Iranian Journal of Blood & Cancer

Iran J Blood Cancer, 2026, 18(2), 70-87
https://ijbc.ir

Original Article

Integrated In Silico Analysis of the Splicing Effects TMSB4X c.-16-1G>A Variant and Prime Editing Design for Splicing Correction in Diffuse Large B- Cell Lymphoma

Eca Desriana Zahwa[1*], Noorwati Sutandyo[1,2,3], Rafika Indah Paramita[1,4,5]

[1]*Master's Programme in Biomedical Sciences, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 4, Jakarta, 10430, Indonesia.*
[2]*Department of Hematology and Medical Oncology, Dharmais National Cancer Center Hospital, Jalan Letjen S.Parman, Jakarta, 11420, Indonesia.*
[3]*Department of Internal Medicine, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 4, Jakarta, 10430, Indonesia.*
[4]*Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 4, Jakarta, 10430, Indonesia.*
[5]*Bioinformatics Core Facilities-IMERI, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 6, Jakarta, 10430, Indonesia*

Scan and read the article online

Citation Zahwa ED, Sutandyo N, Paramita RI. Integrated In Silico Analysis of the Splicing Effects TMSB4X c.-16-1G>A Variant and Prime Editing Design for Splicing Correction in Diffuse Large B- Cell Lymphoma. Iran J Blood Cancer. 2026 June 30;18(2): 70-87.

doi 0

Article info:
Received: 19 Mar 2026
Accepted: 20 Jun 2026
Published: 30 Jun 2026

Abstract

Diffuse Large B-Cell Lymphoma (DLBCL) is an aggressive lymphoid malignancy characterized by diverse genetic alterations, including variants that disrupt pre-mRNA splicing. The TMSB4X c.-16-1G>A variant is a splice-site alteration reported in DLBCL; however, its molecular consequences and potential correction using prime editing have not been comprehensively investigated. This study aimed to evaluate the impact of the TMSB4X c.-16-1G>A variant on splicing, characterize its effects on transcript and protein products, and design an in silico prime editing strategy for variant correction. Variant characterization was performed through genomic, clinical, and population frequency annotation. Splicing effects were assessed using dbscSNV, MaxEntScan, SpliceAI, Pangolin, Human Splicing Finder (HSF), RNA Disease Cryptic Splicing Code (RDCC), and SAI-10k-calc. Transcript and protein consequences were analyzed through open reading frame prediction and protein translation. Prime editing design included sgRNA selection, pegRNA optimization, and off-target assessment. The TMSB4X c.-16-1G>A variant was identified as a splice_acceptor_variant, absent from population databases, and reported in COSMIC as a somatic mutation associated with hematological malignancies. All prediction tools consistently indicated a high probability of splicing disruption, predicting both a 155-bp intron retention event and exon 2 skipping. These aberrant transcripts were predicted to generate truncated proteins compared with the wild-type protein (44 amino acids), producing proteins of 32 and 22 amino acids, respectively. Furthermore, two candidate prime editing systems were successfully designed for targeted correction of the variant. These findings suggest that TMSB4X c.-16-1G>A may induce aberrant splicing and provide a foundation for future experimental validation of prime editing-based correction strategies in DLBCL.

Keywords:
TMSB4X
DLBCL
Splicing
Splice-site variant
Prime editing
Bioinformatics

* Corresponding Author:
Eca Desriana Zahwa
E-mail: **eca.desriana@ui.ac.id**

1. INTRODUCTION

Diffuse Large B-Cell Lymphoma (DLBCL) is the most common subtype of non-Hodgkin lymphoma in adults, accounting for approximately 30–40% of all non-Hodgkin lymphoma cases worldwide. DLBCL is characterized by the rapid proliferation of malignant B lymphocytes and exhibits marked clinical, molecular, and genetic heterogeneity [1–4]. Despite advances in immunochemotherapy, particularly with the introduction of the R-CHOP regimen, which have significantly improved patient response rates and survival outcomes, approximately 30–40% of patients still experience relapse or develop refractory disease following first-line treatment [5,6]. This clinical challenge highlights the incomplete understanding of the molecular mechanisms underlying DLBCL pathogenesis and progression, underscoring the need to identify novel genetic biomarkers and therapeutic targets that may facilitate the development of more effective and precision-based treatment strategies.

Advances in sequencing technologies, particularly Next-Generation Sequencing (NGS), have revealed that DLBCL is a genetically and epigenetically driven malignancy characterized by the accumulation of diverse genetic alterations [3] and epigenetic abnormalities [7]. In addition to mutations that directly alter protein-coding sequences, accumulating evidence suggests that genetic variants affecting gene expression regulation also play a critical role in cancer development. Splice-site variants represent a major cause of aberrant splicing. Mutations occurring at canonical donor or acceptor splice sites can disrupt spliceosome-mediated recognition of exon–intron boundaries, leading to a range of molecular consequences, including exon skipping, intron retention, activation of cryptic splice sites, and reading-frame alterations. These splicing abnormalities may result in loss of protein function, production of truncated proteins, or increased expression of oncogenic isoforms. In recent years, aberrant splicing has emerged as a key molecular hallmark across multiple cancer types, including hematologic malignancies. Consequently, the identification and characterization of splice-site variants are essential for understanding the biological impact of genetic alterations and their potential contribution to cancer pathogenesis [8].

Among the genes implicated in cellular regulation and lymphomagenesis, TMSB4X has emerged as a potential candidate of interest in DLBCL. TMSB4X encodes Thymosin β4 (Tβ4), a highly conserved 44-amino-acid peptide that belongs to the thymosin family. Tβ4 plays a pivotal role in a variety of cellular processes, including actin cytoskeleton dynamics, cell migration, proliferation, differentiation, and tissue repair [9]. Increasing evidence has linked aberrant TMSB4X expression to the initiation and progression of several malignancies, including thyroid cancer [10], breast cancer [11], colorectal cancer [12], prostate cancer [13] and hematologic cancers such as Multiple Myeloma [14] and DLBCL [15]. In particular, elevated TMSB4X expression has been reported in multiple solid tumors and has been associated with enhanced angiogenesis, tumor invasiveness, and metastatic potential [16,17]. Despite these observations, current evidence has largely focused on transcriptional dysregulation of TMSB4X, whereas the functional consequences of genetic variants that may alter transcript processing remain poorly characterized. This knowledge gap is especially evident in DLBCL, where the contribution of TMSB4X splice-altering variants to disease pathogenesis has yet to be systematically investigated.

The TMSB4X c.-16-1G>A splice-site variant represents a potentially functional genetic alteration due to its location at a highly conserved splice acceptor site. Nucleotide substitution at this critical position may impair the recognition of the canonical splice acceptor sequence by the spliceosome, thereby disrupting normal splicing of TMSB4X transcripts. Such alterations can lead to aberrant transcript processing, including the generation of unstable transcripts, changes in RNA architecture, and modifications in the abundance or functional properties of the resulting protein products. Given the established role of TMSB4X in regulating actin cytoskeleton dynamics and cellular behaviors associated with tumor progression, splicing defects affecting this gene may have important biological implications for DLBCL pathogenesis [18,19]. Nevertheless, the molecular consequences of the c.-16-1G>A variant remain insufficiently characterized, and its impact on transcript processing and protein integrity has not yet been systematically investigated.

Bioinformatics approaches have become indispensable for the interpretation of genetic variants, enabling rapid, cost-effective, and comprehensive assessment of their potential molecular consequences. A wide range of computational tools is currently available to predict alterations in splice-site strength, identify alternative splicing events, evaluate nucleotide conservation, assess changes in RNA secondary structure, and infer transcript-level outcomes associated with specific variants. The integration of these predictive methodologies provides valuable preliminary evidence regarding the functional and pathogenic potential of genetic alterations prior to experimental validation. Consequently, an integrated in silico framework represents a highly relevant strategy for investigating the biological impact of

the TMSB4X c.-16-1G>A variant and its potential contribution to DLBCL pathogenesis [20].

Beyond elucidating the molecular consequences of genetic variants, recent advances in genome-editing technologies have created new opportunities for the direct correction of disease-causing mutations. Among these innovations, prime editing has emerged as a versatile CRISPR-based genome-editing platform capable of introducing precise nucleotide substitutions, insertions, and deletions without requiring double-strand DNA breaks. Compared with conventional CRISPR-Cas9 approaches, prime editing offers greater editing precision while reducing the risk of unintended genomic alterations. These advantages make prime editing a particularly promising strategy for correcting splice-site mutations that give rise to aberrant splicing. Nevertheless, the success of this approach depends heavily on the rational design and optimization of prime editing guide RNAs (pegRNAs) and other editing components, which must be tailored to the specific genomic context of the target mutation. [21,22].

In light of these considerations, this study aimed to perform an integrated in silico analysis of the TMSB4X c.-16-1G>A splice-site variant to evaluate its potential effects on RNA splicing and transcript integrity in DLBCL. Furthermore, a targeted prime editing strategy was designed to enable precise correction of the identified mutation. By combining splicing prediction, transcript-level characterization, and genome-editing design, this study seeks to provide insights into the biological significance of TMSB4X splice-site alterations in DLBCL and establish a foundation for the future development of prime editing–based therapeutic approaches targeting aberrant splicing in cancer.

2. MATERIALS & METHODS

2.1. Study Design

An integrated in silico bioinformatics framework was employed to investigate the molecular consequences of the TMSB4X c.-16-1G>A splice-site variant and to design a precision genome-editing strategy for its correction in Diffuse Large B-Cell Lymphoma (DLBCL). The study workflow consisted of variant identification and annotation, computational prediction of splicing defects, characterization of transcript-level consequences, and the design and optimization of a prime editing system targeting the mutant allele.

