HAYATI
Journal of Biosciences

Vol. 33 No. 5, September 2026 1261-1268
DOI:10.4308/hjb.33.5.1261-1268
ISSN: 1978-3019
EISSN: 2086-4094

Short Communication

Check for updates

OPEN ACCESS

Comparative Analysis of the Splice Sites of Different Plant Genomes

Ulkar Huseynova Mustafaeva[1], Orkhan Mustafaev[1,2], Zumrud Abaszade[3], Shader Alizade[4*], Basti Aghayeva[1], Afat Mammadova[2]

[1]Genetic Resources Institute of Ministry of Science and Education of Republic of Azerbaijan, Azadlig, Ave.155, AZ1106, Baku, Azerbaijan
[2]Baku State University, Z.Khalilovstr. 23, AZ1073, Baku, Azerbaijan
[3]Department of Normal Physiology, Azerbaijan Medical University, Samad Vurgun 163A, AZ1022, Baku, Azerbaijan
[4]Western Caspian University, Istiglaliyyat Street 31, AZ1001, Baku, Azerbaijan

ARTICLE INFO

Article history:
Received October 14, 2025
Received in revised form February 4, 2026
Accepted March 4, 2026
Available Online July 20, 2026

KEYWORDS:
exon,
gene regulation,
intron,
pre-mRNA,
splice sites

ABSTRACT

Alternative splicing is a key mechanism that enhances transcriptome and proteome diversity in eukaryotes. While canonical GT–AG splice sites have been extensively characterized, non-canonical combinations remain poorly understood, especially in plants. Previous studies have mostly focused on model species and major crops, leaving many plant lineages unexplored. In this study, we conducted a comparative analysis of splice-site usage across 11 plant species representing diverse crops. Using high-quality RefSeq annotations, we extracted all introns within coding regions and quantified the frequencies and diversities of canonical and non-canonical splice sites. Normalized intron counts revealed that the relative proportions of non-canonical classes are stable across well-annotated genomes, while their diversity increases with the total number of introns. We further observed that non-canonical splice sites are more frequent in introns of extreme lengths, suggesting that intron architecture may influence splicing precision. A small number of recurrent non-canonical combinations (GA–AG, GT–AA, GT–AT, and GT–GG) were consistently detected across species, indicating that these motifs are evolutionarily tolerated and may arise through single-nucleotide substitutions. These results indicate that non-canonical splicing is an evolutionarily conserved feature of plant transcriptomes and provide new insights into splice-site variability beyond model species.

1. Introduction

Splicing represents a fundamental stage in the pathway of gene expression and constitutes a crucial point of its regulation (Alizada 2022; Tyurin *et al*. 2022; Mustafayeva *et al*. 2025). The removal of intronic regions from precursor messenger RNA (pre-mRNA) molecules and the subsequent joining of exons occur through a complex and highly coordinated mechanism mediated by the spliceosome - a dynamic ribonucleoprotein assembly that forms on pre-mRNA substrates (Deyneko *et al*. 2022; Suhorukova *et al*. 2022; Alizade 2023; Alizada & Aliyeva 2024). The spliceosome is sequentially assembled from small nuclear ribonucleoprotein particles (snRNPs) together with a range of non-snRNP splicing factors (Brown & Simpson 1998). Two principal classes of spliceosomes have been identified: the U2- and U12-dependent types, which share similar overall functions but differ in their protein composition (Sharp & Burge 1997; Turunen *et al*. 2013). Both systems recognize highly conserved sequence motifs within introns—typically the canonical GU dinucleotide at the 5′ splice site, a branchpoint adenosine located several nucleotides

*Corresponding Author
E-mail Address: shader622@mail.ru

upstream of the 3′ end, and the AG dinucleotide at the 3′ splice site (Alberts *et al*. 2015). These positions are respectively termed the donor (5′) and acceptor (3′) splice sites (Sanjaya *et al*. 2024). Although these terminal dinucleotides are indispensable for splicing, they alone cannot specify which spliceosomal pathway—major or minor—processes a particular intron (Dietrich *et al*. 1997).

The majority of canonical GT–AG and non-canonical GC–AG introns are processed by the major U2 spliceosome (Parada *et al*. 2014). Genome-wide transcriptomic analyses across a range of taxa, including *Arabidopsis thaliana*, various plant species, and metazoans, demonstrate that approximately 98% of introns conform to the GT–AG type (Cheng *et al*. 2023). Two minor variants of U2-type splice sites are also recognized: GC–AG sites (~0.9%) and the exceedingly rare AT–AC sites, represented by only a few dozen introns in the human genome (Sheth *et al*. 2006; Roca *et al*. 2012).

