Baku State University Journal of Life Sciences & Biology ***2025****, v.2 (4), p. 1-6*

journal homepage: http://bsuj.bsu.edu.az/en

https://doi.org/10.30546/300045.2025.2.4.3001

COMPARATIVE ANALYSIS OF SPLICE SITE USAGE ACROSS PLANT GENOMES

Ulkar Huseynova Mustafayeva[1]*, Orxan Mustafayev[1,2], Shader Alizade[1,2], Latifa Hasanli[1], Basti Qasimli[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, Baku State University, 23 Z. Khalilov St, AZ-1073, Baku, Azerbaijan

Received 07october 2025; accepted 15 november 2025

Abstract

Alternative splicing is a fundamental mechanism that contributes to transcriptome and proteome diversity in eukaryotes. While canonical GT–AG splice sites have been extensively studied, non-canonical splice site combinations remain less understood, particularly in plants. Most research has focused on model organisms or major crops, leaving many plant lineages insufficiently characterized. Here, we present a comparative analysis of splice site usage across 13 plant species representing diverse taxonomic groups. Using high-quality RefSeq annotations, we extracted all introns located within coding regions and quantified the frequency and diversity of canonical and non-canonical splice sites. Normalization of intron counts revealed that the relative proportions of non-canonical classes remain stable across well-annotated genomes, whereas their diversity scales with the total intron number. These findings highlight the evolutionary persistence of non-canonical splicing and provide new insights into splice site variability in plants, broadening the current view of splicing beyond model and agronomically important species.

Keywords: *splicing; RNA; gene regulation; splice sites*

1. Introduction

The genomes of all eukaryotes contain introns, but their number, size, and distribution vary considerably between different species. The correct identification and removal of introns by the splicing machinery is a central, conserved step during gene expression in all eukaryotes, and mutations that alter the sequence of splice sites or elicit splicing errors are often associated with disease. Furthermore, the noncontinuous exon–intron structure of eukaryotic genes allows the formation of alternative mRNA isoforms [1, 2, 5]. Transcription and pre-mRNA splicing are extremely complex multimolecular processes that involve protein–DNA, protein–RNA, and protein–protein interactions [3, 9]. This process allows variation, which provides the basis for quick adaptation to changing conditions. Alternative splicing, e.g., skipping exons, usage of alternative 50 or 30 splice sites, and the retention of introns, results in an enormous diversity of synthesized proteins and, therefore, substantially expands the diversity of products encoded in eukaryotic genomes. The occurrence of alternative splicing provides exciting new possibilities for gene regulation and is responsible for the remarkable

transcriptome and proteome diversity in metazoans [4, 7]. Alternative splicing is regulated by interactions of RNA-binding proteins (RPBs) and splicing factors with sequences in the pre-mRNA, and by base-pairing between complementary RNA sequences in cis and in trans [7, 8, 9].

A critical step in pre-mRNA splicing is the recognition and pairing of 5`- and 3`-splice sites [10]. Splice sites are recognized during the splicing process by a complex of small nuclear RNAs (snRNAs) and proteins: the spliceosome. The spliceosome exists in two main forms, the U2-type and the U12-type, each made up of distinct but functionally comparable protein components. While the terminal dinucleotides play a key role in splicing, they alone do not determine which spliceosome acts on a given intron. Generally, canonical GT-AG introns and most GC-AG variants are processed by the U2 spliceosome. In contrast, the U12 spliceosome usually handles AT-AC introns, though certain AT-AC type II introns are instead processed by the U2 system. This highlights the intricate nature of splicing, which relies on additional sequence cues within the genome [11]. Generally, GC-AG and AT-AC are classified as major non-canonical splice site combinations, while all deviations from these sequences are deemed to be minor non-canonical splice sites [12]. U2-type introns with GA–AG splice junctions, which are evolutionarily conserved, have been reported in FGFR genes, while a functional GT–TG splice site has been discovered in the GNAS gene [15, 16, 17]. A few very uncommon intron end sequences have been detected, and they are frequently considered possible sequencing artifacts, such as introns ending with GT-GG or TT-AG [13].

Although several models have been proposed, the precise mechanism behind non-canonical splicing remains unclear. Earlier investigations into non-canonical splice sites were typically limited to one species or a few species. Many of these studies concentrated primarily on the common GT-AG splice sites, while giving little attention to the non-canonical variants.

Non-canonical splice sites have been reported in various organisms, yet their functional importance and evolutionary conservation in plants are still not fully understood [14]. While research has primarily concentrated on model and agronomically important crops, many other plant groups have rarely been considered in this context. In this study, we systematically investigated splice site usage across 13 plant species spanning different lineages. This comparative framework enabled us to quantify the frequency, diversity, and conservation of non-canonical splice sites, thereby expanding our understanding of splicing variation across plants.