2.2. Sequence Data Retrieval and Genetic Annotation

Genomic sequence data and gene annotations for TMSB4X were retrieved from publicly available databases, including the National Center for Biotechnology Information (NCBI) and the UCSC Genome Browser, using the human reference genome GRCh38/hg38. The reference transcript selected for this study was NM_021109.4 (RefSeq), corresponding to ENST00000451311.7 (Ensembl), which is designated as the MANE Select (Matched Annotation from NCBI and EMBL-EBI) transcript. The use of a MANE Select transcript ensured consistency between RefSeq and Ensembl annotations and facilitated standardized variant reporting. The variant investigated in this study was chrX:g.12976245G>A. Variant coordinates were verified against the reference transcript and corresponding genomic annotations to ensure accurate localization within the TMSB4X gene structure. Reference (wild-type) and mutant genomic sequences were extracted from regions encompassing the canonical splice acceptor site and adjacent nucleotides for subsequent splicing prediction analyses and prime editing design. Genetic information collected included gene identity, chromosomal location, transcript architecture, and exon–intron organization. Variant annotation was performed according to the recommendations of the Human Genome Variation Society (HGVS) to ensure nomenclature consistency, interpretative accuracy, and reproducibility of the findings. The annotated genomic and transcriptomic data served as the foundation for downstream analyses, including splicing impact prediction, assessment of variant pathogenicity, and the development of a prime editing strategy for the correction of splicing abnormalities.

2.3. Variant Annotation and Characterization

Variant annotation and characterization were performed to evaluate the biological significance, population frequency, and potential functional impact of the TMSB4X variant. Comprehensive genomic annotation and functional prediction were conducted using VarSome (https://varsome.com/) [23] and FAVOR (Functional Annotation of Variants Online Resource) (https://favor.genohub.org/) [24] as the primary annotation platforms. Clinical significance data were obtained from the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/), whereas information regarding the association of the variant with cancer was retrieved from COSMIC (the Catalogue of Somatic Mutations in Cancer) (https://cancer.sanger.ac.uk/cosmic). Population allele frequencies were assessed using data from gnomAD v4.1 (https://gnomad.broadinstitute.org/) which integrates both

exome and whole-genome sequencing datasets. Allele frequency (AF) values were used to determine the rarity of the variant and to evaluate its potential relevance as a rare pathogenic alteration within the general population [25]. To assess evolutionary constraint at the variant position, multiple conservation metrics available through the FAVOR platform were analyzed, including phyloP100way, phastCons, GERP (Genomic Evolutionary Rate Profiling), and aPC-Conservation scores. These metrics provide complementary measures of nucleotide conservation across species. Higher conservation scores indicate stronger evolutionary constraint and suggest that the affected nucleotide may play an important biological role, thereby increasing the likelihood that sequence alterations at this position could have functional consequences [24]. The resulting annotations were subsequently integrated to support the interpretation of variant pathogenicity and to provide biological context for downstream analyses of splicing disruption and transcript-level effects.

2.4. Splicing Prediction Analysis

Splicing prediction analysis was performed to evaluate the potential impact of the TMSB4X c.-16-1G>A variant on pre-mRNA splicing and to identify aberrant splicing events that could affect transcript structure and protein products. An integrated in silico approach was employed using multiple prediction tools based on complementary algorithmic frameworks to improve the robustness of the analysis. The effect of the variant on canonical splice-site strength was initially assessed using MaxEntScan (https://genebe.net/tools/maxentscan) which estimates donor and acceptor splice-site strength based on a maximum entropy model. Differences in splice-site scores between the wild-type and mutant alleles were used to evaluate potential disruption of normal splice-site recognition [26]. Deep learning–based splicing prediction was subsequently performed using SpliceAI (https://spliceailookup.broadinstitute.org/), which predicts the probability of splice donor or acceptor gain and loss caused by genetic variants. Higher delta scores were interpreted as indicating a greater likelihood of splicing disruption [27].

To further validate these predictions, Pangolin was used to assess variant-induced changes in splice-site utilization across human tissues [28]. Additional evidence was obtained from dbscSNV v1.1, which integrates AdaBoost and Random Forest models to estimate the probability that a variant affects splicing. Variants with high scores in both models were considered more likely to disrupt normal splicing [29]. Supporting annotations were retrieved from SpliceVarDB (https://splicevardb.org/) a database that compiles splicing-related variant information from multiple predictive and genomic resources, providing additional context for variant interpretation [30].

To characterize the predicted molecular consequences of splicing alterations, SpliceAI outputs were further analyzed using SAI-10k-calc, which translates SpliceAI predictions into specific splicing outcomes, including exon skipping, partial exon deletion, pseudoexonization, intron retention, and cryptic splice-site activation. Predicted transcript size changes associated with these events were also evaluated [31]. Aberrant splicing patterns and transcript-level consequences were characterized using RDCC (RNA Disease Variant Consequence Calculator) (https://rddc.tsinghua-gd.org/). his analysis was used to identify alternative splice sites, determine the type and extent of sequence insertion or deletion, and predict the resulting transcript structures. The integrated results were subsequently used to define the most likely splicing patterns associated with the TMSB4X c.-16-1G>A variant, including intron retention, exon skipping, exon extension or shortening, and cryptic splice-site activation.

2.5. Transcript Structure and Functional Consequence Analysis

Transcript structure and functional consequence analyses were performed based on the splicing patterns predicted by RDCC and SAI-10k-calc, with priority given to predictions exhibiting the highest DanQ scores. Mutant transcript structures were reconstructed according to the predicted splicing alterations to generate the corresponding aberrantly spliced mRNA sequencesOpen reading frame (ORF) analysis was subsequently conducted using ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/) to assess the impact of splicing alterations on coding potential, including the presence of frameshift events, premature termination codons (PTCs), and the likelihood of nonsense-mediated mRNA decay (NMD). Predicted transcript sequences were then translated into protein sequences using ExPASy Translate Tools (https://web.expasy.org/translate/) [32] and Bencling (https://www.benchling.com/). The resulting protein products were compared with the wild-type protein to evaluate changes in protein length, amino acid composition, and overall protein integrity. Particular attention was given to the identification of truncation events and other alterations that could affect protein function as a consequence of aberrant splicing. This integrated analysis was used to characterize the potential downstream effects of the TMSB4X c.-16-1G>A variant at both the transcript and protein levels.

2.6. Desain Prime Editing

A prime editing strategy was designed in silico to correct the TMSB4X c.-16-1G>A variant using PE-Designer (http://www.rgenome.net/pe-designer/) and pegFinder (http://pegfinder.sidichenlab.org/). Input sequences consisted of a 101-bp genomic fragment encompassing the variant site, including 50 bp upstream, the variant nucleotide, and 50 bp downstream. Both wild-type and mutant sequences were analyzed to identify suitable editing targets. Candidate prime editing guide RNAs (pegRNAs) were generated using both platforms and included the essential components required for prime editing, namely the spacer sequence, Primer Binding Site (PBS), and Reverse Transcription Template (RTT). Candidate designs were evaluated based on the presence of an appropriate protospacer adjacent motif (PAM), the position of the target nucleotide relative to the Cas9 nickase cleavage site, PBS and RTT lengths, and the predicted editing efficiency scores provided by each tool. The optimal pegRNA and nicking guide RNA (ngRNA) candidates were selected according to predicted editing efficiency, target-site accessibility, and their ability to precisely restore the mutant allele to the corresponding wild-type sequence. The final editing designs were subsequently visualized to illustrate the relative positions of the variant, PAM sequence, pegRNA components, ngRNA, PBS, and RTT within the TMSB4X genomic locus. [33,34].

2.7. Analysis Off-Target

Potential off-target effects were evaluated using Cas-OFFinder (http://www.rgenome.net/cas-offinder/) and Off-Spotter (https://cm.jefferson.edu/Off-Spotter/). Two web-based platforms widely used for genome-wide off-target prediction. These tools were employed to identify genomic loci exhibiting sequence similarity to the sgRNAs selected for prime editing, thereby enabling assessment of target specificity. Cas-OFFinder was used to identify potential off-target sites based on sequence complementarity between the guide RNA and the human reference genome, whereas Off-Spotter performed genome-wide searches using pre-indexed genomic databases to efficiently detect candidate off-target loci. The results obtained from both platforms were compared to identify off-target sites consistently detected by both methods. Candidate off-target sites were prioritized according to the degree of sequence similarity to the guide RNA, as determined by the number of mismatches, and the presence of compatible PAM sequences. This analysis was performed to evaluate the specificity and potential safety of the proposed prime editing strategy targeting the TMSB4X c.-16-1G>A variant [35,36].

3. RESULTS

3.1. Genomic Characteristics, Functional Significance, and Evolutionary Conservation

Variant annotation identified the analysed alteration as a single-nucleotide variant (SNV) in the TMSB4X gene, consisting of a G>A substitution at the canonical splice acceptor site. Based on the GRCh38/hg38 reference genome, the variant is located on chromosome X at position chrX:12976245G>A. Molecular annotation classified this alteration as a splice_acceptor_variant. Querying the ClinVar database identified the variant as ClinVar Allele ID 3881178 and Variation ID 3765006 (VCV003765006.2), with a clinical classification of uncertain significance and an annotation indicating somatic oncogenicity. In the COSMIC database (COSV66081358), the variant is classified as a driver mutation. To date, the variant has been reported in five cancer cases, including three cases associated with Diffuse Large B-Cell Lymphoma (DLBCL). Population frequency analysis revealed that the TMSB4X c.-16-1G>A variant was absent from both the 1000 Genomes Project Phase 3 and gnomAD v4.1 (exome and genome) datasets across all available populations, including African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) cohorts. A summary of the genomic characteristics, clinical annotations, and population frequency data is presented in **Table 1**. and **Figure 1**.

