Maize Sciences, 2026, xx(x): xx−xx

Method

Identification of Intronic GT/AG Gain Variants Affecting Splicing in an EMS-Mutagenized Maize Population

BAO Yu[1, #], XIA Fan[2, #], QIN Li[1, #], XIE Fugui[1], GUO Xiaolong[1], YIN Can[2], WANG Xiangfeng[2], LU Xiaoduo[1]

([1]*Institute of Molecular Breeding for Maize, Qilu Normal University, Jinan 250200, China;* [2]*State Key Laboratory of Maize Bio-breeding / Frontiers Science Center for Molecular Design Breeding, China Agricultural University, Beijing 100094, China;* [#]*These authors contributed equally to this article.*)

Abstract: Splicing is a critical step in post-transcriptional processing of genes, and its accuracy directly determines whether mature mRNA can be correctly formed. Current strategies for identifying functional mutations from EMS mutant populations have largely focused on premature termination codons and canonical splice sites, while intronic regions—central to splicing regulation—contain potential regulatory variants that create new GT/AG dinucleotides affecting splicing efficiency, which remain largely unexplored. In this study, we developed a workflow integrating in silico screening, PlantCaduceus deep learning-based prediction, and experimental validation to identify intronic GT/AG gain variants in maizeEMSDB mutant library. Screening candidate sites yielded a validation rate of 71.4% (5/7). AG gain variants showed a distinct positional bias: newly created AG sites clustered within 50 bp upstream of canonical acceptors, matching the branchpoint-AG region, and their splicing efficiency decreased as distance from the canonical site increased. This pattern is highly consistent with pathogenic AG-gain variants reported in human genetic studies. By contrast, GT gain variants showed no such positional or distance-dependent effects. Among the validated sites, an AGG→AAG mutation at a canonical splice acceptor site (*cl30719_1*) revealed a competitive splicing phenomenon: the mutation reduced splicing efficiency at the canonical AG, allowing nearby sites to compete and activating a downstream AG that produced aberrant transcripts. Together, this study establishes a screening method for functional gene research in maize and advances understanding of plant splicing regulation.

Key words: maize; EMS mutant library; intron; GT/AG gain variants; splicing

FUNDING

This study was supported by institutional projects financed by the Shandong Provincial Natural Science Foundation (ZR2025QC177).

INTRODUCTION

Ethyl methanesulfonate (EMS) mutagenesis is widely used to generate genome-wide mutant libraries by inducing G/C to A/T transitions, which disrupt gene function and has been instrumental in plant functional genomics research (Lu et al., 2018; Chong et al., 2025; Wang et al., 2025). Recently, several gene-indexed maize EMS mutant libraries have been established, including the B73 mutant library encompassing over 32 000 genes (Lu et al., 2018) and the Zheng58 mutant library with more than two million identified mutations (Chong et al., 2025). These resources provide an important foundation for maize functional genomics and molecular breeding studies.

Accurate splicing depends on conserved *cis*-

Received: 23 March 2026; **Accepted:** 3 April 2026
Corresponding authors: Lu Xiaoduo (lu.xiaoduo@163.com); Wang Xiangfeng (xwang@cau.edu.cn)

https://doi.org/10.2738/MS.2026.0003

elements within introns, including the 5' splice donor site (GT), the 3' splice acceptor site (AG), the branchpoint, and the polypyrimidine tract (Wahl et al., 2009; Choi et al., 2023; Wilkinson et al., 2020; Zhan et al., 2024). Notably, the region between the branchpoint and the AG acceptor is particularly critical for proper spliceosome recognition (Zhang et al., 2023). Current strategies for identifying functional mutations from EMS populations have focused primarily on premature termination codons (stop-gain) and disruption of canonical splice sites (the conserved GT-AG dinucleotides at intron boundaries) (Lu et al., 2018; Zhang et al., 2025; Wang et al., 2025). Although these approaches have identified numerous causal variants, the regulatory potential of intronic regions remains underexplored. Introns account for much of plant genomes and carry regulatory sequences that control pre-mRNA processing fidelity. Among intronic variants, GT/AG gain variants—which create new GT or AG dinucleotides through single nucleotide changes within intronic sequences—are an important class of regulatory variants. These newly created dinucleotides can activate cryptic splice sites, leading to various splicing alterations including pseudoexon inclusion, partial intron retention, and disruption of normal splicing patterns (Zhang et al., 2023). Studies have shown that within the human branchpoint-to-acceptor region, 41.9% of pathogenic splice-altering variants create new AG dinucleotides, with 93.1% of these pathogenic AG-gain variants located within a defined high-risk region. In contrast, systematic identification and functional characterization of GT/AG gain mutations in crop genomes remain relatively limited, offering opportunities to discover regulatory variants with potential breeding value.

Accurate prediction of splicing outcomes is critical for prioritizing candidate mutations. Deep learning methods have greatly improved splicing effect prediction in human genomics, with tools such as SpliceAI showing high accuracy in identifying splice-altering variants (Jaganathan et al., 2019; Chao et al., 2025; Chao et al., 2024; Scalzitti et al., 2021). In plant genomics, the recently developed PlantCaduceus—a DNA language model pre-trained on 16 angiosperm genomes—serves as a specialized tool for plant sequence analysis (Zhai et al., 2025). When fine-tuned with limited *Arabidopsis* data, PlantCaduceus transfers well across species, improving performance by 1.45-fold over existing DNA language models in predicting maize splice donor sites and matching the performance of advanced protein language models in variant effect prediction. This cross-species capability makes it suitable for predicting the functional consequences of intronic mutations in maize.

