**Bridging big omic, genetic and medical data for Precision Medicine implementation in Greece**

***DELIVERABLE WP10.2***

*Technical Report on the Functional Impact Analysis of VUSs*

| | |
|---|---|
| ***Φορέας*** | Hellenic Pasteur Institute |
| ***Τύπος Παραδοτέου*** | Other |
| ***Ημερομηνία Υποβολής Παραδοτέου*** | 15 February 2026 |
| ***Ενότητα Εργασίας*** | Work Package 10<br>***Assessment of the significance of polymorphisms in regulatory and coding regions within gene bodies*** |

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

# 1 Introduction

Variants of uncertain significance (VUSs) remain a major limitation in translational genomics because many fall outside clearly protein-disrupting positions and therefore cannot be interpreted reliably through conventional coding-centric annotation alone. This limitation is particularly acute within gene bodies, where intronic and exon-adjacent variants can perturb pre-mRNA splicing or post-transcriptional regulation without necessarily producing obvious amino-acid changes. In cancer, splice-altering variation can remodel transcript structure, shift isoform usage, trigger nonsense-mediated decay, or activate cryptic splice sites—mechanisms that can influence tumour biology and clinical outcome.

Within EE10's broader objective of functionally annotating intragenic VUSs, and specifically in the context of Deliverable 10.2, this work expands functional impact assessment across two complementary regulatory axes. First, it addresses the splicing/isoform arm by integrating (i) deep-learning-based splice-site disruption prediction, (ii) isoform-resolved RNA-seq evidence, and (iii) clinical outcome association testing. Matched whole-genome sequencing (WGS) variants, RNA-seq, and clinical metadata from TCGA Breast Invasive Carcinoma (BRCA) and TCGA Bladder Urothelial Carcinoma (BLCA) were processed through a unified and reproducible workflow implemented in Snakemake, combining SpliceAI [1] annotation, transcriptome assembly and quantification, variant-aware differential isoform expression testing, integrated variant–transcript matching/classification, and survival analysis (Figure 1). Second, a post-transcriptional module evaluates allele-aware changes in miRNA-mediated targeting to prioritize variants with putative regulatory impact using effect-size scoring and outcome-oriented modelling. By applying the same framework across two tumour types, WP10.2 enables cohort-specific prioritization and cross-cohort comparability, while providing compact candidate sets for downstream mechanistic follow-up.

Overview of the WP10.2 integrative pipeline for splice-impact assessment in TCGA-BRCA and TCGA-BLCA

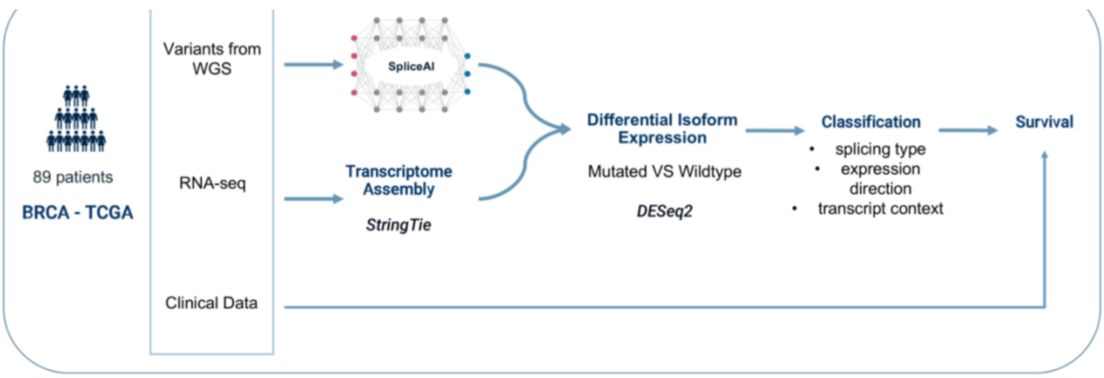


*Figure 1. **Matched multi-omic inputs from TCGA cohorts (BRCA and BLCA) (WGS variant calls, RNA-seq, and clinical data) are integrated in a unified workflow.** Candidate splice-impact variants are first annotated using SpliceAI, while RNA-seq data are processed through transcriptome assembly and quantification. Variant-aware differential isoform expression testing compares Mutated vs Wild-type groups, and results are consolidated into an interpretable classification of splice-event type, expression direction, and transcript context. Finally, prioritized candidates are evaluated using survival analysis to assess clinical association.*