Functional annotation indicated that the TMSB4X c.-16-1G>A variant had a CADD PHRED score of 22.1, corresponding to the top 0.6% of potentially deleterious variants in the human genome. The variant also exhibited a LINSIGHT score of 24.62 (0.3rd genomic percentile), a FATHMM-XF score of 0.405 (91.1st percentile), and a DANN score of 0.99, collectively suggesting a substantial likelihood of functional impact. Within the Annotation Principal Components framework, the highest scores were observed for Epigenetics Active (24.07), Transcription Factor (22.56), and Epigenetics Transcription (20.67), all of which ranked within the lowest genomic percentiles (<1%), indicating strong enrichment for functional regulatory features. A summary of the functional annotation results is provided in **Table 2**. Evolutionary conservation analysis demonstrated that the variant position is located within a highly conserved genomic region. Conservation metrics yielded a phyloP100way score of 3.164, a phastCons score of 0.993, and a GERP score of 13.4. Furthermore, the aPC-

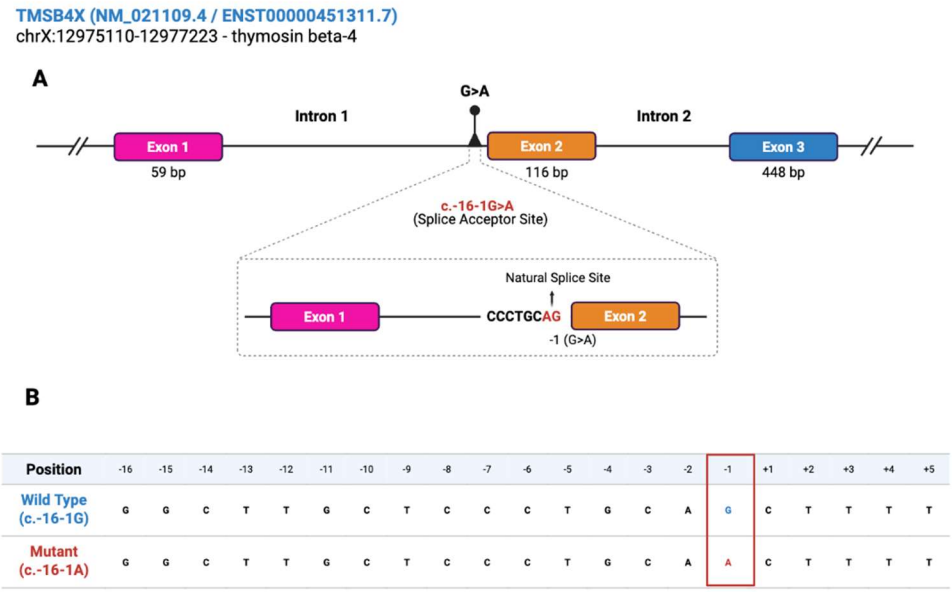

Figure 1. Genomic context and splice-site localization of the TMSB4X c.-16-1G>A variant

A. Schematic representation of the TMSB4X gene structure showing exon–intron organization and the position of the c.-16-1G>A variant at the canonical splice acceptor site of intron 1.

B. Nucleotide-level illustration of the wild-type and mutant sequences highlighting the G>A substitution at position −1 relative to the exon 2 splice junction.

Table 1. Genomic Characteristics, Clinical Annotation, and Population Frequency of the TMSB4X c.-16-1G>A Variant

Annotation	Result
Gene	TMSB4X
Genomic coordinate	chrX:12976245G>A
Reference genome	GRCh38/hg38
Transcript-level HGVS	c.-16-1G>A
Genomic HGVS	NC_000023.11:g.12976245G>A NM_021109.4:c.-16-1G>A NG_146687.1:g.517G>A
Type and length	single nucleotide variant, 1 bp
Molecular consequence	splice_acceptor_variant
ClinVar Allele ID	3881178
Clinical significance (ClinVar Variation ID: 3765006; Accession VCV003765006.2)	Uncertain significance; Somatic - Oncogenicity
COSMIC ID	COSV66081358
COSMIC classification	Driver mutation; Somatic Mutation
COSMIC occurrence	Reported in 5 cancer cases, including 3 cases of diffuse large B-cell lymphoma (DLBCL)
1000 Genomes Project Phase 3	Not detected across AFR, AMR, EAS, EUR, and SAS populations
gnomAD v4.1 Exome	Not detected across AFR, AMR, EAS, EUR, and SAS populations
gnomAD v4.1 Genome	Not detected across AFR, AMR, EAS, EUR, and SAS populations

Abbreviations: AF = Allele Frequency; HGVS = Human Genome Variation Society; AFR = African; AMR = Admixed American; EAS = East Asian; EUR = European; SAS = South Asian.

Table 2. Functional Annotation and Predicted Biological Impact of the TMSB4X c.-16-1G>A Variant

Annotation	Integrative Score (PHRED Scale)	Genome-wide Percentile
aPC-Protein Function	2.969487	50.5%
aPC-Epigenetics Active	24.067011	0.4%
aPC-Epigenetics Repressed	10.698860	8.5%
aPC-Epigenetics Transcription	20.665149	0.9%
aPC-Local Nucleotide Diversity	10.167685	9.6%
aPC-Mutation Density	9.423340	11.4%
aPC-Transcription Factor	22.561253	0.6%
aPC-Mappability	0.519677	88.7%
CADD PHRED	22.1	0.6%
LINSIGHT	24.616715	0.3%
FATHMM-XF	0.405274	91.1%
DANN	0.99	

Abbreviations: aPC = Annotation Principal Components; CADD = Combined Annotation Dependent Depletion; PHRED = Phred-scaled score; LINSIGHT = Linear INSIGHT; FATHMM-XF = Functional Analysis Through Hidden Markov Models–eXtended Features; DANN = Deleterious Annotation of Genetic Variants Using Neural Networks.

Table 3. Evolutionary Conservation Analysis at the TMSB4X c.-16-1G>A Variant Locus

Parameter	Result
phyloP100way	3.164
phastCons score	0.993
GERP score	13.4
aPC-Conservation	19.976269

Table 4. In Silico Splicing Impact Prediction of the TMSB4X c.-16-1G>A Variant Using Multiple Algorithms

Tools	Calibrated Prediction	Score
dbscSNV v.11	Pathogenic Strong	0.9999
MaxEntScan	Pathogenic Strong	8.7504
Splicevardb	Supporting Evidence	3.0
SpliceAI score (max)	Pathogenic Strong	0.99
Pangolin	Pathogenic Strong	0.88

Conservation score was 19.98. Collectively, these findings indicate strong evolutionary constraint at the affected nucleotide position, supporting its potential functional importance. A summary of the conservation analysis is presented in **Table 3**.

3.2. Predicted Splicing Impact of the TMSB4X c.-16-1G>A Variant

Integrated in silico splicing analyses consistently indicated that the TMSB4X c.-16-1G>A variant is highly likely to disrupt normal pre-mRNA splicing (**Table 4**). Four independent prediction tools dbscSNV, MaxEntScan, SpliceAI, and Pangolin classified the variant as having a strong pathogenic effect on splicing, with scores of 0.9999, 8.7504, 0.99, and 0.88, respectively. In addition, SpliceVarDB provided supporting evidence for splicing disruption with a score of 3.0. Collectively, these findings indicate a strong likelihood that the c.-16-1G>A substitution interferes with canonical splice-site recognition and promotes aberrant splicing events.

Analysis using the RNA Disease Cryptic Splicing Code (RDCC) identified multiple potential cryptic splice sites in the vicinity of the variant locus (**Figure 2A**). Among the predicted candidates, a cryptic acceptor site located at c.-16-156 exhibited a high DanQ score (0.999), suggesting a strong probability of utilization as an alternative splice acceptor. Based on this prediction, RDCC identified a potential aberrant splicing event characterized by 155-bp intron retention resulting from activation of the cryptic acceptor site (**Figure 2B**). A second predicted splicing outcome involved complete skipping of exon 2. This finding was further supported by SAI-10k-calc, which predicted the loss of a 100-bp coding exon corresponding to exon 2, accompanied by the elimination of the canonical translation initiation codon (start codon loss) (**Figure 2C**). Taken together, these results suggest that the TMSB4X c.-16-1G>A variant may give rise to multiple aberrant transcript isoforms through distinct splicing mechanisms, including cryptic splice-site activation with intron retention and exon 2 skipping.

3.3. Transcript Structure Analysis and Protein Consequences of Aberrant Splicing

Based on RDCC predictions, the cryptic acceptor site located at c.-16-156 exhibited the highest DanQ score (0.999) and was predicted to induce a 155-bp intron retention event. This aberrant splicing pattern resulted in substantial alterations to transcript structure relative to the wild-type transcript. Subsequent transcript and translation analyses demonstrated that the wild-type TMSB4X transcript encodes a protein consisting of 44 amino acids. In contrast, the transcript containing the predicted 155-bp intron retention event was predicted to generate a truncated protein of 32 amino acids. Similarly, the exon 2–skipping transcript was predicted to encode a markedly shorter protein comprising only 22 amino acids.

Comparison of the translated protein sequences revealed that both aberrant transcripts exhibited not only reductions in protein length but also substantial alterations in amino acid composition relative to the wild-type protein. These findings indicate that the splicing defects predicted for the TMSB4X c.-16-1G>A variant are likely to affect both transcript architecture and protein coding capacity, resulting in structurally altered protein products. Overall, the predicted intron retention and exon-skipping events are expected to generate truncated and compositionally distinct protein isoforms, suggesting a potential loss or alteration of normal TMSB4X function (**Figure 3**).

3.4. Prime Editing Design for Correction of the TMSB4X c.-16-1G>A Variant

A total of four candidate sgRNAs were identified for prime editing of the TMSB4X c.-16-1G>A variant (**Table 5**). The candidate sgRNAs were located between 11 and 20 bp from the target nucleotide and were distributed across both sense and antisense orientations. Based on the predefined design criteria, sgRNA 1 (TTTGTCAGACATGGTTGCGG) and sgRNA 3 (GGGTTTGTCAGACATGGTTG) were selected for subsequent prime editing design. sgRNA 1 exhibited the shortest distance to the target nucleotide (12 bp), whereas sgRNA 3 was positioned 15 bp from the editing site. A summary of sgRNA characteristics and selection criteria is provided in **Table 5** and **Figure 4**.