Despite the fact that mutations in splice-site dinucleotides often disrupt splicing, many introns with non-canonical boundaries are efficiently removed (Parada *et al*. 2014). In most cases, such sites retain similarity to the consensus sequences of U2- or U12-type introns, which enables their recognition by the spliceosome. For example, evolutionarily conserved U2-like introns with GA–AG splice sites, as well as functional GT–TG sites, have been described in various genes, highlighting their regulatory and functional significance (Parada *et al*. 2014). These observations suggest that splice-site strength represents a continuum and that variation in splice-site sequences can contribute to the flexibility of alternative splicing and the generation of novel transcript isoforms.

Moreover, changes in splice-site strength may influence not only the efficiency of intron removal but also the selection of alternative pre-mRNA processing pathways. In particular, it has been demonstrated that canonical splicing and back-splicing compete for the use of the same 5′ and 3′ splice sites: the presence of strong splice sites favors the formation of linear transcripts, whereas their weakening increases the likelihood of circular RNA production (Ashwal-Fluss *et al*. 2014; Pitolli *et al*. 2022). This example illustrates that variation in splice-site sequences can alter the balance between different RNA isoforms and, consequently, affect protein synthesis and gene expression regulation.

The AT–AC introns are excised by a distinct, though related, spliceosomal system composed of U11, U12, U4atac/U6atac, and U5 snRNPs (Burset *et al*. 2000), rather than the canonical U1, U2, U4/U6, and U5 snRNPs that mediate GT–AG splicing.

Although non-canonical splice sites have been documented in diverse organisms, their biological significance and evolutionary persistence in plants remain poorly characterized. Earlier analyses of atypical splice junctions in fungi and animals have typically targeted individual species (Frey and Pucker 2020), and many investigations have emphasized canonical GT–AG sites while overlooking non-canonical variants (Modrek *et al*. 2001).

Given that most previous work has focused on model or agriculturally important crops, numerous plant lineages have been underexplored in this context. In the present study, we conducted a systematic comparative analysis of splice-site usage across 11 plant species representing distinct phylogenetic clades. Species selection was based on two main criteria: (i) availability of high-quality reference genome annotations, and (ii) the fact that these reference genomes were assembled or substantially updated only recently (Table 1). This cross-species framework allowed us to assess the frequency, diversity, and conservation of non-canonical splice sites, thereby providing new insights into the evolutionary dynamics and variability of splicing mechanisms in plants.

2. Materials and Methods

Reference genome annotations for the 11 selected plant species were retrieved from the NCBI RefSeq repository (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/). The selection of species was based on the availability of recently assembled, high-quality reference genomes not included in earlier comparative studies of splice-site usage.

Based on these annotations, all introns entirely contained within coding sequences (CDS) were extracted for downstream analysis. For each identified intron, the terminal dinucleotides at the 5′ and 3′ splice junctions were determined. Data extraction and processing were conducted using a combination of custom Python and PHP scripts, after which a comprehensive summary matrix was generated. In this matrix, rows represented the terminal dinucleotide combinations (e.g., GT–AG), columns denoted the corresponding species, and cell values indicated the total number of introns belonging to each category.

For every species, intron counts were normalized by the total number of CDS-associated introns. Several summary metrics were then calculated, including:

• The proportions of canonical versus non-canonical introns;
• The absolute abundance of non-canonical introns;
• The diversity index, defined as the number of unique terminal combinations observed at least once;
• The relative fraction of non-canonical introns with respect to the overall total.

To evaluate how these parameters correlated with the total intron count, the Spearman rank correlation coefficient (ρ) along with corresponding p-values was computed. The analyses were conducted for both the complete set of non-canonical splice sites (including GC–AG and AT–AC) and, separately, for the minor non-canonical categories (i.e., all terminal combinations other than GT–AG, GC–AG, and AT–AC). For each species, a vector of relative frequencies representing all observed terminal dinucleotide combinations was generated. Pairwise Spearman correlations among these vectors were then used to construct a similarity matrix, which was subsequently visualized as a heatmap to illustrate interspecies relationships in splice-site usage patterns.

All computational analyses were implemented in Python 3.10 (https://www.python.org/downloads/release/python-3100/) utilizing the pandas, numpy, scipy (spearmanr), and matplotlib libraries. Additional PHP scripts were used during the preliminary stage to extract intron sequences from annotated genomic data. All scripts used for data extraction and analysis are publicly available at https://github.com/bioset-org/canonic_intron.