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

# 2 Materials and Methods

## 2.1 Study design, cohorts, and data harmonization

The WP10.2 workflow was designed to functionally interpret gene-body VUSs by linking sequence-level splice-impact candidates to isoform-resolved transcript evidence and clinical outcome data in cohorts with matched multi-omic layers. Two TCGA cohorts were processed using the same end-to-end pipeline to ensure methodological comparability across tumour types: TCGA-BRCA and TCGA-BLCA. For each cohort, three data layers were assembled from the Genomic Data Commons (GDC): (i) somatic Whole Genomic Sequencing (WGS) variant callsets (used to derive the intragenic VUS candidate sets described in WP10.1), (ii) RNA-seq alignment files (BAM) for transcriptome reconstruction and isoform quantification, and (iii) clinical metadata for outcome modelling and covariate extraction.

To harmonize these layers, sample and patient identifiers were standardized to stable TCGA keys, enabling reproducible joins between variant tables, RNA-seq samples, and clinical records while minimizing ambiguity from aliquot- or file-level identifiers. Survival endpoints were derived from curated clinical follow-up fields (time-to-event from days-to-death, censoring from days-to-last-follow-up) and used to support outcome-oriented prioritization of splice-impact candidates. Where mutation-stratified analyses were performed, stability safeguards were applied (e.g., minimum carrier counts and cohort-specific feasibility checks) to reduce artefactual inference driven by ultra-rare events or sparse group sizes.

## 2.2 Variant preprocessing and splice-impact prediction (SpliceAI)

Somatic variant callsets were processed through a standardized preprocessing layer intended to support robust annotation, integration, and reporting across cohorts. Incoming VCFs were validated for format integrity and indexed to enable scalable querying. Where required, records were normalized to ensure consistent handling of multiallelic sites and allele representations and to preserve traceability from derived summary tables back to originating variant records.

Splice-impact prioritization was performed by annotating variants with SpliceAI, a deep-learning model that predicts variant-induced changes in splice acceptor and donor site usage from local sequence context. For each variant, SpliceAI reports four Δ-scores: Acceptor Gain (AG), Acceptor Loss (AL), Donor Gain (DG), and Donor Loss (DL), together with the Δ-position of maximum predicted change. Candidate splice-impact events were prioritized using a $\Delta \geq 0.20$ threshold, which is commonly used to balance sensitivity and specificity for downstream evaluation. SpliceAI annotations were retained in VCF INFO fields and parsed into structured tabular representations, including the maximal Δ score per variant and the corresponding event class, enabling downstream event summarization and integration with transcriptome evidence.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

### 2.3 RNA-seq processing, transcriptome reconstruction, and isoform quantification

RNA-seq alignments were processed to obtain isoform-resolved expression profiles suitable for detecting transcriptomic consequences of predicted splice-impact events. Alignments were generated using *STAR* [2] (two-pass mode consistent with standard processing), after which transcriptome assembly and quantification were performed using *StringTie* (v2.2.1) [3]. Transcript models were assembled per sample and subsequently merged into a cohort-level, non-redundant transcriptome using *StringTie --merge*, enabling a consistent transcript index for quantification and downstream comparisons within each cohort. The merged transcriptome was compared to GENCODE (v47) [4] using *gffcompare*, providing transcript-level and locus-level classifications (known versus novel structures) and concordance metrics. Isoform-level count matrices were generated using *prepDE.py* for statistical testing. Differential isoform expression analysis was then performed using *DESeq2* (v1.44.0) [5] to test whether predicted splice-impact variants are associated with shifts in transcript abundance or isoform usage patterns as reflected by transcript-level abundance differences in the reconstructed isoform space. For each candidate splice-impact variant, samples were partitioned into Mutated (carriers) and Wild-type (non-carriers) groups, and transcript-level counts were compared between groups. *SpliceAI* predictions were integrated with transcriptome-derived evidence to organize results into interpretable event-centric representations by combining event class (AG/AL/DG/DL), predicted Δ-position and maximal Δ-score with isoform-level differential expression outputs, yielding structured tables for downstream inspection and prioritization.