Using the two selected sgRNAs, pegRNAs were designed with varying lengths of the primer binding site (PBS) and reverse transcription template (RTT). For sgRNA 1, a total of 21 pegRNA candidates were generated, with PBS lengths ranging from 12-14 nucleotides and RTT lengths ranging from 14-20 nucleotides (**Table 6**). For sgRNA 3, 12 pegRNA candidates were generated, incorporating PBS lengths of 12-14 nucleotides and RTT lengths of 17-20 nucleotides (**Table 7**). The GC content of the PBS sequences ranged from 42.9% to 50.0%, whereas the RTT sequences exhibited GC contents between 52.6% and 63.2%. Several candidates generated design warnings due to the presence of cytosine (C) as the first nucleotide of the RTT sequence, while the remaining candidates satisfied all design criteria. Overall, 33 pegRNA candidates were generated, representing a range of PBS-RTT combinations suitable for subsequent optimization.

Following sgRNA selection and pegRNA optimization, two final prime editing designs were selected for correction of the TMSB4X c.-16-1G>A variant (**Table 8**). The first design utilized sgRNA 1 in the antisense orientation, with a predicted nicking position at nucleotide 62, a GC content of 50.0%, and an edit-to-nick distance of 12 bp. The second design employed sgRNA 3 in the sense orientation, with a nicking position at nucleotide 32, a GC content of 60.0%, and an edit-to-nick distance of 20 bp. For each sgRNA, a single optimized extension sequence comprising the RTT and PBS was selected and combined with the standard pegRNA scaffold to generate a complete pegRNA construct. Both final designs satisfied the established criteria for edit distance, GC content, and overall prime editing architecture, supporting their suitability for experimental validation. Detailed characteristics of the sgRNAs, extension sequences, scaffolds, and complete pegRNA constructs are summarized in **Table 8**.

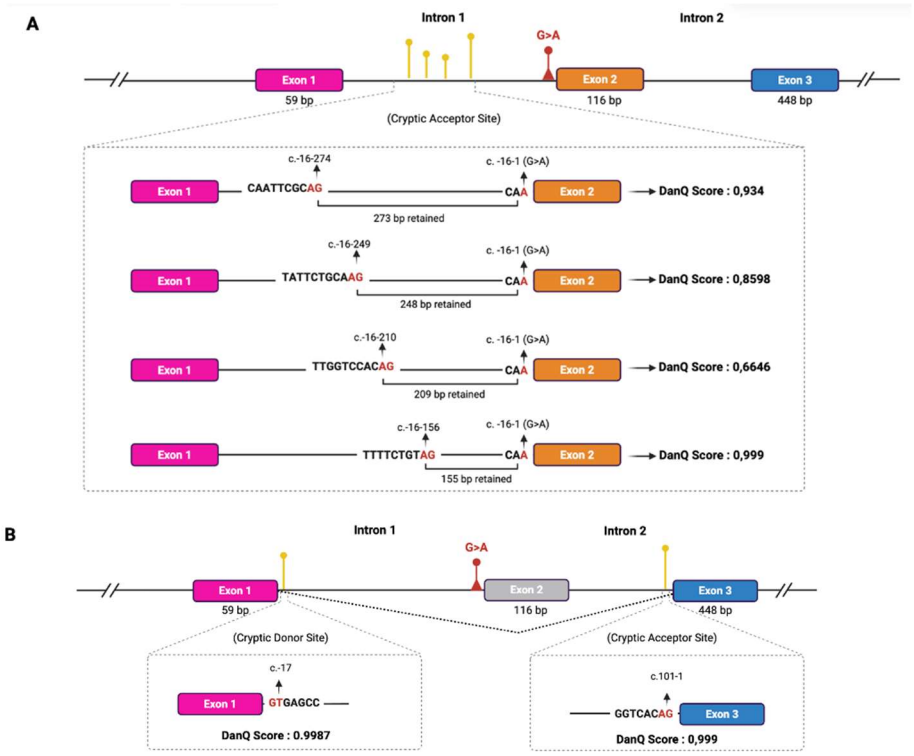

Figure 2. Predicted aberrant splicing consequences induced by the TMSB4X c.-16-1G>A variant

A. Computational prediction of cryptic splice acceptor site activation within intron 1 leading to intron retention events of varying lengths, together with corresponding DanQ scores.

B. Predicted exon 2 skipping and alternative splice site usage resulting in transcript disruption, including loss of canonical splicing signals.

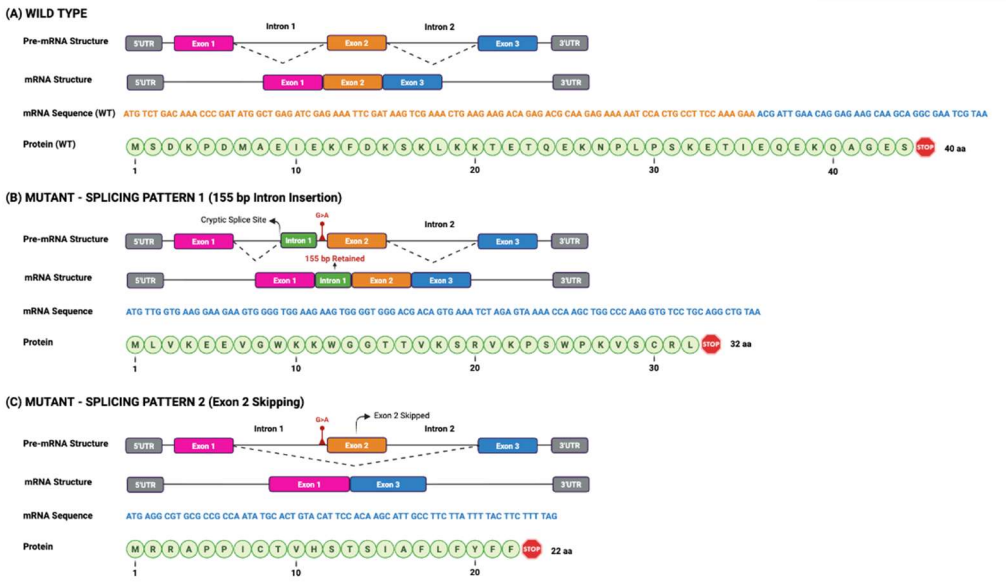

Figure 3. Transcript and protein-level consequences of aberrant splicing in TMSB4X c.-16-1G>A

A. Wild-type pre-mRNA and mRNA structure of TMSB4X, showing canonical exon composition and normal splicing pattern.

B. Predicted intron retention (155 bp) event resulting in frameshifted transcript and truncated protein product (32 amino acids).

C. Predicted exon 2 skipping event leading to severe truncation and production of a shortened protein (22 amino acids).

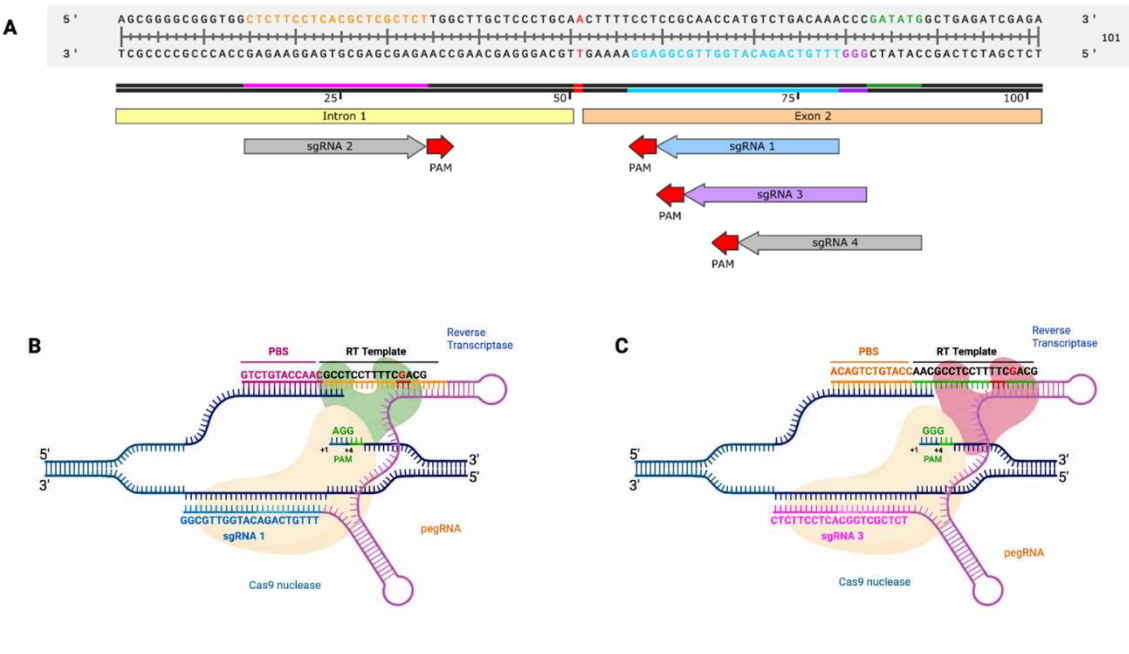

Figure 4. Prime editing strategy for correction of the TMSB4X c.-16-1G>A variant

A. Genomic overview of sgRNA positioning relative to the target site, protospacer adjacent motif (PAM), and editing window within the TMSB4X locus.

B. Schematic architecture of sgRNA 1 based prime editing system showing pegRNA components including primer binding site (PBS), reverse transcription template (RTT), and Cas9 nickase complex formation.

C. Schematic architecture of sgRNA 3 based prime editing system showing pegRNA components including primer binding site (PBS), reverse transcription template (RTT), and Cas9 nickase complex formation.