2. Materials and Methods

Reference gene annotations for 13 plant species were obtained from the NCBI RefSeq database (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/). From these annotations, all introns fully located within coding sequences (CDS) were extracted. For each intron, the two nucleotides at the 5′ and 3′ ends were identified. These data were retrieved using custom Python and PHP scripts, after which a summary table was constructed, with rows corresponding to terminal dinucleotide combinations (e.g., GT–AG), columns to species, and cell values representing the number of introns of each type.

For each species, the counts of introns in each class were normalized by the total number of CDS introns. The following measures were calculated: the proportion of canonical and non-canonical introns; the absolute number of non-canonical introns; the number of distinct combinations observed at least once (diversity); the fraction of non-canonical introns relative to the total.

To assess the relationship of these measures with the total number of introns, the Spearman rank correlation coefficient (ρ) with associated p-values was used. Analyses were performed both for all non-canonical splice sites (including GC–AG and AT–AC) and separately for “minor” combinations (all except GT–AG, GC–AG, and AT–AC). For each species, a vector of relative frequencies of all terminal combinations was constructed. Pairwise Spearman correlations between these vectors were used to generate a similarity matrix, which was visualized as a heatmap.

All analyses were performed in Python 3.10 (https://www.python.org/downloads/release/python-3100/) using the libraries *pandas*, *numpy*, *scipy (spearmanr)*, and *matplotlib*. Additional PHP scripts were used for the initial extraction of intron sequences.

3. Results and Discussion

Our analysis of 13 species, comprising >2.2 million CDS introns (on average ~171,000 per species), shows that the canonical GT–AG splice site dominates (96.6%). The major non-canonical classes GC–AG and AT–AC together account for ~1.5%, whereas other minor combinations contribute ~1.8%. The absolute number of non-canonical introns strongly scales with the total number of annotated introns (Spearman's $\rho = 0.879$, $p = 7.25\times10^{-9}$), but their proportion relative to all introns does not depend on annotation size ($\rho = -0.208$, $p = 0.319$) (Table 1). This indicates that the increase in non-canonical introns primarily reflects scale rather than enrichment. Diversity (the number of distinct terminal dinucleotide combinations observed at least once) moderately increases with annotation size ($\rho = 0.614$, $p = 1.08\times10^{-3}$), consistent with the expected sampling effect: rare splice site types are more likely to be detected in larger datasets. The same trends hold when considering only minor classes: the absolute number is correlated with annotation size ($\rho = 0.786$, $p = 3.17\times10^{-6}$), whereas the proportion shows no dependence ($\rho = -0.033$, $p = 0.875$).

Table1. Correlation of intron number with non-canonical splice-site usage across 13 plant species

Dataset	Metric	Spearman ρ	p-value
Noncanonical all	Occurrences vs total introns	0.879	7.25e-09
	Types vs total introns	0.614	1.08e-03
	Fraction vs total introns	–0.208	3.19e-01
Noncanonical minor	Occurrences vs total introns	0.786	3.17e-06
	Types vs total introns	0.614	1.08e-03
	Fraction vs total introns	–0.033	8.75e-01

Our results are consistent with broader surveys: Pucker & Brockington (2017) reported ~98.7% GT–AG introns in plants, compared to ~99.2% in mammals. Notably, GC–AG occurs almost twice as frequently in plants, and minor non-canonical combinations are also more common (0.09% versus 0.02% in mammals). Taken together, these observations indicate that both major and minor non-canonical classes represent a more prominent component of splicing in plants than in animals, reflecting the greater complexity of the plant splicing landscape [18].

Correlation analysis of frequency vectors across all splice site combinations further highlights the conserved ranking of site usage. The Spearman correlation matrix shows predominantly high positive values: the heatmap is dominated by warm colors with no pronounced blue regions (Fig.1). This indicates that the relative ordering of splice site classes is stable (GT–AG ≫ GC–AG ≫ others), even when certain rare combinations are absent in individual species.

Despite well-known sources of annotation and mapping artifacts [19, 17], several studies have demonstrated that a subset of non-canonical splice sites is evolutionarily conserved. They recur at homologous loci across different species and are preserved within species [20, 13]. This supports the notion that many such events are maintained by selection, while others represent transient variants that are subsequently eliminated.