### 2.4 Event classification, survival analysis, and reproducibility

Candidate splice-impact variants were further evaluated for clinical relevance using survival analyses based on mutation status. For each candidate variant meeting minimum group size constraints (≥5 samples per group), overall survival differences between Mutated and Wild-type groups were tested using the log-rank test, and Kaplan-Meier curves were generated to visualize representative associations. Multiple hypothesis testing across variants was controlled using Benjamini-Hochberg false discovery rate (FDR) correction, with FDR $< 0.05$ used as the criterion for statistical significance. The workflow was implemented as a reproducible computational pipeline orchestrated with *Snakemake*, combining Python/R-based processing with standard genomics utilities for file handling, interval operations, and statistical analysis. Execution was organized to be cohort-repeatable with consistent directory structures, stable sample identifiers, and traceable intermediate artifacts to support auditing and re-execution. Outputs produced include *SpliceAI*-annotated variant tables, cohort-level transcriptome assemblies and annotation reports, isoform-level count matrices, *DESeq2* isoform results under variant-aware grouping, event-classified integrated result tables, and survival analysis summaries with corresponding Kaplan-Meier visualizations.

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης NextGenerationEU

## 2.5 miRNA-mediated post-transcriptional regulatory impact analysis

In parallel to splice-impact analysis, a post-transcriptional regulatory module was applied to evaluate whether intragenic variants—particularly those localized in 3'Untranslated Regions (UTRs) and, where relevant, Coding Sequences (CDS)/5'UTR contexts—may alter miRNA-mediated regulation. Candidate variants were restricted to those overlapping genomic windows defined around seed-based binding coordinates from highly expressed miRNAs (expression-guided selection using DIANA-miTED [6], a tissue-specific microRNA expression atlas developed by DIANA lab). These candidates were then annotated with curated miRNA response element (MRE) coordinates and region context using TarBase v9 [7], the reference database with experimental evidence of microRNA-gene interactions, enabling region-aware organization of the regulatory candidate set. Candidate catalogues and per-variant scoring tables are reported in WP10.1; WP10.2 focuses on functional interpretation, ranking logic, and outcome-oriented evaluation of high-impact candidates.

Variant-aware miRNA targeting effects were quantified using DIANA microT-CNN predictions [8] by comparing reference and alternative allele sequence contexts. For each reference–alternative allele pair, targeting scores were recalculated to estimate the direction and magnitude of predicted change, allowing classification of variants into (i) MRE-disrupting (loss-of-site), (ii) MRE-creating (gain-of-site), or (iii) modulatory events. A normalized disruption score using an appropriate custom scoring formula was computed per candidate interaction to support consistent ranking and prioritization across cohorts. Ranked candidate tables were produced for downstream interpretation and, where clinical metadata and minimum carrier counts permitted, candidates were further evaluated using outcome-oriented analyses analogous to the splice-impact module.

## 2.6 Disruption score computation and prioritization logic

To prioritize candidates in a comparable way across loci, miRNAs, and regulatory contexts, a normalized MRE disruption score was computed for each candidate variant based on allele-aware microT-CNN targeting predictions. The score summarizes predicted regulatory perturbation while preserving interpretability for downstream reporting by integrating (i) the direction of targeting change (gain vs loss) and (ii) the baseline strength of the affected interaction. Specifically, for each miRNA–candidate interaction, the disruption score was computed as:

$$MRE\ Disruption\ Score = \frac{(MT - WT)\ \times\ \log_2(1 + \frac{MT + WT}{2})}{\log_2(1.5)}$$

where WT and MT denote the predicted microT-CNN targeting scores for the reference and alternative allele, respectively. Positive values indicate predicted gain of targeting, whereas negative values indicate predicted loss of targeting, and the log-weighting upweights perturbations occurring in higher-confidence binding contexts.