Table 5. Candidate sgRNAs for Prime Editing of the TMSB4X c.-16-1G>A Variant

No.	sgRNA Sequence (5'–3')	Cut Position	Orientation	Distance to Edit Start (bp)	PAM	Selected
1	TTTGTCAGACATGGTTGCGG	62	Antisense	11	AGG	✓
2	CTCTTCCTCACGCTCGCTCT	32	Sense	19	TGG	–
3	GGGTTTGTCAGACATGGTTG	65	Antisense	14	CGG	✓
4	CATATCGGGTTTGTCAGACA	71	Antisense	20	TGG	–

Table 6. Candidate pegRNA Designs with Variable PBS and RTT for sgRNA 1.

No.	Extension Sequence (RTT + PBS)	PBS Length (nt)	PBS GC (%)	RTT Length (nt)	RTT GC (%)	Warnings
1	CAGCTTTTCCTCCGCAACCATGTCTG	12	50.0	14	57.1	First base is 'C'
2	GCAGCTTTTCCTCCGCAACCATGTCTG	12	50.0	15	60.0	–
3	TGCAGCTTTTCCTCCGCAACCATGTCTG	12	50.0	16	56.3	–
4	CTGCAGCTTTTCCTCCGCAACCATGTCTG	12	50.0	17	58.8	First base is 'C'

5	CCTGCAGCTTTTCCTCCG CAACCATGTCTG	12	50.0	18	61.1	First base is 'C'
6	CCCTGCAGCTTTTCCTCCG CAACCATGTCTG	12	50.0	19	63.2	First base is 'C'
7	TCCCTGCAGCTTTTCCTCCG CAACCATGTCTG	12	50.0	20	60.0	–
8	CAGCTTTTCCTCCG CAACCATGTCTGA	13	46.2	14	57.1	First base is 'C'
9	GCAGCTTTTCCTCCG CAACCATGTCTGA	13	46.2	15	60.0	–
10	TGCAGCTTTTCCTCCG CAACCATGTCTGA	13	46.2	16	56.3	–
11	CTGCAGCTTTTCCTCCG CAACCATGTCTGA	13	46.2	17	58.8	First base is 'C'
12	CCTGCAGCTTTTCCTCCG CAACCATGTCTGA	13	46.2	18	61.1	First base is 'C'
13	CCCTGCAGCTTTTCCTCCG CAACCATGTCTGA	13	46.2	19	63.2	First base is 'C'
14	TCCCTGCAGCTTTTCCTCCG CAACCATGTCTGA	13	46.2	20	60.0	–
15	CAGCTTTTCCTCCG CAACCATGTCTGAC	14	50.0	14	57.1	First base is 'C'
16	GCAGCTTTTCCTCCG CAACCATGTCTGAC	14	50.0	15	60.0	–
17	TGCAGCTTTTCCTCCG CAACCATGTCTGAC	14	50.0	16	56.3	–
18	CTGCAGCTTTTCCTCCG CAACCATGTCTGAC	14	50.0	17	58.8	First base is 'C'
19	CCTGCAGCTTTTCCTCCG CAACCATGTCTGAC	14	50.0	18	61.1	First base is 'C'
20	CCCTGCAGCTTTTCCTCCG CAACCATGTCTGAC	14	50.0	19	63.2	First base is 'C'
21	TCCCTGCAGCTTTTCCTCCG CAACCATGTCTGAC	14	50.0	20	60.0	–

Table 7. Candidate pegRNA Designs with Variable PBS and RTT for sgRNA 3.

No.	Extension Sequence (RTT + PBS)	PBS Length (nt)	PBS GC (%)	RTT Length (nt)	RTT GC (%)	Warnings
1	CAGCTTTTCCTCCGCAA CCATGTCTGACA	12	50.0	17	52.9	First base is 'C'
2	GCAGCTTTTCCTCCGCAA CCATGTCTGACA	12	50.0	18	55.6	–
3	TGCAGCTTTTCCTCCGCAA CCATGTCTGACA	12	50.0	19	52.6	–
4	CTGCAGCTTTTCCTCCGCAA CCATGTCTGACA	12	50.0	20	55.0	First base is 'C'
5	CAGCTTTTCCTCCGCAA CCATGTCTGACAA	13	46.2	17	52.9	First base is 'C'
6	GCAGCTTTTCCTCCGCAA CCATGTCTGACAA	13	46.2	18	55.6	–
7	TGCAGCTTTTCCTCCGCAA CCATGTCTGACAA	13	46.2	19	52.6	–
8	CTGCAGCTTTTCCTCCGCAA CCATGTCTGACAA	13	46.2	20	55.0	First base is 'C'
9	CAGCTTTTCCTCCGCAA CCATGTCTGACAAA	14	42.9	17	52.9	First base is 'C'
10	GCAGCTTTTCCTCCGCAA CCATGTCTGACAAA	14	42.9	18	55.6	–

11	TGCAGCTTTTCCTCCGCAA CCATGTCTGACAAA	14	42.9	19	52.6	–
12	CTGCAGCTTTTCCTCCGCAA CCATGTCTGACAAA	14	42.9	20	55.0	First base is 'C'

Table 8. Final Selected sgRNA and pegRNA Designs for Correction of the TMSB4X c.-16-1G>A Variant by Prime Editing.

Target Sequence sgRNA 1	Target position	Cleavage position	Orientation	GC (%) Content	Editing position
TTTGTCAGACATGGTTG CGGAGG	57	62	Antisense (-)	50,0 %	12
Extension Sequence sgRNA 1 (RTT + PBS)	GCAGCTTTTCCTCCGCAACCATGTCTG				
Scaffold sequence	GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGT GGCACCGAGTCGGTGC				
full pegRNA sequence sgRNA 1 (5' to 3')	TTTGTCAGACATGGTTGCGG GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGT GGCACCGAGTCGGTGC GCAGCTTTTCCTCCG CAACCATGTCTG				
Target Sequence sgRNA 3	**Target position**	**Cleavage position**	**Orientation**	**GC (%) Content**	**Editing position**
GGGTTTGTCAGACATGG TTG	60	65	Antisense (-)	50,0 %	15
Extension Sequence sgRNA 3 (RTT + PBS)	GCAGCTTTTCCTCCGCAA CCATGTCTGACA				
Scaffold sequence	GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGT GGCACCGAGTCGGTGC				
full pegRNA sequence sgRNA 3 (5' to 3')	GGGTTTGTCAGACATGGTTG GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGT GGCACCGAGTCGGTGC GCAGCTTTTCCTCCGCAA CCATGTCTGACA				

Table 9. Predicted Off-Target Sites for sgRNA Candidates in the Prime Editing Design Targeting TMSB4X.

sgRNA	Target Sequence (5'→3')	Gene	Hit DNA	Chromosome Position	Mismatches
sgRNA1	CCGCAACCATGTCTG ACAAA	CRTC1	CaGCAACCATGcCTGAtAAA-AGG	chr19, + 18844247	3
		LUZP1	CtaCAACCATGgCTGACAAA-TGG	chr1, + 23466736	3
		PPAP2B	CaGaAACCATGTCTcAgAAA-AGG	chr1, - 57041381	4
		GRIK2	CaGCAAaCATcTCTGAgAAA-AGG	chr6, + 101850347	4
		SPATA5	CaGCAACtATGTgTGACAAc-AGG	chr4, + 123930355	4
		YME1L1	tCGCAgCaATGTCTGACAAg-AGG	chr10, + 27407745	4
sgRNA3	CAACCATGTCTGACAA ACCC	DCDC1	ggACCATGTCTaAgAAACCC-TGG	chr11, + 31271080	4
		SLC16A3	gAACCATtTCTGggAAACCC-GGG	chr17, - 80192235	4
		CACNB2	CAAgCcTtTCTGAtAAACCC-GGG	chr10, + 18779749	4

3.5. Off-Target Analysis of Candidate sgRNAs

Off-target analysis was performed for the two sgRNA candidates selected for prime editing design (**Table 9**). The prediction results identified a limited number of potential off-target sites for both sgRNAs, with all detected sites containing multiple sequence mismatches relative to the intended target sequence. For sgRNA 1, six potential off-target sites were identified. Two sites were located in the CRTC1 and LUZP1 genes, each containing three mismatches relative to the guide sequence. Four additional off-target sites were detected within PPAP2B, GRIK2, SPATA5, and YME1L1, all of which exhibited four mismatches. No off-target sites with zero, one, or two mismatches were identified for this sgRNA.

For sgRNA 3, three potential off-target sites were detected in the DCDC1, SLC16A3, and CACNB2 genes. All identified sites contained four mismatches relative to the guide sequence, indicating a lower degree of sequence similarity to the intended target locus. Overall, both sgRNAs demonstrated favorable specificity profiles, as all predicted off-target sites contained at least three mismatches. Notably, sgRNA 3 exhibited the lowest number of predicted off-target loci, with only three candidate sites identified, all containing four mismatches. These findings suggest a relatively low risk of unintended genome editing for both candidates, with sgRNA 3 displaying a slightly higher predicted specificity than sgRNA 1.

4. DISCUSSION

The TMSB4X c.-16-1G>A variant is located at the −1 position of the intron 1 splice acceptor site and is consequently classified as a splice acceptor variant. Variants occurring within canonical splice donor and acceptor sites are well recognized for their potential to disrupt highly conserved sequence motifs that define exon–intron boundaries during pre-mRNA processing. Accurate recognition of the 5′ and 3′ splice sites by the spliceosome is essential for the generation of mature transcripts, and alterations within these conserved regions may impair spliceosome assembly and splice-site recognition, ultimately leading to aberrant RNA processing. Previous studies have demonstrated that pathogenic splicing mutations are most frequently observed at the +1 and +2 positions of the 5′ donor site and the −1 and −2 positions of the 3′ acceptor site, which represent critical nucleotides required for splice-site recognition. Consequently, nucleotide substitutions affecting these canonical positions are often associated with substantial disruption of normal splicing and may result in a wide spectrum of transcript abnormalities, including exon skipping, intron retention, and activation of cryptic splice sites [37].