3. Results

A comprehensive examination of 11 plant genomes revealed a consistent distribution of splice-site classes across species. In total these annotations contain 2.824.011 coding-sequence introns. Across this set the sum of detected non-canonical introns is 89.350, yielding a weighted non-canonical fraction of 3.16% (canonical splice sites ≈ 96.84%). The per-species proportion of canonical (GT–AG) splice sites ranges from 96.28% to 97.36% (Table 2). The mean non-canonical fraction across species is ~3.23%.

Table 1. Information of the samples used and the number of introns analyzed per sample

Spesies	Family	Clade	Assembly	RefSeq	Total introns
Arachis hypogaea	Fabaceae	Eudicot	arahy.Tifrunner.gnm2.J5K5	GCF_003086295.3	351.561
Arachis stenosperma	Fabaceae	Eudicot	arast.V10309.gnm1.PFL2	GCF_014773155.1	159.003
Coffea arabica	Rubiaceae	Eudicot	Coffea Arabica ET-39 HiFi	GCF_036785885.1	274.499
Coffea eugenioides	Rubiaceae	Eudicot	Ceug_1.0	GCF_003713205.1	132.868
Hordeum vulgare	Poaceae	Monocot	MorexV3_pseudomolecules_assembly	GCF_904849725.1	145.956
Lolium perenne	Poaceae	Monocot	Kyuss_2.0	GCF_019359855.2	202.166
Lolium rigidum	Poaceae	Monocot	APGP_CSIRO_Lrig_0.1	GCF_022539505.1	233.944
Triticum aestivum	Poaceae	Monocot	IWGSC CS RefSeq v2.1	GCF_018294505.1	545.841
Triticum dicoccoides	Poaceae	Monocot	WEW_v2.1	GCF_002162155.2	351.023
Triticum urartu	Poaceae	Monocot	Tu2.1	GCF_003073215.2	185.707
Vicia villosa	Fabaceae	Eudicot	Vvil1.0	GCF_029867415.1	241.443

Table 2. Summary of canonical and non-canonical splice-site proportions across 11 plant genomes

RefSeq	Canonical (GT–AG), %	Non-canonical count	Non-canonical, (%)
GCF_003086295.3	97.27	9.600	2.73
GCF_014773155.1	97.19	4.474	2.81
GCF_036785885.1	96.92	8.459	3.08
GCF_003713205.1	96.31	4.900	3.68
GCF_904849725.1	96.35	5.318	3.64
GCF_019359855.2	96.28	7.510	3.71
GCF_022539505.1	96.29	8.679	3.70
GCF_018294505.1	96.94	16.716	3.06
GCF_002162155.2	96.75	11.414	3.25
GCF_003073215.2	96.82	5.900	3.17
GCF_029867415.1	97.36	6.376	2.64

In our study, Spearman correlation was selected because it provides a quantitative assessment of similarity or dissimilarity in the structure of relative splice-site frequency profiles between species (Pucker and Brockington 2018). Correlation analysis of the splice-site frequency vectors across species (Spearman rank correlation) reveals consistently high similarity in the relative usage of splice-site classes: the mean pairwise Spearman ρ is ~0.657 (median ρ ≈ 0.659). All pairwise correlations among these 11 species are highly significant (pairwise p-values < 0.001). The heatmap of Spearman correlations (Figure 1) shows predominantly warm colors, reflecting the conserved ordering of classes (GT–AG >> GC–AG >> other minor combinations) across genomes.

Correlation analysis showed a strong positive relationship between total intron number and the absolute count of non-canonical introns ($\rho \approx 0.92, p < 10^{-4}$), whereas the proportion of non-canonical sites was uncorrelated or slightly negatively associated with annotation size ($\rho \approx -0.45, p \approx 0.17$). Similar trends were observed for minor non-canonical classes (Table 3).

To compare the length distributions of introns with canonical and non-canonical splice site combinations, we combined all introns from the 11 analyzed species (Figure 2). The resulting distributions are largely similar across most regions but differ noticeably at both extremes of the length range. Introns with non-canonical splice sites (NCSS) are more frequent among very short introns (<100 bp) and very long introns (>5 kb), whereas canonical introns (CSS) dominate in the central region around 200 bp, corresponding to the most common intron length in plants. Long introns are hotspots for alternative splicing. Non-canonical splice sites (NCSS) are generally considered weaker than canonical splice sites (CSS), making them more susceptible to regulatory modulation. By harboring a non-canonical site, the cell