In this study, we developed a workflow to identify, predict, and validate intronic GT/AG gain variants from maizeEMSDB. This approach integrates: (i) genome-wide screening of intronic variants that create new GT or AG dinucleotides, (ii) PlantCaduceus-based prediction of splicing probability changes between wild-type and mutant sequences, and (iii) experimental validation through RT-PCR and sequencing to confirm splicing alterations. Applying this pipeline to seven candidate mutations, we found that 71.4% (5/7) induced detectable splicing variants, including canonical splice site shifts, competitive activation of downstream AG sites, and activation of cryptic splice sites. This work provides a methodological framework for mining intronic regulatory mutations from EMS populations, yields new experimental insights into the complexity of splice site selection in plants, and has potential applications in functional gene discovery and molecular breeding.

RESULTS AND ANALYSIS

Overview of the screening workflow for intronic GT/AG gain variants

To identify intronic GT/AG gain variants that affect splicing in our maize EMS population, we developed a workflow combining computational screening, deep learning prediction, and experimental validation (Fig. 1).

The workflow consists of three steps: First, we screened the EMS library for variants that create new GT or AG dinucleotides within introns; second, we first used PlantCaduceus zero-shot prediction to prioritize deleterious variants, then predicted splicing effects to identify those altering splice site usage; and third, we validated candidates by RT-PCR and sequencing to analyze splicing patterns and protein-level consequences. This approach efficiently identifies functional intronic regulatory variants for maize functional genomics and molecular breeding.

Genome-wide screening and prediction of intronic GT/AG gain variants

Following the workflow described above, we

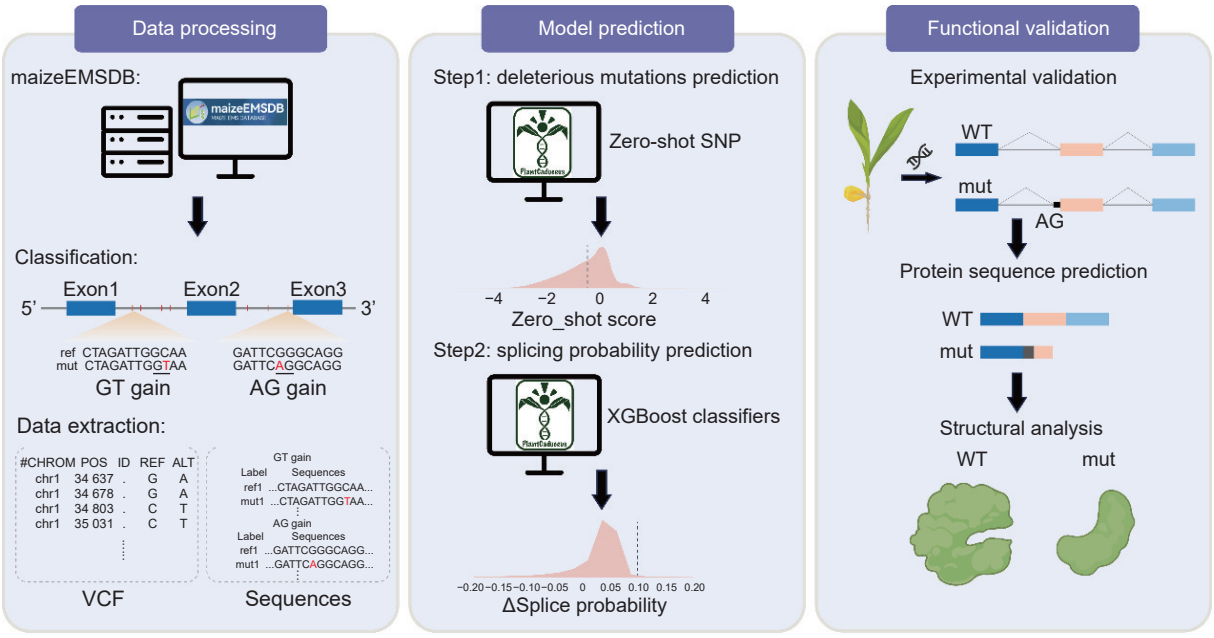

Fig. 1. Workflow for identifying functional intronic GT/AG gain variants.
The workflow involves three steps: variant screening, prediction, and validation. First, intronic GT/AG gain variants are extracted from the maize EMS database (maizeEMSDB) together with VCF records and genomic sequences. Next, PlantCaduceus predicts deleteriousness and splice site usage; variants passing both thresholds are selected as candidates. Candidate mutants are grown to identify homozygous or heterozygous plants, and transcript analysis detects splicing alterations. Finally, amino acid changes and protein structures are predicted to assess functional impact.

screened for intronic GT/AG gain variants in the maizeEMSDB. Using whole-genome sequencing data and genome annotation (GFF), we identified all intronic variants that created new GT or AG dinucleotides. This yielded 89 173 GT gain and 167 389 AG gain variants (Fig. 2-A). All mutation data are publicly accessible through the maizeEMSDB website, allowing users to search for mutations in genes of interest and request corresponding seeds for functional studies.