According to clinical annotation databases, the TMSB4X c.-16-1G>A variant is currently classified as a variant of uncertain significance (VUS) in ClinVar, indicating that the available evidence remains insufficient to establish a definitive association with a specific disease phenotype. Nevertheless, ClinVar also includes an annotation of somatic oncogenicity, suggesting that the variant has been identified in the context of human malignancies. This observation is consistent with data from COSMIC, where the variant is classified as a driver mutation and annotated as a somatic oncogenic alteration. Notably, the variant has been reported in five cases of hematologic malignancies, including three cases directly associated with Diffuse Large B-Cell Lymphoma (DLBCL), supporting its potential biological relevance in lymphoid tumorigenesis [38,39].

The apparent discrepancy between the current VUS classification and the available cancer-related evidence highlights a common challenge in the interpretation of non-coding and splice-site variants. While clinical classification frameworks often require substantial functional and clinical evidence before assigning pathogenic status, emerging genomic data increasingly demonstrate that variants affecting RNA splicing can exert significant biological effects despite limited clinical documentation. In this context, the recurrent detection of TMSB4X c.-16-1G>A in hematologic cancers, together with its localization at a canonical splice acceptor site, suggests that the variant may contribute to disease pathogenesis through disruption of normal transcript processing.

Furthermore, the absence of this variant from population databases and its exclusive identification in cancer-associated datasets are consistent with previous observations that deleterious splice-site mutations are frequently underrepresented in healthy populations. Such variants may be subject to negative selection in normal tissues because of their detrimental effects on gene function, yet confer a selective growth advantage when acquired somatically in malignant clones [8]. Collectively, these findings support the hypothesis that TMSB4X c.-16-1G>A represents a biologically relevant somatic alteration with potential implications for DLBCL development and progression.

The functional significance of non-coding variants in disrupting RNA splicing and contributing to human disease pathogenesis has been widely documented. For example, germline intronic mutations in COL6A1 have been shown to induce pseudoexon formation, leading to abnormal transcript processing and the development of collagen-related muscular dystrophy [40]. Analisis genomik skala luas

pada mutasi non-coding yang telah dikarakterisasi dalam penelitian terdahulu yang telah berhasil mengidentifikasi 228 mutasi non-coding pengubah pola splicing pada 219 gen yang tersebar di seluruh genom, di mana 17 di antaranya berkaitan erat dengan onkogenesis. Similarly, a large-scale analysis of experimentally validated non-coding variants identified 228 splicing-altering mutations across 219 genes, including 17 genes associated with cancer. Among these, proximal intronic G>A mutations in STK11 promoted the insertion of a novel 130-bp out-of-frame exon, resulting in reduced expression of this tumor suppressor gene and contributing to tumor progression [41]. These findings demonstrate that non-coding variants can exert significant biological effects through aberrant splicing mechanisms. Therefore, the localization of TMSB4X c.-16-1G>A at a canonical splice acceptor site further supports its potential to disrupt normal splicing and contribute to DLBCL pathogenesis.

The strong pathogenic potential of this variant is supported by several high functional prediction scores. A CADD PHRED score of 22.1 (corresponding to the 0.6th genomic percentile) and a DANN score of 0.99 indicate that the G-to-A nucleotide substitution is likely to have a substantial deleterious effect on gene function. This CADD score is consistent with previous empirical validations demonstrating that variants with CADD scores above 20 belong to the top 1% of variants predicted to be the most functionally deleterious in the human genome [42]. Although the use of a universal CADD threshold across the genome has been widely adopted, this approach was not originally recommended by the developers of the tool due to its potential to introduce predictive bias, particularly for splicing variants. Recent studies have suggested that a more appropriate threshold for splice-altering variants is a CADD score of 25.5, a cutoff that has been robustly validated using Mann–Whitney U testing, Benjamini–Hochberg false discovery rate (FDR) correction (<0.05), and significant receiver operating characteristic (ROC) curve analyses. Consequently, the CADD score of 22.1 observed for the c.-16-1G>A variant does not meet the pathogenicity criterion when evaluated exclusively against this splicing-specific threshold [43].

Therefore, the integration of complementary functional prediction metrics such as DANN is essential to mitigate this limitation and provide a more comprehensive assessment of pathogenicity. DANN analysis yielded an exceptionally high score of 0.99. Given that DANN scores range from 0 to 1, with values approaching 1 indicating an extremely high probability of deleterious functional effects, the score of 0.99 provides strong independent evidence supporting the pathogenic potential of this variant. Unlike CADD, which relies primarily on conventional linear models, DANN is capable of capturing complex non-linear relationships within non-coding regions, including intronic sequences, without introducing bias toward specific variant classes. Accordingly, the convergence of a near-maximal DANN score (0.99) with other functional metrics, including the LINSIGHT genomic percentile of 0.3% and the FATHMM-XF percentile of 91.1%, collectively supports the hypothesis that this somatic variant has the potential to disrupt normal TMSB4X gene function and expression [44].

Mutations affecting the canonical splice acceptor site (the AG dinucleotide) are widely recognized for their ability to disrupt exon–intron boundary recognition by the spliceosome complex. Consistent with this mechanism, all five in silico prediction algorithms employed in this study (dbscSNV, MaxEntScan, SpliceAI, Pangolin, and SpliceVarDB) unanimously classified the TMSB4X c.-16-1G>A variant as having a strong pathogenic effect on splicing. The reliability of this multi-algorithm approach is supported by a comparative study by Jang et al. (2022), which demonstrated that the combined use of SpliceAI and MaxEntScan achieved an accuracy exceeding 90% in predicting authentic splicing defects. The application of multiple prediction tools in the present study was therefore intended to provide comprehensive functional evidence and strengthen the assessment of the mutagenic effects associated with this variant [45].

Disruption of the canonical splice acceptor site is expected to force the splicing machinery to utilize alternative processing pathways, which in the present study were predicted to generate two deleterious aberrant transcript isoforms. This observation is consistent with the established behavior of splice-site mutations, whereby inactivation of a canonical acceptor site promotes the activation of nearby cryptic splice sites. Such alterations can result in varying degrees of intron retention or exon skipping. Consequently, variants producing these structural transcript abnormalities are frequently classified as deleterious or disruptive in large-scale cancer genomics studies [46].

The first predicted aberrant transcript resulted from a 155-bp intron retention event mediated by activation of an alternative cryptic acceptor site located at c.-16-156, which exhibited a near-maximal DanQ score of 0.999. Previous studies have shown that when a canonical acceptor site becomes nonfunctional, the spliceosome can shift to nearby cryptic sites with sufficient functional strength. Activation of cryptic splice sites caused by non-coding mutations has been identified as a common mechanism underlying the

incorporation of intronic sequences into mature mRNA transcripts [47].

The second predicted splicing abnormality involved exon 2 skipping, resulting in the exclusion of a 100-bp coding sequence (CDS). Experimental evidence has demonstrated that loss of the canonical translation initiation codon due to exclusion of a protein-coding exon frequently leads to complete translational failure or severe protein truncation. This structural consequence arises from the inability of the ribosomal scanning machinery to identify a valid translation initiation site and reading frame. Similar observations have been reported for the tumor suppressor gene PTEN, in which exon 3 skipping generated functionally defective transcripts. These findings indicate that splicing abnormalities affecting essential coding exons can produce dysfunctional gene products that directly contribute to tumor initiation and progression [48,49].

The transcript alterations resulting from aberrant splicing are predicted to directly modify the architecture of the resulting Thymosin β4 protein. The wild-type transcript encodes a functional protein consisting of 44 amino acids and containing a conserved G-actin-binding motif. In contrast, the transcript generated by the predicted 155-bp intron retention event is expected to undergo a frameshift, resulting in truncation of the protein to only 32 amino acids. The exon 2-skipping transcript is predicted to produce an even shorter protein of 22 amino acids, accompanied by substantial alterations in amino acid composition. Several studies have demonstrated that splice acceptor site variants can disrupt normal splicing mechanisms and lead to the generation of aberrant transcripts. For example, a study of Li–Fraumeni syndrome identified a splice acceptor mutation within intron 3 of the TP53 gene affecting a critical invariant residue. This alteration disrupted recognition of the canonical splice site and resulted in the production of two distinct aberrant transcripts [50].

From a biological perspective, the severe truncation and marked alterations in amino acid sequence observed in the predicted mutant proteins are likely to result in loss of function of Thymosin β4. This destructive consequence is consistent with previous structural studies demonstrating that the integrity of the N-terminal amino acid region of Thymosin β4 is essential for maintaining its actin polymerization-inhibitory activity. In particular, amino acid residues spanning positions 13–23 are critical for stable binding to G-actin. The affinity of Thymosin β4 for G-actin has been reported to be completely abolished when either the first 23 amino acids of the N-terminus or the last 26 amino acids of the C-terminus are deleted or substantially altered. Furthermore, the α-helical structure located within the N-terminal domain constitutes an essential binding architecture that mediates inhibition of actin polymerization through steric hindrance. In addition, the first three amino acids of the highly conserved "LKKTET" motif function as the principal electrostatic contact site responsible for initiating interactions between Thymosin β4 and G-actin, while simultaneously inducing the functional folding of both the N-terminal and C-terminal helical regions. Given that the transcript generated by the predicted 155-bp intron retention event encodes a protein of only 32 amino acids, whereas exon 2 skipping produces a protein of merely 22 amino acids, the c.-16-1G>A variant is expected to eliminate a substantial portion of these critical binding domains, including the conserved "LKKTET" motif and other essential hydrophobic anchoring residues. Consequently, the truncated proteins predicted in this study would be expected to lose the structural basis required for interaction with actin and therefore be unable to perform their physiological role in regulating the cytoskeleton of malignant cells [51].