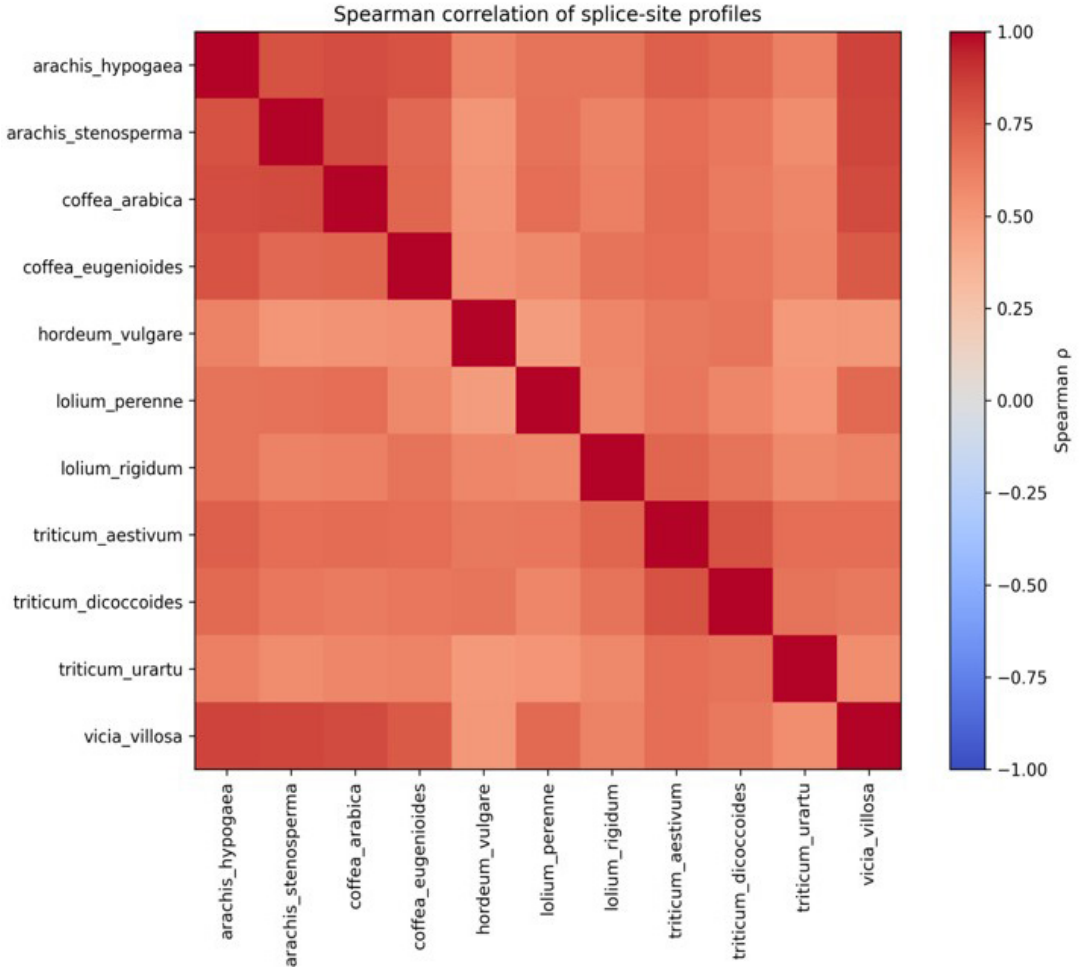

Figure 1. Spearman correlation matrix of splice-site frequency profiles across 11 plant species

Table 3. Spearman correlation of intron number with occurrence, diversity, and proportion of non-canonical splice sites

Dataset	Parameter compared with total introns	Spearman's ρ	p-value
All Non-canonical	Occurrences (absolute count)	0.918	6.66×10^{-5}
	Types (distinct combinations)	0.301	0.368
	Fraction (relative proportion)	–0.445	0.170
Minor Non-canonical	Occurrences (absolute count)	0.927	3.97×10^{-5}
	Types (distinct combinations)	0.301	0.368
	Fraction (relative proportion)	–0.309	0.355