To gauge potential deleteriousness, we used PlantCaduceus zero-shot scores as measures of evolutionary conservation, with more negative scores indicating higher conservation and greater potential deleteriousness (Zhai et al., 2025). To evaluate the consistency of zero-shot predictions across the four models (l20, l24, l28, l32), we first analyzed cross-model correlations (Fig. S1-A). The results showed high inter-model correlations (Pearson R = 0.85–0.92), indicating that the four models provide highly consistent evaluations of variant deleteriousness potential. Given this high consistency, we used the average zero-shot score across the four models as a reliable summary for initial filtering. Given the absence of a unified threshold standard for deleteriousness prediction in plant intronic regions, we selected a relatively lenient threshold (zero-shot score < −0.5) as an initial filtering criterion to maximize the retention of candidate sites that might affect splicing (Fig. 2-B). At this cutoff, we identified 89 173 potentially deleterious mutations from 167 389 AG gain variants (53.3%) and 135 970 potentially deleterious mutations from 261 796 GT gain variants (51.9%) (Fig. 2-A), which were subsequently subjected to fine-scale screening through splicing probability change prediction.

Focusing on these candidates, we used PlantCaduceus to compare splicing probabilities between wild-type and mutant sequences (Δscore probability = mutant probability - wild-type probability) for each of the four models. We then analyzed cross-model correlations of ΔSplice probability (Fig. S1-B). The results showed moderate inter-model correlations (Pearson R = 0.51–0.60), indicating that while the models share some agreement, their predictions are not fully consistent. To obtain robust predictions that mitigate model-specific biases, we therefore calculated the average ΔSplice probability across the four models as an integrated score for each variant. The results showed that 850 GT gain and 2796 AG gain variants exhibited substantial changes in splicing probability (Fig. 2-C; Table S1), suggesting they may create new splice sites. For variants predicted to potentially create new

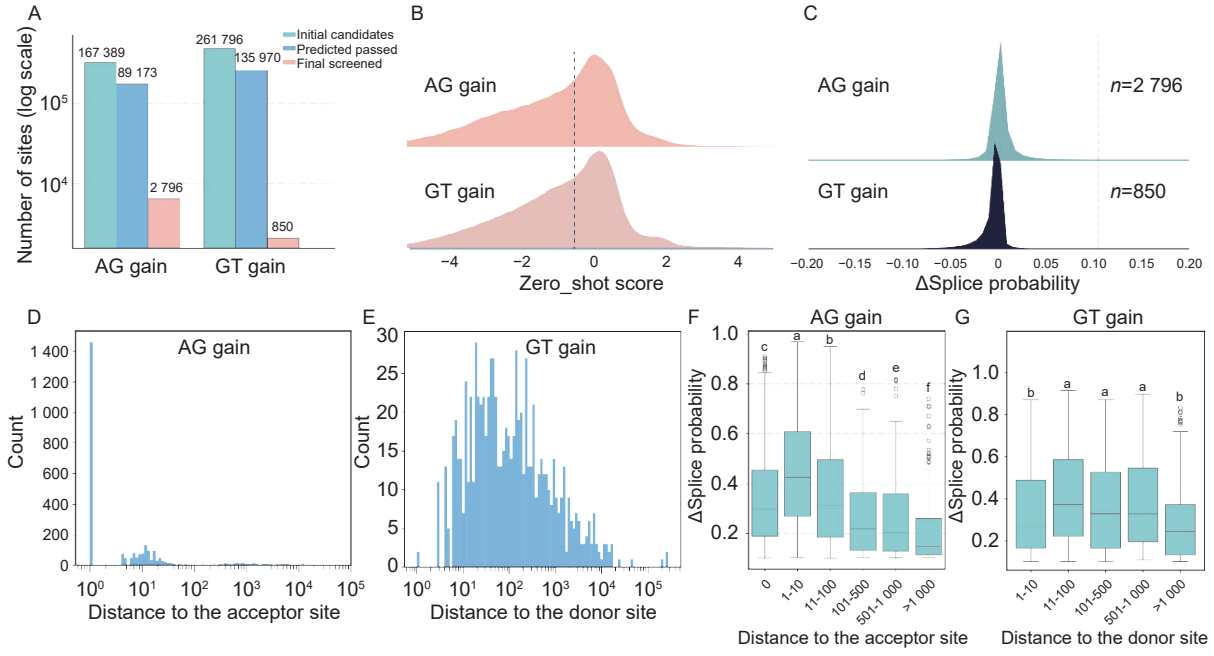

Fig. 2. Genome-wide screening and prediction of intronic GT/AG gain variants.

A. Numbers of GT and AG gain variants retained at each screening step: total variants identified from the maize EMS library, potentially deleterious variants (zero-shot score < -0.5), and variants predicted to alter splicing (ΔSplice probability > 0.1). **B**. Zero-shot score distribution for all GT and AG gain variants. The dashed line marks the threshold for potentially deleterious variants (score < -0.5). **C**. Distribution of splicing probability changes (ΔSplice probability) for GT and AG gain variants. The dashed line marks the threshold for predicted splice-altering effects (ΔSplice probability > 0.1). **D**. Distances of newly created AG sites from canonical AG acceptors for AG gain variants predicted to affect splicing. The shaded region highlights the branchpoint-AG interval (< 50 bp). **E**. Distances of newly created GT sites from canonical GT donors for GT gain variants predicted to affect splicing. **F**. Boxplots show ΔSplice probability of AG gain variants grouped by distance from the canonical AG. Statistical significance among groups was assessed using Kruskal-Wallis test followed by Dunn's post-hoc test for multiple comparisons. Different letters indicate significant differences ($P < 0.01$). **G**. Boxplots show ΔSplice probability of GT gain variants grouped by distance from the canonical GT. Statistical significance among groups was assessed using Kruskal-Wallis test; no significant differences were detected ($P > 0.05$).

splice sites (Δscore probability > 0.1), we analyzed distances to canonical splice sites. AG gain variants were mainly found within 50 bp upstream of the canonical AG, clustered between the branchpoint and acceptor (Fig. 2-D). This finding is highly consistent with observations from human genetic studies (Zhang et al., 2023), suggesting that AG gain variants may affect splicing by interfering with recognition between the branchpoint and the acceptor. By contrast, GT gain variants showed a dispersed distribution of distances from the canonical GT, predominantly ranging from 10 to 1000 bp, with no significant regional enrichment (Fig. 2-E), implying that the mechanism of action for GT gain variants may differ from that of AG gain variants.

