 18(21-22):1800093.

Stoeger, T., Gerlach, M., Morimoto, R. I., and Nunes Amaral, L. A. (2018). Large-scale investigation of the reasons why potentially important genes are ignored. PLoS biology, 16(9):e2006643.

Upadhyaya, S.R., Bayer, P.E., Tay Fernandez, C.G., Petereit, J., Batley, J., Bennamoun, M., Boussaid, F. and Edwards, D. (2022). Evaluating plant gene models using machine learning. Plants, 11(12):1619.

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., and Vandepoele, K. (2022). Plaza 5.0: extending the scope and power of comparative and functional genomics in plants. Nucleic Acids Research, 50(D1):D1468–D1474.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S. and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science, 350(6264):1096-1101.

Washburn, J. D., Mejia-Guerra, M. K., Ramstein, G., Kremling, K. A., Valluru, R., Buckler, E. S., and Wang, H. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance from dna sequence. Proceedings of the National Academy of Sciences, 116(12):5542–5549.

Woodhouse, M. R., Cannon, E. K., Portwood, J. L., Gardiner, J. M., Hayford, R. K., Haley, O., & Andorf, C. M. (2025). Tools and resources at the maize genetics and genomics database (MaizeGDB). Cold Spring Harbor Protocols, 2025(1), pdb-over108430.

Xiao, N., Cao, D.-S., Zhu, M.-F., and Xu, Q.-S. (2015). protr/protrweb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics, 31(11):1857–1859.

Zhai, J., Gokaslan, A., Schiff, Y., Berthel, A., Liu, Z.-Y., Lai, W.-Y., Miller, Z. R., Scheben, A., Stitzer, M. C., Romay, M. C., et al. (2025). Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained dna language model. Proceedings of the National Academy of Sciences, 122(24):e2421738122.

Figure 1. Differences between premature stop enriched-common genes (PSE-Cs) and phenotype-associated genes (PAGs). A). Gene models in the B73 reference genome contain alternate alleles that introduce premature stop codons, totaling 713 models. A threshold of 0.25 alternate allele frequency defined 162 premature stop enriched-common genes (PSE-Cs), which were used as a second comparator category of gene models in this study. B). Principal component analysis (PCA) of PSE-Cs (*n*=162) and PAGs (*n*=272). The first two principal components explain 28.62% of the variance. Data included in the PCA included a set of 2,675 previously published features that were downloaded from MaizeGDB (Sen et al., 2023). C). Difference in the number of exons per gene model between PSE-Cs and PAGs. D). Difference in the length of the three prime untranslated regions (3′UTRs) of gene models between PSE-Cs and PAGs. E). Difference in the number of isoforms per gene model between PSE-Cs and PAGs. F). Difference in the proportion of GC content between PSE-Cs and PAGs. ** p-value < 0.01, *** p-value < 0.001, n.s non-significant (two-tailed Mann-Whitney *U* test).

Figure 2. Model performance using different subsets of total features. A). ROC curves are generated from three models: one using all 2,675 features (non-genomic, genomic sequence and comparative genomics-derived), one using 2,290 genomic sequence and comparative genomics derived features, and one using the top 100 genomic sequence and comparative genomics derived features. Performance was evaluated on 48 withheld gene models. B). Predicted probability from the three models in Panel A across all maize gene models.

Figure 3. Phenotype-associated genes (PAGs) have higher predicted probabilities than uncategorized gene models in maize, rice, and *Arabidopsis*. A). Distribution of predicted probabilities of PAGs, premature stop enriched-common (PSE-C) gene models and

uncharacterized maize gene models. **B**). Distribution of predicted probabilities of a subset of validated gene models (*n*=96) retrieved from the literature in rice (Huang et al., 2022) using a model trained on maize data with 99 genomic sequence and comparative genomics derived features. C). Distribution of predicted probabilities of a subset of validated gene models (*n*=2,018) retrieved from the literature in *Arabidopsis* (Lloyd and Meinke, 2012) using a model trained on maize data with 99 genomic sequence and comparative genomics derived features. **D).** Predicted probability distribution of newly validated PAGs (*n* = 27, curated from MaizeGDB since November 27, 2021) compared to hold-out PAGs, hold-out premature stop enriched-common genes (PSE-Cs), premature stop enriched-rare genes (PSE-Rs) and uncategorized gene models.

Figure 4. Genes with higher estimated effects on plant phenotypes are conserved in the maize population. A). Bottom and top-quartile gene models were defined from the predicted probability distribution derived from the model trained with the top 100 genomic sequence and comparative genomics derived features. B). Bottom-quartile gene models show ~4.4-fold higher odds of containing premature stop codons compared to top-quartile models (p-value < 2.26×10^{-16}, Fisher's exact test) C). Variant frequency per kilobase in coding sequences of the primary transcript of the gene models, separated by effect type. Bottom-quartile models have higher frequencies of high effect (p-value = $9.03 * 10^{-33}$) and moderate effect variants (p-value = 0), whereas top-quartile models have higher frequencies of low effect variants (p-value = 0; two-tailed Mann-Whitney *U* test). D). Proportional distribution of high, moderate, and low effect variants between quartiles. E). Average absolute zero-shot score for variants estimated from a pre-trained DNA language model, PlantCaduceus (Zhai et al. 2025), in different parts of the gene

models between the bottom-quartile gene models and top-quartile gene models from panel A. ** $0.01 < \text{p-value} < 1 * 10^{-20}$, *** $\text{p-value} < 1 * 10^{-200}$ (two-tailed Mann-Whitney *U* test).