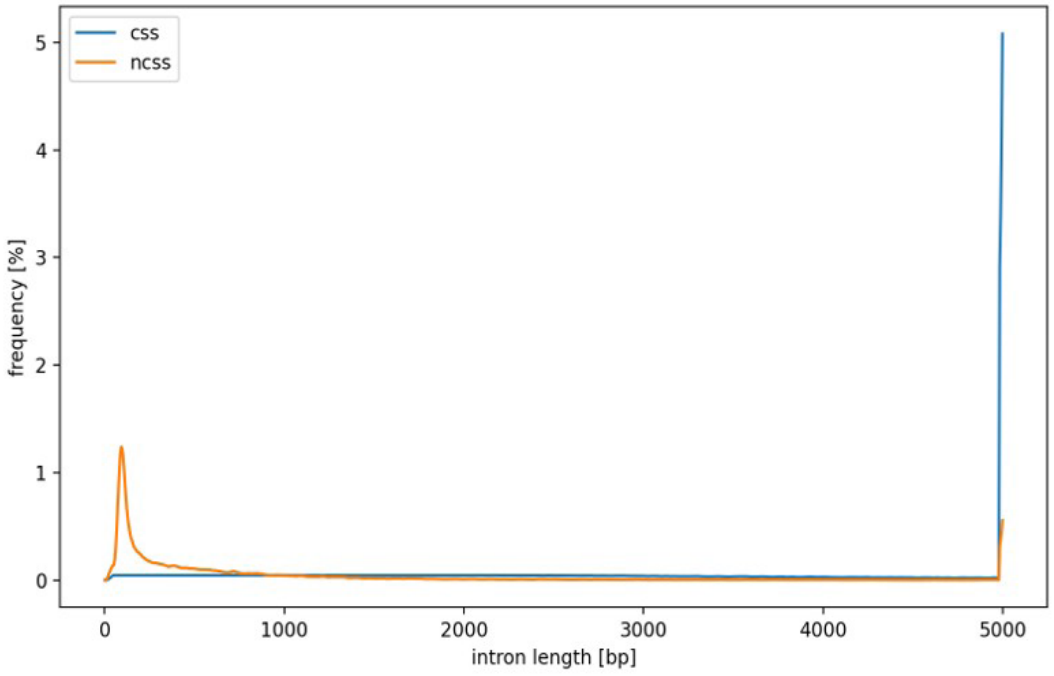

Figure 2. Comparison of intron length distributions between canonical and non-canonical splice sites

may gain additional flexibility in splice-site selection, allowing exon inclusion or skipping to respond to specific cellular or environmental cues. The predominance of CSS at ~200 bp likely reflects structural constraints on exon definition. Spliceosome assembly is strongly influenced by the spatial organization of exons and introns, and experimental evidence shows that spliceosome formation is strongly inhibited when exon lengths are artificially expanded beyond ~300 nucleotides (De Conti *et al*. 2013). This observation highlights the importance of optimal spacing between splice sites for efficient recognition. At intermediate lengths, where splice-site pairing is most efficient, there is strong selective pressure to maintain the canonical GT–AG consensus to ensure rapid and accurate processing (Sibley *et al*. 2016).

Although both CSS and NCSS classes display largely similar overall shapes, the NCSS curve is slightly elevated around the modal intron length (~200 bp), suggesting a modest enrichment of non-canonical splice sites in introns of typical size. Both distributions show additional divergence at the extremes — NCSS are more common among very short introns (<100 bp) and somewhat more frequent in very long introns (>5 kb), indicating a broader overall length distribution.

In our study, the intron length distribution appears highly uneven because this pattern is not an analytical artifact but a well-known architectural feature of plant genomes described in previous studies (Pucker and Brockington 2018; Thakur *et al*. 2025). Across all analyzed species, the same overall distribution shape is preserved, characterized by a pronounced peak of short introns and a long right-skewed tail, indicating the universality of this pattern. In this study, introns were pooled across all species to highlight the general trend and to compare the distributions of canonical and non-canonical splice sites, rather than to perform interspecies comparisons of intron length.

To examine which minor non-canonical splice-site variants are most common in plants, we compared the frequency of individual dinucleotide pairs across the 11

analyzed genomes (Figure 3). The four most recurrent non-canonical combinations (GA-AG, GT-AA, GT-AT, and GT-GG) were detected across all species, with median counts of approximately 10-40 introns per genome. Their relatively high frequencies suggest that these splice-site motifs arise repeatedly and are tolerated by the splicing machinery. The "others" category, which aggregates numerous rare dinucleotide combinations, displayed the widest dispersion and several high-value outliers, indicating substantial heterogeneity among infrequent site types. Overall, these data show that plant genomes share a conserved set of dominant minor splice-site classes, while additional rare variants occur sporadically and are likely species-specific.

4. Discussion

The stable pattern of splice-site usage across different plant genomes shows that the main rules of splice-site recognition are highly conserved. Although plant species differ in genome size and intron number, the proportions of canonical and non-canonical splice sites stay nearly the same. This suggests that non-canonical splicing is a regular and evolutionarily conserved part of gene expression in plants. The strong correlation between the total intron count and the absolute number of non-canonical introns reflects scaling effects rather than genuine enrichment, while the lack of association between non-canonical fraction and annotation size supports the robustness of these proportions across datasets.

Previous large-scale surveys have reported similar distributions: Pucker and Brockington (2017) found approximately 98.7% GT–AG introns in plants and 99.2% in mammals, highlighting the slightly higher prevalence of GC–AG and other minor combinations in plants. This greater diversity is consistent with the complex regulatory architecture of plant transcriptomes, where alternative splicing and RNA surveillance play central roles in environmental adaptation (Reddy *et al*. 2013).