Where multiple miRNA–MRE evaluations were available for the same locus (e.g.,

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης NextGenerationEU

overlap with more than one candidate interaction context), locus-level summaries were derived by aggregating across miRNAs using a consistent rule, reporting the maximum absolute disruption score per locus (and retaining the corresponding signed value to distinguish gain- from loss-dominated events). Candidates were then ranked by disruption magnitude and organized into tractable "high-priority" subsets (upper tail of the distribution), enabling concise reporting of top loci and facilitating targeted downstream evaluation. Disruption scores were retained alongside region context (e.g., 3'UTR/5'UTR/CDS labeling via curated MRE annotation and regulatory compartment labeling where applicable), enabling stratified summaries of whether stronger effects concentrate in specific gene-body contexts.

This scoring step converts large candidate tables into a ranked, effect-size–ordered landscape suitable for: (i) descriptive reporting of gain/loss score distributions, (ii) selection of top candidates for mechanistic evaluation, and (iii) downstream outcome-oriented analysis when clinical metadata are available.

### 2.7 Survival analysis and penalized modeling

To evaluate whether prioritized loci show association with clinical outcome, overall survival analyses were performed using mutation status as the stratification variable. Overall survival time was derived from TCGA clinical fields (days-to-death and days-to-last-follow-up), harmonized into a single time-to-event endpoint with censoring status defined accordingly. For each candidate locus, individuals were stratified into Carrier (variant present) and Wild-type (variant absent) groups.

Primary screening was performed using Kaplan–Meier survival curves and the log-rank test to assess differences between Carrier and Wild-type survival distributions. To reduce instability from extremely rare variants, analyses were restricted to loci meeting minimum carrier constraints (e.g., ≥5 carriers or a predefined minimum per group), and the resulting p-values were adjusted for multiple testing using Benjamini–Hochberg false discovery rate (FDR) correction where a large number of loci were screened.

In addition to non-parametric screening, Cox proportional hazards regression was used to estimate effect sizes as hazard ratios (HRs) with confidence intervals. Age at diagnosis (or age at index) was included as a covariate where available to reduce confounding. In settings where sparse carrier counts or separation could bias standard Cox estimates, a penalized likelihood approach (e.g., Firth correction) was used to improve coefficient stability. For multivariable prioritization across many correlated candidates, penalized Cox regression with LASSO regularization was applied to derive compact predictor sets, selecting the penalty parameter via cross-validation on partial likelihood. Model outputs were summarized through coefficient profiles and ranked predictor lists, supporting interpretation of directionality (risk vs protective) and relative contribution.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

## 2.8 Penalized Survival Analysis on the Integrated Set of MRE/Splicing Variants and Clinical Data

To assess the joint prognostic contribution of intragenic variants affecting MREs and splice-site regions, a two-stage penalized survival modelling strategy was applied, extending the approach established in WP9.

Candidate variants were drawn from two distinct functional classes. MRE variants were identified through the WP10 intragenic variant pipeline, retaining only those with prior evidence of predicted miRNA binding alteration from the microT-CNN disruption-score analysis. Splice-site variants (both SNPs and indels) were collected from the WP10 splicing analysis, each recording binary carrier status across the cohort. The two variant sets were merged with source labels to maintain provenance and deduplicated on genomic coordinates to avoid collinear predictors.

Prior to modelling, variants were filtered by a minimum carrier count (≥ 5 patients) to reduce instability from ultra-rare predictors. A binary mutation matrix was constructed and merged with clinical data, retaining patients with complete overall-survival endpoints and age at diagnosis. LASSO-penalized Cox proportional hazards regression was then applied to the combined predictor set (all retained variants plus age) using 10-fold cross-validation; the lambda.min solution was retained to define the sparse candidate set for downstream evaluation.

In the second stage, predictors selected by LASSO were refit using Firth's penalized Cox regression to obtain stable hazard ratio estimates, profile-likelihood confidence intervals, and p-values under rare-feature and separation-prone conditions where standard maximum-likelihood Cox estimation can diverge. Both univariate and multivariate Firth models were fitted, with the multivariate model including age as a baseline covariate. Global model significance was assessed using a penalized likelihood ratio test. Per-variant carrier and wildtype counts were computed alongside the Cox results, enabling transparent assessment of the support underlying each effect-size estimate.