Further analysis revealed a significant negative correlation between the distance from the canonical AG and the splicing probability change for AG gain variants: the greater the distance, the smaller the change (Fig. 2-F; Fig. S2-A). These findings revealed a mechanistic divergence between the two variant types: while AG gain variants exhibited strict distance dependency with significantly higher splicing efficiency within the 50 bp branchpoint-to-acceptor region, GT gain variants showed comparable splicing efficiency across the <10 to 1000 bp range, with significant attenuation only observed at distances beyond 1000 bp (Fig. 2-G; Fig. S2-B).

Experimental validation reveals diverse splicing patterns and competitive mechanisms

From these predictions, we selected 10 high-confidence candidates considering splicing probability changes, splice-site positions, and gene functional annotation. Through RT-PCR and sequencing analysis of splicing products from mutant materials, results were obtained for seven sites, among which five (71.4%) exhibited clear splicing alterations, validating the effectiveness of our screening pipeline (Fig. 3; Table 1).

Notably, among AG gain variants, AGG→AAG mutations at canonical splice acceptor sites warranted special attention. We identified 3284 such variants genome-wide. PlantCaduceus predicted that ~1460

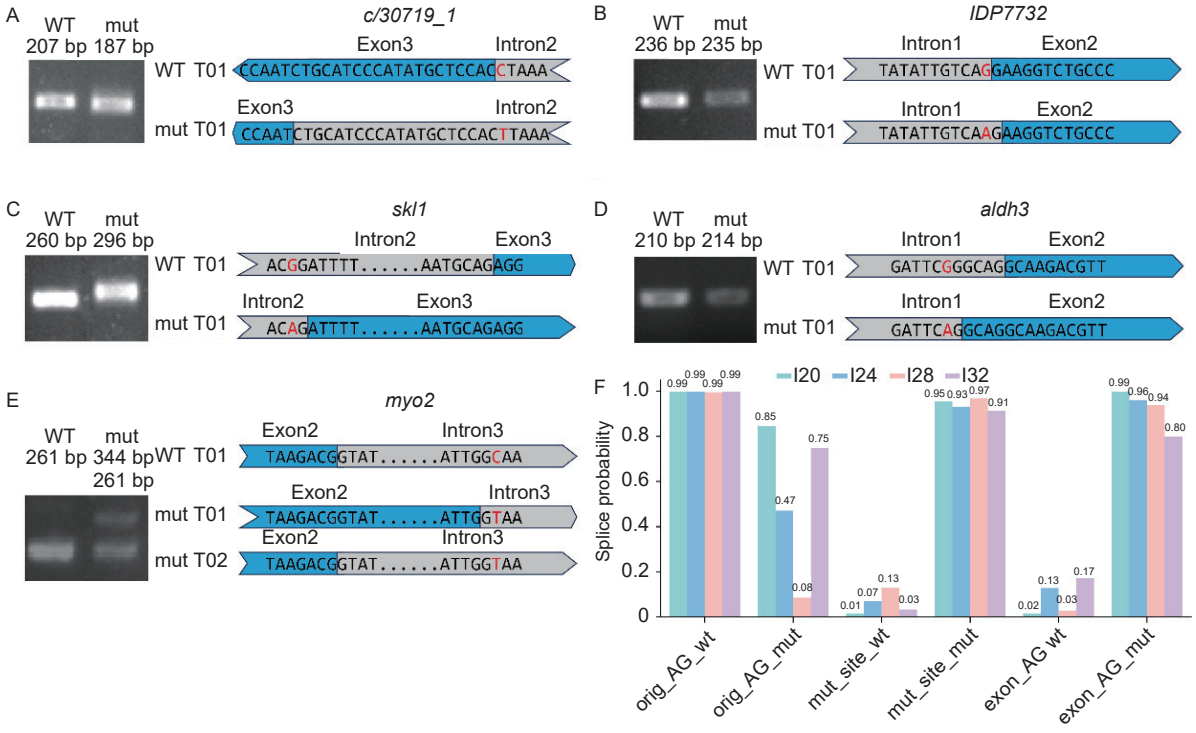

Fig. 3. Experimental validation of candidate sites.

A-E. RT-PCR gel electrophoresis and schematic diagrams illustrating splicing alterations for five positively validated candidate sites, including *cl30719_1*, *IDP7732*, *skl1*, *aldh3*, and *myo2*. **F**. Splicing probability changes before and after mutation for three AG sites (canonical AG, AG gain mutation site, and downstream exonic AG site) at the *cl30719_1* locus, as predicted by four PlantCaduceus models (l20/l24/l28/l32).