Spearman is a nonparametric method that does not require assumptions about the data distribution and is suitable for comparing ranked profiles of relative frequencies (Pucker and Brockington 2018). Correlation analysis of splice-site frequency profiles across species revealed high similarity in the relative ordering of classes, underscoring the evolutionary conservation of the splicing hierarchy (GT–AG >> GC–AG >> minor types). Such conservation implies strong selective constraints acting on core spliceosomal recognition signals. However, non-canonical sites themselves appear to follow a dual evolutionary trajectory. Some of them are conserved and functionally relevant—maintained

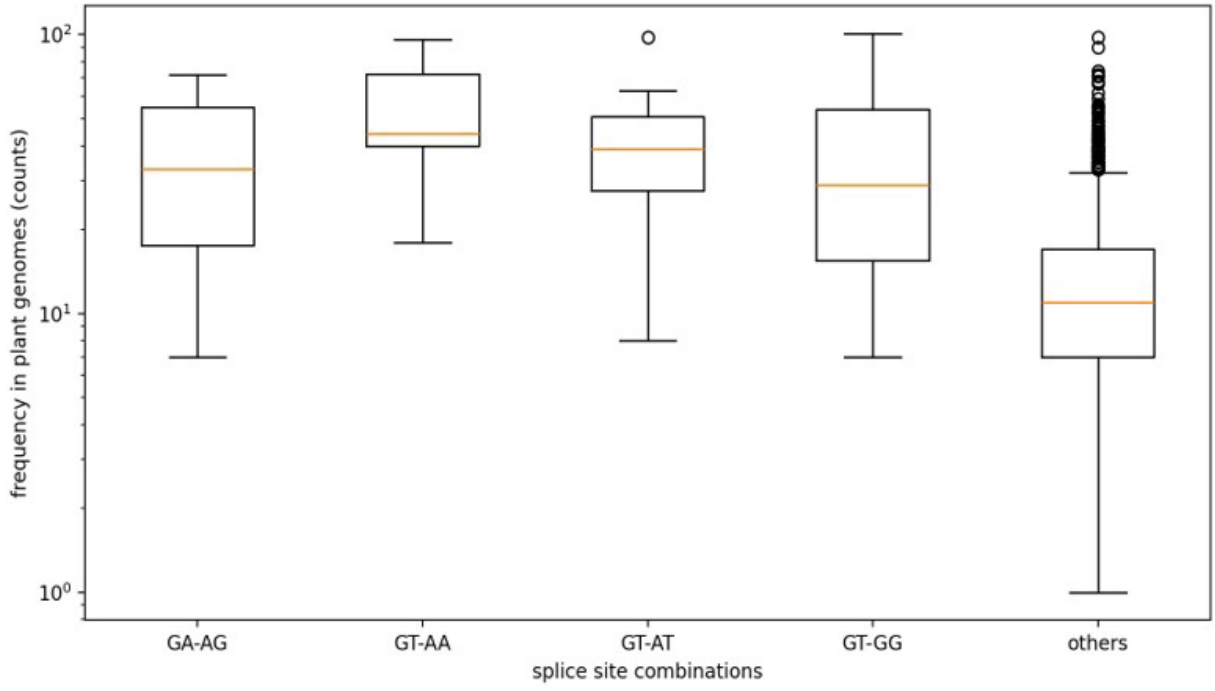

Figure 3. Occurrence of recurrent and rare non-canonical splice-site combinations in plant genomes

by selection due to specific roles in gene regulation or transcript processing (Rogozin *et al*. 2003; Pucker *et al*. 2017)—whereas others are likely transient, arising through local mutations or transcriptional noise and eventually eliminated from the genome.

Methodologically, the stability of non-canonical splice-site proportions has important implications. The observed independence from annotation completeness indicates that differences across datasets are primarily determined by annotation quality rather than size. Many non-canonical and cryptic introns remain undetected because they are weakly expressed or subject to post-transcriptional quality-control pathways, such as nonsense-mediated decay (NMD) and nuclear degradation. Standard RNA-seq protocols, typically biased toward poly(A)+ RNA and long exons, underrepresent these transcripts. As noted by Sibley *et al*. (2016), activation of cryptic splice sites can introduce premature stop codons and trigger NMD, explaining their systematic underrepresentation in most transcriptome annotations.