Mutation-frequency distributions were visualised stratified by variant source, and cohort-level summary statistics, ncluding sample sizes, event rates, and variant counts at each filtering step, were recorded. The workflow generated standardised outputs per cohort: LASSO cross-validation curves, coefficient summaries, multivariate Firth Cox result tables with carrier counts, significance-annotated coefficient plots, and mutation-frequency diagnostics, forming the technical basis for the WP10 reporting package.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

# 3 Results

## 3.1 TCGA-BRCA results

### 3.1.1 Cohort assembly and analysis scale

After harmonization across molecular layers, a matched TCGA–BRCA cohort was assembled by requiring concurrent availability of somatic WGS variant callsets, RNA-seq alignments, and clinical metadata for overall survival modelling. From an initial pool of 113 candidate matched cases, filtering for complete data and consistent identifier linkage yielded a final integrative cohort of 89 BRCA cases (Table 1), which constituted the analysis set for splice-impact prediction, isoform-resolved transcriptome reconstruction, variant-aware differential isoform testing, and survival association analysis. This cohort definition ensured that each splice-impact candidate could be evaluated against both transcriptomic evidence and clinical outcome in a consistent per-patient framework, while maintaining sufficient sample size for mutation-stratified comparisons under minimum carrier-count constraints used in downstream analyses.

*Table 1. **Cohort assembly and final analysis set for TCGA–BRCA.** Summary of case counts from the initial matched candidate pool to the final integrative cohort retained after harmonization across WGS, RNA-seq, and clinical layers.*

| Stage | # cases (BRCA) | Description |
|---|---|---|
| **Initial matched candidate set** | 113 | Cases with matched molecular layers prior to final harmonization filters |
| **Final integrative cohort** | 89 | Cases retained after requiring complete WGS + RNA-seq + clinical metadata and consistent identifier linkage |

### 3.1.2 SpliceAI prediction yields a cohort-wide catalogue of splice-altering candidates

Splice-impact prioritization was performed by annotating the harmonized TCGA–BRCA somatic variant set with SpliceAI and retaining candidate events using a $\Delta \geq 0.20$ threshold (maximal $\Delta$ across AG/AL/DG/DL). This procedure identified 3,238 predicted splice-impact variants in the final BRCA cohort (Table 2), comprising 1,519 SNVs (47%) and 1,719 indels (53%). Indels showed a higher mean maximal $\Delta$-score than SNVs ($0.45 \pm 0.16$ vs $0.37 \pm 0.14$), consistent with stronger local sequence perturbations from multi-base changes. The resulting catalogue constitutes the primary splice-impact candidate set advanced to transcriptome-level interrogation and downstream event classification.

*Table 2. **SpliceAI splice-impact candidate catalogue in TCGA–BRCA.** Candidate variants were defined as those with maximal SpliceAI $\Delta$-score $\geq 0.20$ across the four event channels (AG/AL/DG/DL). Counts are reported by variant class (SNV vs indel), together with summary statistics of maximal $\Delta$-scores.*

| Variant class | # candidates | % of candidates | Mean max Δ-score | SD max Δ-score |
|---|---|---|---|---|
| **SNV** | 1519 | 47.0 | 0.37 | 0.14 |
| **Indel** | 1719 | 53.0 | 0.45 | 0.16 |
| **Total** | 3238 | 100.0 | NA | NA |

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης NextGenerationEU

### 3.1.3 Isoform-resolved transcriptome reconstruction supports transcript-level interrogation

To enable transcript-level evaluation of predicted splice-impact variants, an isoform-resolved transcriptome was reconstructed from the matched TCGA RNA-seq alignments using a reference-guided *StringTie* workflow. Briefly, RNA-seq BAMs (aligned to GRCh38 with STAR in two-pass mode via GDC standard processing) were assembled per sample and then merged into a cohort-level, non-redundant transcriptome (StringTie --merge). The resulting merged annotation was benchmarked against *GENCODE v47* using *gffcompare*, providing transcript class labels (known vs novel structures) and global concordance metrics. In the BRCA cohort, the merged assembly contained 447,219 transcript models across 74,558 loci, including 4,664 novel loci (~6%), while concordance with GENCODE was high (exon-level precision >91% and sensitivity ~90%) (Table 3). This transcriptome backbone was then used as the common reference for isoform quantification (re-quantification against the merged GTF) and for downstream variant-aware differential isoform testing and variant–transcript matching.