Table 1. Experimental validation of intronic GT/AG gain mutations

Loc	Variant	Mut type	Dist.(bp)	Gene	Zero-shot score	Δscore	Geno.	Result
chr1:1 268 188 084	G>A	AG gain	0	*IDP7732*	−4.10	0.83	aa	Positive
chr3:237 065 791	C>T	AG gain	0	*cl30719_1*	−3.24	0.88	aa	Positive[1]
chr6:133 977 128	C>T	AG gain	0	*rpl23a*	−1.49	0.35	-	-
chr9:161 760 385	C>T	AG gain	0	*ftd1*	−7.15	0.34	-	-
chr9:34 632 636	C>T	AG gain	0	*asn1*	−7.32	0.29	-	-
chr1:154 801 155	G>A	AG gain	36	*skl1*	−0.53	0.72	aa	Positive
chr3:227 308 106	G>A	AG gain	4	*aldh3*	−2.49	0.53	aa	Positive
chr5:224 092 037	C>T	GT gain	96	*myo2*	−1.61	0.48	Aa	Positive
chr10:150 787 348	C>T	AG gain	8	*kea2*	−3.13	0.34	aa	Negative
chr1:289 863 209	G>A	AG gain	28	*uce2*	−0.50	0.54	aa	Negative

Note: Dist. (bp), distance from the newly created splice site to the canonical splice site; Geno., genotype (aa, homozygous mutant; Aa, heterozygous mutant; –, not determined). [1] *cl30719_1* showed an atypical splicing pattern: the mutation activated a downstream exonic AG site instead of the predicted AG gain mutation site.

would create new AG sites causing 1-bp acceptor shifts and frameshifts. We selected five sites from this category for experimental validation and successfully obtained clear results for two sites (Fig. 3-A to -B; Table 1).

IDP7732 behaved as expected: the mutation caused a 1-bp downstream shift of the splice acceptor site, deleting one nucleotide from mature mRNA and creating a frameshift (Fig. 3-B). By contrast, the *cl30719_1* displayed a different splicing pattern. Instead of activating the predicted AG gain variant site, it activated an endogenous AG site in exon 3,

leading to partial exon sequence excision (Fig. 3-A). This suggests that an AGG→AAG mutation can trigger competitive splicing among AG sites near the canonical AG. To test this, we examined acceptor splicing probabilities before and after mutation for three types of AG sites: the canonical AG, the AG gain mutation site, and the downstream exonic AG site. After mutation, the splicing probability of the canonical AG decreased substantially, while those of both the AG gain mutation site and the downstream exonic AG site increased markedly. In the PlantCaduceus_l20 model, the AG gain mutation site reached 0.95, and the downstream exonic AG site reached 0.99 (Fig. 3-F). This pattern suggests that the two sites may compete for spliceosome recognition, with the exonic AG site being preferentially selected—consistent with our experimental observation that the exonic AG outcompeted the mutant site. By contrast, PlantCaduceus_l28 and l32 predicted the opposite pattern, favoring the AG gain mutation site over the exonic one. These discrepancies indicate that while model predictions can serve as useful references for evaluating competitive splicing outcomes, results may vary across models, underscoring the need for combined consideration to reduce single-model bias.

We also tested five intronic GT/AG gain variants outside canonical splice sites. The results showed that three of these (60%) exhibited activation of new splice sites, all producing in pseudoexon inclusion (Fig. 3-C to -E; Table 1). Overall, five of seven candidates (71.4%) tested positive, confirming our pipeline is effective. The competitive splicing observed at AGG→AAG sites, together with the variability among model predictions, highlights the complexity of plant splicing regulation and demonstrates that integrating multiple models enhances prediction reliability.

Protein structural impacts of splicing variants and optimization of screening strategy

To evaluate the impact of splicing variants on protein function, we analyzed protein sequence and structure for two representative sites. For *skl1*, activation of a new splice site led to a 36-nucleotide insertion between exons 2 and 3, resulting in the addition of 12 amino acids after residue 63 (Fig. 4-A). Domain annotation revealed that the Shikimate kinase/gluconokinase domain of the skl1 protein is located between amino acids 87 and 215, with the 12-amino acid insertion falling outside this region, leaving the functional domain intact. Three-dimensional structure prediction and comparison of the wild-type and mutant proteins showed an RMSD value of 0.274, indicating no significant structural change and suggesting that protein function is likely unaffected. By contrast, for *myo2*, a C-to-T mutation 97 bp downstream of exon 2 in myo2 created a new GT site and activated a cryptic donor. This activation led to a frameshift in the reading frame, with a premature termination codon encountered at the 34th amino acid following the frameshift point, producing a truncated protein and consequently resulting in loss of protein function (Fig. 4-B).

The contrast between these two cases highlights that detecting splicing variants alone is not sufficient; the ultimate impact on protein sequence and structure must also be considered. Frameshifts typically disrupt protein function and should be prioritized in reverse genetic screens. In-frame insertions, however, have uncertain effects and require case-by-case evaluation

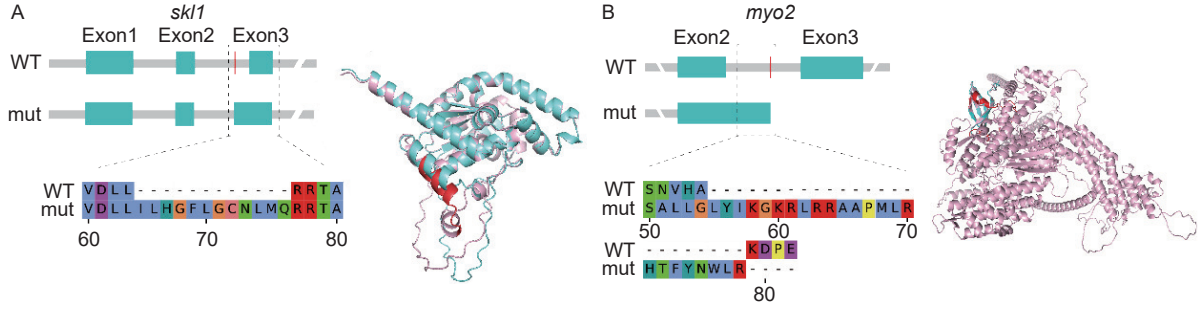

Fig. 4. Structural impacts of splicing variants on protein function.