In addition to frequency patterns, the analysis of intron length distributions revealed that splice-site composition is not uniform across the intron size range. Non-canonical splice sites were more common among very short (< 100 bp) and very long (> 5 kb) introns, whereas canonical sites dominated around the modal intron length (~ 200 bp). This trend suggests that extreme intron sizes may reduce splicing precision and favor the usage of alternative recognition motifs. Similar tendencies have been described in previous studies, in which long or structurally complex introns were associated with reduced spliceosomal fidelity and an increased occurrence of non-canonical junctions (Burset *et al*. 2000; Parada *et al*. 2014). Such length-dependent variability implies that intron architecture itself can influence splice-site selection, potentially contributing to the emergence and maintenance of rare non-canonical events in plant genomes.

Analysis of minor non-canonical splice-site combinations revealed a conserved subset of recurrent variants (GA–AG, GT–AA, GT–AT, and GT–GG) shared across all examined plant species. Their repeated occurrence in independent genomes suggests that these alternative dinucleotide pairs can arise frequently through single-nucleotide mutations in canonical GT–AG sites, as previously noted for similar cases in plants (Pucker and Brockington 2018) and mammals (Burset *et al*. 2000). The spliceosome can correctly recognize splice sites that slightly differ from the canonical GT–AG sequence when supported by nearby enhancer elements or favorable snRNA interactions (Roca *et al*. 2013; Sibley *et al*. 2016). Most of the very rare variants, however, appear transient or artifactual and typically lack evolutionary conservation (Rogozin *et al*. 2003). Together, these observations indicate that a small number of minor non-canonical motifs are stably maintained across plants, while many others reflect sporadic splicing noise or annotation artifacts.

Overall, these findings demonstrate that non-canonical splicing is a reproducible and conserved phenomenon across plant species. While its absolute frequency scales with genome complexity, its relative proportion remains constant, pointing to deep evolutionary and mechanistic constraints. Expanding analyses beyond model and agronomically important plants will help clarify the functional relevance of rare splice-site variants and their contribution to transcriptome plasticity in diverse plant lineages.

Acknowledgements

Authors thanks all the teams who have supported this research.

References

Alizada, S., 2022. Role of miRNAs in cotton salt stress responses. Advances in Biology & Earth Sciences. 7, 80-84.

Alizade, Sh.A., 2023. Evaluation of the GhMAPK gene expression level under salt stress in cotton cultivars. *Plant Biotechnology and Breeding*. 6, 40-47. https://doi.org/10.30901/2658-6266-2023-4-o6

Alizada, S., Aliyeva, K., 2024. Comparative analysis of expression of antiporter encoding gene (GhNHX1) under different concentrations of NaCl in cotton (*Gossypium hirsutum* L.). *Advances in Biology & Earth Sciences*. 9, 168-174. https://doi.org/10.62476/abes9168

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., 2015. *Essential Cell Biology*, fourth ed. Garland Science, New York

Ashwal-Fluss, R., Meyer, M., Pamudurti, N.R., Ivanov, A., Bartok, O., Hanan, M., Evantal, N., Memczak, S., Rajewsky, N., Kadener, S., 2014. circRNA biogenesis competes with pre-mRNA splicing. *Molecular Cell*. 56, 55-66. https://doi.org/10.1016/j.molcel.2014.08.019

Brown, J.W.S., Simpson, C.G., 1998. Splice site selection in plant PRE-mRNA splicing. *Annu. Rev. Plant Physiol. Plant Mol. Biol*. 11, 77-95.

Burset, M., Seledtsov, I.A., Solovyev, V.V., 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*. 28, 4364–4375. https://doi.org/10.1093/nar/28.21.4364

Cheng, W., Hong, C., Zeng, F., Liu, N., Gao, H., 2023. Sequence variations affect the 5′ splice site selection of plant introns. *Plant Physiology*. 193, 1281-1296. https://doi.org/10.1093/plphys/kiad375

De Conti, L., Baralle, M., Buratti, E., 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdisciplinary Reviews: RNA*. 4, 49-60. https://doi.org/10.1002/wrna.1140

Deyneko, I.V., Mustafaev, O.N., Tyurin, A.A., Zhukova, K.V., Varzari, A., Goldenkova-Pavlova I.V., 2022. Modeling and cleaning RNA-seq data significantly improve detection of differentially expressed genes. *BMC bioinformatics*. 23, 1-14. https://doi.org/10.1186/s12859-022-05023-z

Dietrich, R.C., Incorvaia, R., Padgett, R.A., 1997. Terminal intron dinucleotide sequences do not distinguish between U2-and U12-dependent introns. *Mol. Cell*. 1, 151-160. https://doi.org/10.1016/S1097-2765(00)80016-7