To precisely correct the c.-16-1G>A point mutation without introducing potentially harmful double-strand breaks (DSBs), a prime editing (PE) strategy was implemented in the present study. Initial screening identified 33 functional pegRNA candidates derived from two selected sgRNAs (sgRNA 1 and sgRNA 3), encompassing primer binding site (PBS) lengths of 12–14 nucleotides and reverse transcription template (RTT) lengths of 14–20 nucleotides. The GC content of the PBS (42.9–50.0%) and RTT (52.6–63.2%) components fell within the thermodynamically optimal range. Consistent with the prime editing design guidelines proposed by Anzalone et al. (2019), balanced PBS GC content is expected to promote stable annealing to the target DNA strand following nicking while avoiding excessive stabilization that could impair dissociation after reverse transcription. The two final prime editing designs selected in this study therefore possess several advantageous structural characteristics. From a structural standpoint, it has been recommended that the first nucleotide of the 3′ pegRNA extension should not be cytosine (C). This hypothesis suggests that the presence of a cytosine residue at this position may facilitate aberrant base pairing with G81 within the gRNA scaffold. Such non-specific interactions could disrupt the canonical RNA conformation and consequently reduce the efficiency of Cas9 binding and editing activity [52].

The editing distance between the sgRNA cleavage site and the target nucleotide position c.-16-1G>A revealed strategically distinct spatial configurations between the two designs. sgRNA 1 cleaves the antisense strand with a very

short editing distance of 12 bp, whereas sgRNA 3 cleaves the sense strand with an editing distance of 15 bp. From the perspective of DNA repair kinetics, the close spatial proximity observed for sgRNA 1 (12 bp) offers a potentially significant functional advantage. An editing distance of less than 15 bp is theoretically expected to minimize exposure of the single-stranded DNA flap to degradation by cellular exonucleases before reverse transcription is fully completed. In contrast, although sgRNA 3 and sgRNA 4 exhibited greater editing distances, these designs would require additional consideration to ensure efficient editing and maintain synthetic viability.

In addition to editing efficiency, genomic safety through the minimization of off-target effects represents a critical parameter in prime editing design. Molecular predictive analysis based on the updated data presented in **Table 9** revealed markedly contrasting risk profiles between the two sgRNA candidates. sgRNA 1 identified six potential off-target sites distributed across several functional genomic loci. These included two sites with three mismatches located within the CRTC1 and LUZP1 genes, as well as four sites with four mismatches located within PPAP2B, GRIK2, SPATA5, and YME1L1. Although the accumulation of mismatches outside the seed region can substantially reduce cleavage activity, the presence of six potential off-target sites nevertheless warrants careful consideration. However, the comparative study by Anzalone et al. (2019) demonstrated that prime editing possesses an additional layer of specificity, whereby sgRNA binding alone is insufficient to generate permanent genomic modifications in the absence of complementary pairing between the pegRNA primer binding site (PBS) and the target sequence. Consequently, these findings suggest that the predicted off-target interactions of sgRNA 1 at loci such as CRTC1 and LUZP1 may not necessarily result in permanent collateral mutations or indel formation [52,53].

In contrast, sgRNA 3 exhibited a substantially safer profile for long-term clinical translation. According to the off-target analysis, sgRNA 3 possessed only three potential off-target sites, representing a 50% reduction compared with sgRNA 1. Furthermore, all predicted off-target sites associated with sgRNA 3 contained four mismatches and were specifically located within the DCDC1, SLC16A3, and CACNB2 genes. As demonstrated in a CRISPR specificity study by Dey et al. (2024), the presence of four mismatches within a non-target sequence is generally sufficient to markedly impair the thermodynamic binding affinity of the Cas9–guide RNA complex, thereby substantially reducing cleavage activity. Therefore, although sgRNA 1 provides an advantage in terms of editing distance and repair kinetics, sgRNA 3 can be considered the more promising overall candidate. sgRNA 3 offers a substantially greater safety margin by limiting potential off-target activity to only three loci, all characterized by the maximum observed level of sequence mismatch, consistent with the stringent criteria applied in the selection of guide RNAs for clinical applications [54].

This study provides a strong theoretical and computational foundation regarding the pathogenic consequences of the c.-16-1G>A variant and its potential correction through prime editing. Nevertheless, the functional conclusions presented herein are based primarily on in silico modeling and predictive algorithms. Experimental validation using mutant DLBCL cell lines is therefore essential to confirm the relative utilization of the predicted intron-retention and exon-skipping transcripts through approaches such as RT-qPCR or RNA sequencing. Furthermore, although predictive analyses have delineated the off-target profiles of the candidate sgRNAs, with sgRNA 3 demonstrating a more favorable safety profile than sgRNA 1, empirical laboratory-scale validation remains indispensable. Such validation is necessary to confirm the absence of unintended secondary nicking events in the target cells before advancing this platform toward in vivo evaluation using xenograft models and, ultimately, future personalized medicine applications.

5. CONCLUSION

The TMSB4X c.-16-1G>A variant is a splice-site alteration located at the canonical splice acceptor position that is absent from population databases and has been reported in DLBCL cases. Multiple splicing prediction algorithms consistently indicated a strong potential for disruption of normal RNA splicing. In silico analyses predicted two major aberrant splicing events, namely a 155-bp intron retention and exon 2 skipping. Both events were associated with alterations in transcript structure and open reading frame (ORF), resulting in truncated protein products compared with the 44-amino-acid wild-type Thymosin β4 protein. The predicted protein products comprised 32 amino acids in the intron-retention transcript and 22 amino acids in the exon-skipping transcript. In addition, this study successfully generated two candidate prime editing designs consisting of sgRNA–pegRNA combinations for precise correction of the TMSB4X c.-16-1G>A variant. Off-target analysis demonstrated favorable specificity profiles for both sgRNAs, with sgRNA 3 exhibiting the more advantageous off-target profile. Collectively, these findings provide a foundation for future experimental validation and support the development of prime editing–based strategies for the correction of pathogenic splice-site variants in TMSB4X.

Acknowledgment

The authors are grateful to the Master's Programme in Biomedical Sciences, Faculty of Medicine, Universitas Indonesia for the academic support provided during this research. We are also highly indebted to the Indonesia Endowment Fund for Education (LPDP) for the financial support that funded this project. Additionally, we express our gratitude to the developers of the bioinformatic tools utilized throughout this in silico study.

Conflict of interest

The authors declare no conflict of interest regarding the publication of this manuscript.

Funding

This research was funded by RIIM-BRIN/LPDP Research Grant (Grant number: 122/IV/KS/11/2023; 139/PKS/WRIII-DISTP/UI/2023; Grant receiver name: Noorwati Sutandyo).

Ethical statement

Ethical approval and patient informed consent were not required for this study as the research was conducted entirely in silico. The investigation focused on computational modeling of the TMSB4X c.-16-1G>A variant and predictive design of prime editing guide RNAs using publicly accessible genomic databases and bioinformatic tools. No human tissues, animal subjects, or live clinical samples were utilized in this study.

References

[1] Li S, Young KH, Medeiros LJ. Diffuse large B-cell lymphoma. Pathology 2018;50:74–87. https://doi.org/10.1016/j.pathol.2017.09.006.

[2] Testoni M, Zucca E, Young KH, Bertoni F. Genetic lesions in diffuse large B-cell lymphomas. Annals of Oncology 2015;26:1069–80. https://doi.org/10.1093/annonc/mdv019.

[3] Pasqualucci L, Dalla-Favera R. Genetics of diffuse large b-cell lymphoma. Blood 2018;131:2307–19. https://doi.org/10.1182/blood-2017-11-764332.

[4] Dobashi A. Molecular Pathogenesis of Diffuse Large B-Cell Lymphoma. Journal of Clinical and Experimental Hematopathology 2016;56:71–8. https://doi.org/10.3960/jslrt.56.71.

[5] Larouche J-F, Berger F, Chassagne-Clément C, Ffrench M, Callet-Bauchu E, Sebban C, et al. Lymphoma Recurrence 5 Years or Later Following Diffuse Large B-Cell Lymphoma: Clinical Characteristics and Outcome. Journal of Clinical Oncology 2010;28:2094–100. https://doi.org/10.1200/JCO.2009.24.5860.

[6] Vannata B, Conconi A, Winkler J, Cascione L, Margiotta Casaluci G, Nassi L, et al. Late relapse in patients with diffuse large B-cell lymphoma: impact of rituximab on their incidence and outcome. Br J Haematol 2019;187:478–87. https://doi.org/10.1111/bjh.16106.

[7] Jiang Y, Melnick A. The Epigenetic Basis of Diffuse Large B-Cell Lymphoma. Semin Hematol 2015;52:86–96. https://doi.org/10.1053/j.seminhematol.2015.01.003.

[8] Jayasinghe RG, Cao S, Gao Q, Wendl MC, Vo NS, Reynolds SM, et al. Systematic Analysis of Splice-Site-Creating Mutations in Cancer. Cell Rep 2018;23:270-281.e3. https://doi.org/10.1016/j.celrep.2018.03.052.

[9] Xing Y, Ye Y, Zuo H, Li Y. Progress on the Function and Application of Thymosin β4. Front Endocrinol (Lausanne) 2021;12. https://doi.org/10.3389/fendo.2021.767785.

[10] Kuo C-Y, Jhuang J-Y, Huang W-C, Cheng S-P. Aberrant Expression of Thymosin Beta-4 Correlates With Advanced Disease and BRAF V600E Mutation in Thyroid Cancer. Journal of Histochemistry & Cytochemistry 2022;70:707–16. https://doi.org/10.1369/00221554221138370.

[11] Yoon SY, Lee HR, Park Y, Kim JH, Kim SY, Yoon SR, et al. Thymosin β4 expression correlates with lymph node metastasis through hypoxia inducible factor-α induction in breast cancer. Oncol Rep 2011;25:23–31.