A. Protein structure analysis of skl1. Left panel: Schematic diagrams illustrating gene structure alterations and protein sequence changes before and after mutation. Right panel: Three-dimensional structure comparison of wild-type (blue) and mutant (pink) proteins, with the 12-amino acid insertion highlighted in red. **B**. Protein structure analysis of myo2. Left panel: Schematic diagrams illustrating gene structure alterations and protein sequence changes before and after mutation. Right panel: Three-dimensional structure comparison of wild-type (pink) and mutant (blue) proteins, with altered amino acid residues in the mutant highlighted in red.

based on domain location and function. Therefore, we recommend prioritizing sites predicted to cause frameshift mutations for experimental validation, as this approach enhances the efficiency of obtaining loss-of-function mutants.

DISCUSSION

In this study, we developed a pipeline that integrates in silico screening, deep learning-based prediction, and experimental validation to systematically identify intronic GT/AG gain variants affecting splicing in a maizeEMSDB mutant library. Experimental validation of seven candidate sites yielded a positive rate of 71% (5/7), indicating that the pipeline successfully identified functionally relevant splicing mutations. Analysis of the distribution, splicing effects, and protein-level impacts of these mutations revealed distinct mechanistic features between AG and GT gain variants, providing practical insights for refining screening strategies.

AG and GT gain variants differed markedly in both distance distribution and splicing effects. Mutations creating new AG sites were predominantly located within 50 bp upstream of the canonical AG, a region corresponding to the branchpoint-AG interval. This pattern closely mirrors findings from human genetic studies (Zhang et al., 2023), in which 93.1% of pathogenic AG-gain variants cluster in a defined high-risk region downstream of the branchpoint and upstream of the acceptor. The cross-species conservation of this feature suggests a shared mechanism: newly created AG sites in this region may compete with U2AF35 binding, interfering with U2AF65 recognition of the polypyrimidine tract and disrupting normal spliceosome assembly (Wahl et al., 2009; Wilkinson et al., 2020). In contrast, mutations creating new GT splice sites showed no significant regional enrichment in their distance distribution, and their splicing probability changes exhibited no significant differences across the 10-1000 bp range. This suggests that the mechanism of action for GT gain variants may differ from that of AG gain variants. As the core dinucleotide of the splice donor site, GT recognition primarily depends on base complementarity with U1 snRNP. Newly created GT sites, as long as they are located within intronic regions, may be recognized by U1 snRNP regardless of their distance from the canonical GT. However, whether they successfully compete with the original donor site may depend more on factors such as sequence context and RNA secondary structure (Bartys et al., 2019; Shepard and Hertel, 2008).

While the averaging approach provided a robust and effective initial filter in this study (71.4% validation rate), future applications could benefit from more sophisticated ensemble strategies. For instance, a criterion requiring a positive prediction from the l32 model (which has the largest receptive field and model capacity) supported by at least one positive from the other three models, or a weighted averaging scheme favoring l32, could further reduce false positives and improve the precision–recall trade-off for competitive splicing event prediction. Systematic evaluation of such strategies on benchmark datasets will be an important direction for future method development.

We also observed evidence of competitive splicing. Two AGG→AAG mutations (IDP7732 and cl30719_1) produced different outcomes: one resulted in splice site shifting, while the other activated an AG site within the downstream exon. Comparative analysis of splicing probabilities for the canonical AG, the mutant AG site, and the downstream exonic AG site showed that upon mutation, the probability of the canonical AG dropped sharply, while those of the other two sites increased. These results support a competitive splicing model: upon mutation, recognition efficiency of the original splice site decreases, while newly created or pre-existing cryptic AG sites gain enhanced competitiveness, with the spliceosome ultimately selecting the site based on relative strength (Smith et al., 1993; Chua and Reed, 2001). Notably, different prediction models (l20/l24/l28/l32) showed discrepancies in their splicing probability predictions for the same AG site, reflecting current limitations of deep learning models and suggesting that in practical applications, predictions from multiple models should be considered comprehensively for decision-making.

Differences in protein-level consequences further underscore the importance of evaluating beyond splicing alone. In skl1, a 12-amino acid in-frame insertion fell outside the Shikimate kinase/gluconokinase domain and caused no significant structural change (RMSD = 0.274), suggesting it is a tolerated variant. In myo2, a frameshift mutation led to a truncated protein and likely loss of function. This contrast illustrates that frameshift mutations—which almost invariably disrupt function—warrant higher priority in screening

for loss-of-function alleles. In-frame insertions, by contrast, require case-by-case assessment based on domain context. We therefore recommend prioritizing frameshift-predicted sites for experimental validation to improve the efficiency of recovering loss-of-function mutants. It is worth noting that current structure prediction tools, such as AlphaFold3, primarily provide static three-dimensional structures and evaluate variant effects using metrics like RMSD (Nam, 2026). However, RMSD alone may not fully capture the functional impact of a variant. On one hand, variants can affect protein dynamics (e.g., local flexibility, conformational changes) that are not readily reflected in static structures (Schaefer and Rost, 2012; Ramasamy et al., 2026). On the other hand, variants may alter post-translational modification sites (e.g., phosphorylation, ubiquitination), thereby regulating protein activity, stability, or interactions—effects that cannot be directly assessed by RMSD (Ramasamy et al., 2026; Nam, 2026). Therefore, for variants with minimal RMSD changes, such as skl1, additional experimental approaches (e.g., molecular dynamics simulations, post-translational modification assays) are needed to fully evaluate functional consequences.