Frey, K., Pucker, B., 2020. Animal, Fungi, and Plant Genome Sequences Harbor Different Non-Canonical Splice Sites. *Cells*. 9, 1-19. https://doi.org/10.3390/cells9020458

Mustafayeva, U.H., Mustafayev, O., Alizade, Sh., Hasanli, L., Qasimli, B., Mammadova, A., 2025. Comparative analysis of splice site usage across plant genomes. *Baku State University Journal of Life Sciences & Biology*. 2, 1-6 https://doi.org/10.30546/300045.2025.2.4.3001

Modrek, B., Resch, A., Grasso, C., Lee, C., 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*. 29, 2850-2859. https://doi.org/10.1093/nar/29.13.2850

Parada, G.E., Munita, R., Cerda, C.A., Gysling, K., 2014. A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Research*. 42, 10564-10578. https://doi.org/10.1093/nar/gku744

Pitolli, C., Marini, A., Sette, C., Pagliarini, V., 2022. Non-canonical splicing and its implications in brain physiology and cancer. *International Journal of Molecular Sciences*. 23, 2811. https://doi.org/10.3390/ijms23052811

Pucker, B., Holtgräwe, D., Weisshaar, B., 2017. Consideration of non-canonical splice sites improves gene prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence. *BMC Res Notes*. 10, 1-6. https://doi.org/10.1186/s13104-017-2985-y

Pucker, B., Brockington, S.F., 2018. Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. *BMC Genomics*. 19, 1-13. https://doi.org/10.1186/s12864-018-5360-z

Reddy, A.S., Marquez, Y., Kalyna, M., Barta, A., 2013. Complexity of the alternative splicing landscape in plants. *Plant Cell*. 25, 3657-3683. https://doi.org/10.1105/tpc.113.117523

Roca, X., Akerman, M., Gaus, H., Berdeja, A., Bennett, C.F., Krainer, A.R., 2012. Widespread recognition of 5' splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides. *Genes Dev*. 26, 1098-1109. https://doi.org/10.1101/gad.190173.112

Roca, X., Krainer, A.R., Eperon, I.C., 2013. Pick one, but be quick: 5' splice sites and the problems of too many choices. *Genes Dev*. 27, 129-44. https://doi.org/10.1101/gad.209759.112

Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G., Koonin, E.V., 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*. 13, 1512-1517. https://doi.org/10.1016/S0960-9822(03)00558-X

Sanjaya, A., Nishijima, R., Fujii, Y., Asano, M., Ishii, K., Kazama, Y., Abe, T., Fujiwara, M.T., 2024. Rare occurrence of cryptic 5’ splice sites by downstream 3’ splice site/exon boundary mutations in a heavy-ion-induced egy1-4 allele of *Arabidopsis thaliana*. *Front. Plant Sci*. 15, 1-9. https://doi.org/10.3389/fpls.2024.1388040

Sharp, P.A., Burge C.B., 1997. Classification of introns: U2-type or U12-type. *Cell*. 91, 875-879. https://doi.org/10.1016/S0092-8674(00)80479-1

Suhorukova, A.V., Tyurin, A.A., Pavlenko, O.S., Mustafayev, O.N., Sinelnikov, I.G., Goldenkova-Pavlova, I.V., 2022. Development of dual reporter vector system for estimating translational activity of regulatory elements. *BMC Plant Biology*. 22, 1-10. https://doi.org/10.1186/s12870-022-03735-1

Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., Sachidanandam, R., 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Research*. 34, 3955-3967

Sibley, C.R., Blazquez, L., Ule, J., 2016. Lessons from non-canonical splicing. *Nat Rev Genet.* 17, 407-421. https://doi.org/10.1038/nrg.2016.46

Thakur, K.P., Butenko, A., Karásek, F., Svobodová, M., Faktorová, D., Pavlisková, H., Varga, V., Aleš Horák, A., Lukeš, J., Staněk, D., 2025. Splice site diversity and abundance of noncanonical introns in diplonemids (Diplonemea, Euglenozoa). *RNA*. 31, 1826-1840.

Turunen, J.J., Niemelä, E.H., Verma, B., Frilander, M.J., 2013. The significant other: Splicing by the minor spliceosome. Wiley Interdiscip. *Rev. RNA*. 4, 61-76. https://doi.org/10.1002/wrna.1141

Tyurin, A.A., Mustafaev, O., Suhorukova, A.V., Pavlenko, O.S., Fridman, V.A., Demyanchuk, I.S., Irina V Goldenkova-Pavlova I.V., 2022. Modulation of the translation efficiency of heterologous mRNA and target protein stability in a plant system: the case study of interferon-αA. *Plants*. 11, 1-19. https://doi.org/10.3390/plants11192450