[12] Tang M-C, Chan L-C, Yeh Y-C, Chen C-Y, Chou T-Y, Wang W-S, et al. Thymosin beta 4 induces colon cancer cell migration and clinical metastasis via enhancing ILK/IQGAP1/Rac1 signal transduction pathway. Cancer Lett 2011;308:162–71. https://doi.org/10.1016/j.canlet.2011.05.001.

[13] Wu Y, Clark KC, Nguyen E V., Niranjan B, Horvath LG, Taylor RA, et al. Proteomic characterisation of prostate cancer intercellular communication reveals cell type-selective signalling and TMSB4X-dependent fibroblast reprogramming. Cellular Oncology 2022;45:1311–28. https://doi.org/10.1007/s13402-022-00719-z.

[14] Caers J, Hose D, Kuipers I, Bos TJ, Van Valckenborgh E, Menu E, et al. Thymosin 4 has tumor suppressive effects and its decreased expression results in poor prognosis and decreased survival in multiple myeloma. Haematologica 2010;95:163–7. https://doi.org/10.3324/haematol.2009.006411.

[15] Yang Y, Qin Z, Liu X, Yang H, Wang X, Mai C, et al. The implications of TMSB4X in TIM3 hypermethylation and CD8+ T cell exhaustion in diffuse large B-cell lymphoma. Sci Rep 2026. https://doi.org/10.1038/s41598-026-54670-2.

[16] Chen C, Li M, Yang H, Chai H, Fisher W, Yao Q. Roles of Thymosins in Cancers and Other Organ Systems. World J Surg 2005;29:264–70. https://doi.org/10.1007/s00268-004-7817-2.

[17] Cha H-J. Role of Thymosin 4 in Tumor Metastasis and Angiogenesis. CancerSpectrum Knowledge Environment 2003;95:1674–80. https://doi.org/10.1093/jnci/djg100.

[18] González-Rincón J, Méndez M, Gómez S, García JF, Martín P, Bellas C, et al. Unraveling transformation of follicular lymphoma to diffuse large B-cell lymphoma. PLoS One 2019;14:e0212813. https://doi.org/10.1371/journal.pone.0212813.

[19] Hahn CN, Venugopal P, Scott HS, Hiwase DK. Splice factor mutations and alternative splicing as drivers of hematopoietic malignancy. Immunol Rev 2015;263:257–78. https://doi.org/10.1111/imr.12241.

[20] Jian X, Boerwinkle E, Liu X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. Genetics in Medicine 2014;16:497–503. https://doi.org/10.1038/gim.2013.176.

[21] Anzalone A V., Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, et al. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature 2019;576:149–57. https://doi.org/10.1038/s41586-019-1711-4.

[22] Lopes da Costa B, Helms KM, Theodore K, Tsai Y-T, Caruso SM, Liu S, et al. Prime editing for the investigation of aberrant splicing defect associated with a pathogenic PRPH2 variant. Mol Ther Nucleic Acids 2025;36:102740. https://doi.org/10.1016/j.omtn.2025.102740.

[23] Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics 2019;35:1978–80. https://doi.org/10.1093/bioinformatics/bty897.

[24] Zhou H, Arapoglou T, Li X, Li Z, Zheng X, Moore J, et al. FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. Nucleic Acids Res 2023;51:D1300–11. https://doi.org/10.1093/nar/gkac966.

[25] Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Hum Mutat 2022;43:1012–30. https://doi.org/10.1002/humu.24309.

[26] Shamsani J, Kazakoff SH, Armean IM, McLaren W, Parsons MT, Thompson BA, et al. A plugin for the Ensembl Variant Effect Predictor that uses MaxEntScan to predict variant spliceogenicity. Bioinformatics 2019;35:2315–7. https://doi.org/10.1093/bioinformatics/bty960.

[27] de Sainte Agathe J-M, Filser M, Isidor B, Besnard T, Gueguen P, Perrin A, et al. SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation. Hum Genomics 2023;17:7. https://doi.org/10.1186/s40246-023-00451-1.

[28] Zeng T, Li YI. Predicting RNA splicing from DNA sequence using Pangolin. Genome Biol 2022;23:103. https://doi.org/10.1186/s13059-022-02664-4.

[29] Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res 2014;42:13534–44. https://doi.org/10.1093/nar/gku1206.

[30] Sullivan PJ, Quinn JMW, Wu W, Pinese M, Cowley MJ. SpliceVarDB: A comprehensive database of experimentally validated human splicing variants. The American Journal of Human Genetics 2024;111:2164–75. https://doi.org/10.1016/j.ajhg.2024.08.002.

[31] Canson DM, Davidson AL, de la Hoya M, Parsons MT, Glubb DM, Kondrashova O, et al. SpliceAI-10k calculator for the prediction of pseudoexonization, intron retention, and exon deletion. Bioinformatics 2023;39. https://doi.org/10.1093/bioinformatics/btad179.

[32] Gasteiger E. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31:3784–8. https://doi.org/10.1093/nar/gkg563.

[33] Chow RD, Chen JS, Shen J, Chen S. A web tool for the design of prime-editing guide RNAs. Nat Biomed Eng 2020;5:190–4. https://doi.org/10.1038/s41551-020-00622-8.

[34] Hwang G-H, Jeong YK, Habib O, Hong S-A, Lim K, Kim J-S, et al. PE-Designer and PE-Analyzer: web-based design and analysis tools for CRISPR prime editing. Nucleic Acids Res 2021;49:W499–504. https://doi.org/10.1093/nar/gkab319.

[35] Pliatsika V, Rigoutsos I. "Off-Spotter": very fast and exhaustive enumeration of genomic lookalikes for designing CRISPR/Cas guide RNAs. Biol Direct 2015;10:4. https://doi.org/10.1186/s13062-015-0035-z.

[36] Bae S, Park J, Kim J-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics 2014;30:1473–5. https://doi.org/10.1093/bioinformatics/btu048.

[37] Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. J Appl Genet 2018;59:253–68. https://doi.org/10.1007/s13353-018-0444-7.

[38] Reichel J, Chadburn A, Rubinstein PG, Giulino-Roth L, Tam W, Liu Y, et al. Flow sorting and exome sequencing reveal the oncogenome of primary Hodgkin and Reed-Sternberg cells. Blood 2015;125:1061–72. https://doi.org/10.1182/blood-2014-11-610436.

[39] González-Rincón J, Méndez M, Gómez S, García JF, Martín P, Bellas C, et al. Unraveling transformation of follicular lymphoma to diffuse large B-cell lymphoma. PLoS One 2019;14:e0212813. https://doi.org/10.1371/journal.pone.0212813.

[40] Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med 2017;9. https://doi.org/10.1126/scitranslmed.aal5209.

[41] Cao S, Zhou DC, Oh C, Jayasinghe RG, Zhao Y, Yoon CJ, et al. Discovery of driver non-coding splice-site-creating mutations in cancer. Nat Commun 2020;11:5573. https://doi.org/10.1038/s41467-020-19307-6.

[42] Schubach M, Maass T, Nazaretyan L, Röner S, Kircher M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. Nucleic Acids Res 2024;52:D1143–54. https://doi.org/10.1093/nar/gkad989.

[43] Tenywa J-F, Lamouche J-B, Baer S, Nicaise S, Le Béchec A, Piton A, et al. Genome region aware CADD thresholds for noncoding variant prioritization. NAR Genom Bioinform 2025;7. https://doi.org/10.1093/nargab/lqaf157.

[44] Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 2015;31:761–3. https://doi.org/10.1093/bioinformatics/btu703.

[45] Jang W, Park J, Chae H, Kim M. Comparison of In Silico Tools for Splice-Altering Variant Prediction Using Established Spliceogenic Variants: An End-User's Point of View. Int J Genomics 2022;2022:1–6. https://doi.org/10.1155/2022/5265686.

[46] Calabrese C, Davidson NR, Demircioğlu D, Fonseca NA, He Y, Kahles A, et al. Genomic basis for RNA alterations in cancer. Nature 2020;578:129–36. https://doi.org/10.1038/s41586-020-1970-0.

[47] Sibley CR, Blazquez L, Ule J. Lessons from non-canonical splicing. Nat Rev Genet 2016;17:407–21. https://doi.org/10.1038/nrg.2016.46.

[48] Okumura N, Yoshida H, Kitagishi Y, Nishimura Y, Matsuda S. Alternative splicings on p53, BRCA1 and PTEN genes involved in breast cancer. Biochem Biophys Res Commun 2011;413:395–9. https://doi.org/10.1016/j.bbrc.2011.08.098.

[49] Agrawal S, Eng C. Differential expression of novel naturally occurring splice variants of PTEN and their functional consequences in Cowden syndrome and sporadic breast cancer. Hum Mol Genet 2006;15:777–87. https://doi.org/10.1093/hmg/ddi492.

[50] Varley J, McGown G, Thorncroft M, White G, Tricker K, Kelsey A, et al. A novel TP53 splicing mutation in a Li-Fraumeni syndrome family: a patient with Wilms' tumour is not a mutation carrier. Br J Cancer 1998;78:1081–3. https://doi.org/10.1038/bjc.1998.631.

[51] GOLDSTEIN A, HANNAPPEL E, KLEINMAN H. Thymosin β: actin-sequestering protein moonlights to repair injured tissues. Trends Mol Med 2005;11:421–9. https://doi.org/10.1016/j.molmed.2005.07.004.

[52] Anzalone A V., Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, et al. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature 2019;576:149–57. https://doi.org/10.1038/s41586-019-1711-4.

[53] Swartjes T, Staals RHJ, van der Oost J. Editor's cut: DNA cleavage by CRISPR RNA-guided nucleases Cas9 and Cas12a. Biochem Soc Trans 2020;48:207–19. https://doi.org/10.1042/BST20190563.

[54] Dey D, Chakravarti R, Bhattacharjee O, Majumder S, Chaudhuri D, Ahmed KT, et al. A mechanistic study on the tolerance of PAM distal end mismatch by SpCas9. Journal of Biological Chemistry 2024;300:107439. https://doi.org/10.1016/j.jbc.2024.107439.