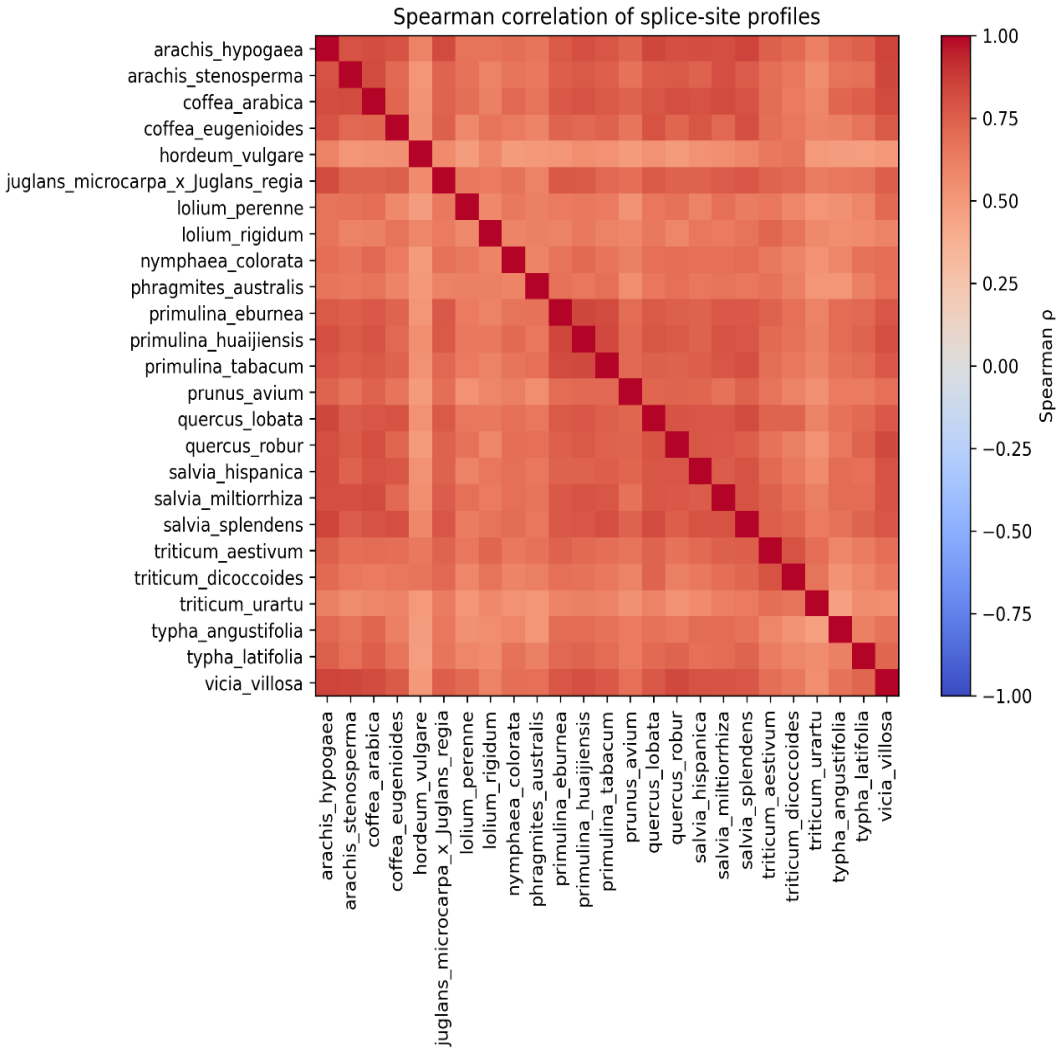

Fig. 1. Cross-species similarity of splice-site profiles (Spearman correlation heatmap)

The strong similarity of splice site profiles across species and the stable proportion of non-canonical introns indicate that the general hierarchy of splice site usage is maintained by core recognition mechanisms and RNA quality control. Rare non-canonical splice sites, however, fall into two categories: some are functionally important and preserved by selection, while others are transient and gradually removed. The independence of the proportion of non-canonical introns from annotation completeness has important methodological implications. Many non-canonical and cryptic events are systematically underrepresented: such transcripts are typically lowly expressed, rapidly degraded by NMD or nuclear decay, and therefore poorly detected by standard RNA-seq protocols, which are usually biased toward poly(A)+ RNA and long exons. As noted by Sibley et al. (2016), [21], cryptic splice sites and pseudoexons are rarely included in annotations, although their activation can lead to premature stop codons and the initiation of NMD. This is illustrated by the Nd-1 (Arabidopsis thaliana) genome: in its first annotation release (v1.0), non-canonical introns were entirely absent, but after incorporating Araport11-based hints and RNA-seq evidence, more than 1,200 cases were identified (v1.1). For comparison, in the high-quality Araport11 annotation the class proportions are stable (~98.9% GT–AG, ~1.0% GC–AG, ~0.1% AT–AC, and ~0.1% others), and nearly identical values are observed in Nd-1 v1.1. [22, 23]. Thus, while ab initio annotations systematically underestimate non-canonical events, annotations of comparable quality yield very similar proportions of non-canonical splice sites. This supports the view that the observed differences are driven primarily by annotation quality rather than annotation size.