The maize EMS mutant library, originally established for forward and reverse genetics, has proven to be a rich resource for exploring regulatory variation (Lu et al., 2018). The pipeline developed here for GT/AG gain variants can be readily adapted to identify other classes of regulatory variants, such as mutations affecting branchpoints, poly-pyrimidine tracts, or exonic splicing enhancers. As the library continues to be phenotypically characterized, it will serve as an enduring platform for dissecting gene function and discovering alleles with breeding potential. The seed stocks corresponding to each mutation are publicly available, enabling the broader community to validate and utilize these variants in their own research programs.

MATERIALS AND METHODS

Screening and Prediction of GT/AG Gain variants

Based on sequencing data from our previously established maizeEMSDB mutant library, we first performed functional annotation of genome-wide SNP sites using the SnpEff software and extracted mutation sites located within intronic regions. Taking each mutation site as the center, we extracted 3-bp sequences (including one base upstream and one base downstream of the mutation) according to the gene transcription direction to determine whether the mutation created new GT or AG dinucleotides. GT gain variants and AG gain variants satisfying these criteria were saved as separate VCF files. The zero_shot_score.py script from the PlantCaduceus model was used to calculate zero-shot scores for each mutation site, which reflect the degree of evolutionary conservation, with more negative scores indicating higher conservation and greater potential deleteriousness (Zhai et al., 2025). In this study, a threshold of −0.5 was applied, and mutation sites with zero-shot scores less than −0.5 were selected as potentially deleterious mutations for subsequent analysis.

For the selected potentially deleterious mutations, sequence files were generated according to mutation type. The sequence extraction rule was as follows: For each selected variant, a 512 bp reference sequence was extracted centered on the key signal site (the newly created GT or AG dinucleotide or the canonical splice site), with coordinates adjusted according to gene strand orientation to ensure that the key signal was positioned at fixed index positions consistent with model input requirements (for genes on the forward strand, GT signals were located at positions 256-257 and AG signals at positions 254-255; for genes on the reverse strand, AC signals [corresponding to forward strand GT] were located at positions 255-256 and CT signals [corresponding to forward strand AG] at positions 257-258). For GT/AG gain variants, sequences were extracted centered on the start coordinate of the newly created GT or AG. The predict_XGBoost.py script from the PlantCaduceus model was used to predict splice site probabilities for wild-type and mutant sequences, and the difference between them (ΔSplice probability = mutant − wild-type) was calculated to evaluate the impact of mutations on splicing. Mutations with ΔSplice probability greater than 0.1 were selected as candidate sites potentially capable of activating new splice sites. PlantCaduceus consists of four models (l20, l24, l28, l32). For each variant, the average ΔSplice probability across the four models was calculated as an integrated score. Variants with an average ΔSplice probability > 0.1 were considered candidate sites potentially creating new splice sites.

Plant Material and Growth Conditions

Seeds corresponding to target mutation sites were

obtained from the maizeEMSDB mutant library. For each line, 20 seeds were sown at room temperature and maintained under standard management practices. Whole seedlings were collected at 7 days post-sowing (seedling stage), immediately frozen in liquid nitrogen, and stored at −80°C until further use.

Genomic DNA Extraction and Genotyping

Genomic DNA was extracted from approximately 100 mg of leaf tissue using a modified CTAB method. Target mutation sites were amplified by PCR using gene-specific primers (Table S2), and the amplification products were subjected to third-generation sequencing. Genotypes (homozygous mutant, heterozygous mutant, or wild-type) were determined by aligning the sequencing reads with the reference sequence.

RNA Extraction and Splicing Variant Analysis

Total RNA was extracted from leaves of genotyped homozygous or heterozygous mutant plants using TRIzol reagent and reverse-transcribed into cDNA. RT-PCR was performed using primers spanning the splice region flanking the target site (Table S3), and amplification products were analyzed by agarose gel electrophoresis to detect splicing pattern alterations. PCR products were subsequently subjected to third-generation sequencing (PacBio platform, Tsingke Biotechnology Co., Ltd., Beijing, China) to precisely characterize the splicing variant types.

Generation of Mutant Protein Sequences and Structural Comparison

Based on the experimentally validated splicing variant types, mutant CDS sequences were obtained and translated into amino acid sequences to generate mutant proteins. Taking skl1 as an example, the protein sequence was submitted to the InterPro website for domain prediction to obtain information on conserved domain locations and functional annotations (Blum et al., 2024; Paysan-Lafosse et al., 2024). AlphaFold3 (Abramson et al., 2024) was used to predict three-dimensional structures of wild-type and mutant proteins, and the resulting PDB files were imported into PyMOL software for structural superposition and comparison. Root mean square deviation (RMSD) was calculated to evaluate the impact of mutations on protein three-dimensional structure.

CONCLUSION

We developed a workflow to identify intronic GT/AG gain mutations that affect splicing in a maize EMS mutant library, combining genome-wide screening, PlantCaduceus predictions, and experimental validation. The method successfully identified five validated splice-altering variants from seven tested candidates, giving a 71.4% validation rate. Two main observations emerged from the analysis. First, AG and GT gain mutations differ in their positional effects: AG gain mutations cluster within 50 bp upstream of the canonical AG and show a clear distance-dependent effect on splicing, while GT gain mutations do not. Second, an AGG→AAG mutation at a canonical splice acceptor site (cl30719_1) revealed competitive splicing—the mutation weakened the original AG, allowing a nearby AG in the downstream exon to be preferentially used. Protein structure analysis further showed that frameshift mutations disrupt protein function, whereas in-frame insertions may be tolerated depending on domain context. This suggests that screening for splicing mutations should consider not only the splicing event itself but also the resulting protein sequence. The pipeline and the publicly available mutant resources described here provide a practical approach for discovering and validating regulatory mutations in maize. The findings also contribute to understanding how plants regulate alternative splice site selection.

SUPPLEMENTAL DATA

Table S1. Intronic GT/AG gain variants predicted to affect splicing by PlantCaduceus.

Table S2. Primers used for genotyping of target mutation sites.

Table S3. Primers for cDNA amplification and splicing analysis.

Table S4. Database access and resource availability.

Fig. S1. Cross-model consistency analysis of PlantCaduceus predictions.

Fig. S2. Distance-grouped comparison of ΔSplice probability for AG and GT gain variants.

DECLARATION OF COMPETING INTEREST

The authors declare no conflict of interest.

REFERENCES

Abramson J, Adler J, Dunger J, et al. 2024. Accurate structure

prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**: 493-500.

Bartys N, Kierzek R, Lisowiec-Wachnicka J. 2019. The regulation properties of RNA secondary structure in alternative splicing. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, **1862**: 194401.

Blum M, Andreeva A, Florentino L C, et al. 2024. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, **53**: D444-D456.

Chao K-H, Mao A, Liu A, et al. 2025. OpenSpliceAI provides an efficient, modular implementation of SpliceAI enabling easy retraining across nonhuman species. *eLife*, **14**: RP107454.

Chao K-H, Mao A, Salzberg S L, et al. 2024. Splam: a deep-learning-based splice site predictor that improves spliced alignments. *Genome Biology*, **25**: 243.

Choi S, Cho N, Kim K K. 2023. The implications of alternative pre-mRNA splicing in cell signal transduction. *Experimental & Molecular Medicine*, **55**: 755-766.

Chong L, Su H, Liu Y, et al. 2025. Creating a gene-indexed EMS mutation library of Zheng58 for improving maize genetics research. *Theoretical and Applied Genetics*, **138**: 83.

Chua K, Reed R. 2001. An upstream AG determines whether a downstream AG is selected during catalytic step II of splicing. *Mol Cell Biol*, **21**(5): 1509-1514.

Jaganathan K, Panagiotopoulou S K, McRae J F, et al. 2019. Predicting splicing from primary sequence with deep Learning. *Cell*, **176**(3): 535-548. e24.

Lu X, Liu J, Ren W, et al. 2018. Gene-indexed mutations in maize. *Molecular Plant*, **11**(3): 496-504.

Nam K H. 2026. Evaluation of AlphaFold3 prediction for post-translational modification, oligomeric assembly, and quenchable metal binding of fluorescent proteins. *Journal of Molecular Graphics and Modelling*, **142**: 109169.

Paysan-Lafosse T, Andreeva A, Blum M, et al. 2024. The Pfam protein families database: embracing AI/ML. *Nucleic Acids Research*, **53**(D1): D523-D534.

Ramasamy P, Zuallaert J, Martens L, et al. 2026. Assessing the relation between protein phosphorylation, AlphaFold3 models, and conformational variability. *Protein Science*, **35**(1): e70376.

Scalzitti N, Kress A, Orhand R, et al. 2021. Spliceator: multi-species splice site prediction using convolutional neural networks. *BMC Bioinformatics*, **22**(1): 561.

Schaefer C, Rost B. 2012. Predict impact of single amino acid change upon protein structure. *BMC Genomics*, **13**(Suppl 4): S4.

Shepard P J, Hertel K J. 2008. Conserved RNA secondary structures promote alternative splicing. *RNA*, **14**: 1463-1469.

Smith C W, Chu T T, Nadal-Ginard B. 1993. Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Mol Cell Biol*, **13**(8): 4939-4952.

Wahl M C, Will C L, Lührmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**(4): 701-718.

Wang B, Liu J, Chen X, et al. 2025. A barley SS2a single base mutation at the splicing site leads to obvious changes in starch. *Journal of Integrative Agriculture*, **24**(4): 1359-1371.

Wilkinson M E, Charenton C, Nagai K. 2020. RNA splicing by the spliceosome. *Annual Rev Biochem*, **89**: 359-388.

Zhai J, Gokaslan A, Schiff Y, 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.

Zhan X, Lu Y, Shi Y. 2024. Molecular basis for the activation of human spliceosome. *Nature Communications*, **15**: 6348.

Zhang P, Chaldebas M, Ogishi M, et al. 2023. Genome-wide detection of human intronic AG-gain variants located between splicing branchpoints and canonical splice acceptor sites. *Proceedings of the National Academy of Sciences*, **120**(46): e2314225120.

Zhang X, Li H, Zhao C, et al. 2025. ZmMYB92 modulates secondary wall cellulose synthesis in maize. *The Plant Journal*, **122**(6): e70296.

(Managing Editor: 孙蕾)