4. Conclusion

In this study, we conducted a systematic analysis of splice site usage in 13 plant species. Our results demonstrate that non-canonical splice sites, although rare, are a consistent and reproducible feature of plant genomes. The stability of their relative proportions across well-annotated genomes indicates that non-canonical splicing is not a random artifact but an evolutionarily conserved phenomenon. At the same time, the scaling of splice site diversity with intron number suggests that genome complexity contributes to the emergence of rare splicing variants. By extending splice site studies beyond model and crop plants, this work provides a broader perspective on splicing variability and establishes a framework for future functional investigations into the biological roles of non-canonical splicing in plants.

References

[1] Gehring NH, Roignant JY. Anything but Ordinary – Emerging Splicing Mechanisms in Eukaryotic Gene Regulation. *Trends in Genetics* .2020, 37(4), p. 1-18. doi: 10.1016/j.tig.2020.10.008

[2] Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet*. 2018, 59, p. 253–268.

[3] Kornblihtt RA, De La Mata M, Fededa JP, Mucoz JM, Nogues G. Multiple links between transcription and splicing. *RNA*, 2004, 10(10), p.1489–1498.

[4] Nilsen TW, Graveley BR, Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 2010, 463, p. 457–463.

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

[6] Liu Y, Gonzàlez-Porta M, Santos S, Brazma A , Marioni JC, Aebersold R, Venkitaraman AR, Wickramasinghe VO. Impact of alternative splicing on the human proteome. *Cell Rep.* 2017, 20, p. 1229–1241.

[7] Lee Y, Rio DC. Mechanisms and regulation of alternative pre-mRNA splicing. *Annu. Rev. Biochem.* 2015, 84, p. 291–323.

[8] Fu XD, Ares JrM. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* 2014, 15, p. 689–701.

[9] Alizada S, Aliyeva K. Comparative analysis of expression profiles of antiporter encoding gene (*GhNHX1*) under different concentrations of NaCl in cotton (*Gossypium hirsutum* L.). *Advances in Biology & Earth Sciences*, 2024, 9, 168-174.

[10] Hertel KJ. Combinatorial Control of Exon Recognition. *The Journal of Biological Chemistry*. 2008, 283(3), p. 1211–1215.

[11] Frey K, Pucker, B. Animal, Fungi, and Plant Genome Sequences Harbor Different Non- Canonical Splice Sites. *Cells*, 2020, *9*(2), p. 1-19. doi: 10.3390/cells9020458

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

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

[14] Mamedova AO. Bioindication of environmental quality based on plant mutational and modification variability. Cytol. Genet., 2009, pp.123-125, https://link.springer.com/article/10.3103/S009545270902008X

[15] Brackenridge S, Wilkie AO, Screaton GR Efficient use of a 'dead-end' GA 5' splice site in the human fibroblast growth factor receptor genes. *EMBO J.*, 2003, 22, p. 1620–1631.

[16] Pollard AJ, Krainer AR, Robson SC, Europe-Finner GN. Alternative splicing of the adenylyl cyclase stimulatory G-protein G alpha(s) is regulated by SF2/ASF and heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1) and involves the use of an unusual TG 3'-splice site. *J. Biol. Chem.*, 2002, 277, p. 15241–15251.

[17] Parada GE, Munita R, Cerda CA, Gysling K, A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Research.* 2014, 42(16) p. 10564–10578. https://doi.org/10.1093/nar/gku744

[18] Reddy AS, Marquez Y, Kalyna M, Barta A. Complexity of the alternative splicing landscape in plants. *Plant Cell*. 2013, 25(10), p. 3657–3683. doi: 10.1105/tpc.113.117523

[19] Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* 2000, 28(21), p. 4364–4375. doi: 10.1093/nar/28.21.4364

[20] Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*. 2003, 13(17), p. 1512–1517. doi: 10.1016/S0960-9822(03)00558-X

[21] Sibley CR, Blazquez L, Ule J. Lessons from non-canonical splicing. *Nat Rev Genet.* 2016, 17(7), p. 407–421. doi: 10.1038/nrg.2016.46

[22] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*. 2000, 408, p. 796–815. doi: 10.1038/35048692

[23] Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J.* 2017, 89(4), p. 789–804. doi: 10.1111/tpj.13415