*Table 3.* ***Cohort-level transcriptome assembly summary for TCGA-BRCA.*** *RNA-seq alignments were assembled with StringTie and merged into a unified transcriptome, which was compared to GENCODE v47 using gffcompare. Reported are the size of the merged assembly, the number of loci, the fraction of novel loci, and global exon-level concordance metrics (precision and sensitivity).*

| Metric | Value |
|---:|:---:|
| Cohort | TCGA-BRCA |
| Merged transcriptome models (StringTie --merge) | 447,219 |
| Loci represented | 74,558 |
| Novel loci (count; ~%) | 4,664 (~6%) |
| Exon-level precision vs GENCODE v47 | >91% |
| Exon-level sensitivity vs GENCODE v47 | ~90% |

### 3.1.4 Variant-aware differential isoform expression identifies transcriptomic consequences of splice-impact candidates

To evaluate whether predicted splice-impact variants are associated with measurable transcriptomic changes, isoform-level differential expression testing was performed under a variant-aware grouping strategy. For each SpliceAI-prioritized candidate, samples were stratified into Mutated (carriers) and Wild-type (non-carriers) groups, and transcript-level counts derived from the merged cohort transcriptome were compared using *DESeq2*. This design enables detection of variant-associated shifts in transcript abundance and isoform usage patterns that are consistent with splice perturbation mechanisms (e.g., altered splice-site selection, exon skipping, or transcript destabilisation), and that may be missed by gene-level analyses.

In the BRCA cohort, significant isoform-level associations (FDR-adjusted) were detected for both SNVs and indels (Table 4). Specifically, the SNV result set contains 257 significant isoform hits, spanning 242 transcripts, 180 genes, and 199 distinct variant loci (chr:position). The indel result set contains 290 significant

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

isoform hits, spanning 281 transcripts, 200 genes, and 213 distinct variant loci. These variant-linked isoform signals were advanced to the integration and event-classification layer to relate SpliceAI-predicted event types (AG/AL/DG/DL) to expressed transcript structures and directionality of expression change.

*Table 4.* ***Summary of significant variant-associated isoform signals in TCGA–BRCA.*** *Isoform-level DESeq2 testing was performed for splice-impact candidates using Mutated (carrier) vs Wild-type (non-carrier) grouping. The table summarises the number of significant isoform hits (rows in the results tables) and the corresponding diversity of affected transcripts, genes, and distinct variant loci (chr:position) observed in the significant SNV and indel result sets.*

| Result set | Significant isoform hits | Unique transcripts | Unique genes | Distinct variant loci |
|---|---|---|---|---|
| **SNVs** | 257 | 242 | 180 | 199 |
| **Indels** | 290 | 281 | 200 | 213 |

### 3.1.5 Integrated variant–transcript matching and event classification produces an interpretable functional atlas

To consolidate splice-impact predictions with transcript-level evidence, SpliceAI-prioritized variants were integrated with the cohort transcriptome assembly and isoform-level differential expression outputs. For each event, the dominant SpliceAI class (acceptor gain/loss; donor gain/loss) and its predicted splice-site position were assessed against reconstructed transcript structures to determine whether the predicted splice-site change was supported by an observed transcript context. Variant–transcript associations were then grouped into three transcript-context categories: predicted (the predicted splice-site change aligns to a splice junction present in the differentially expressed transcript), other (alignment occurs in an alternative expressed isoform of the locus but not in the differentially expressed transcript), and none (no matching splice-site structure detected in the assembled models). These context assignments were further stratified by expression direction (up/down) and variant class (SNV vs indel), yielding an event-centric catalogue suitable for systematic prioritization and inspection.

Across significant isoform-linked events, the largest fraction of supported candidates fell into the **“**none**”** context class for both SNVs and indels (Table 5), indicating that many variant-associated isoform changes do not map cleanly to an explicitly reconstructed junction at the predicted SpliceAI Δ-position in short-read assemblies. Nonetheless, a non-trivial subset of candidates was classified as predicted or other, representing higher-confidence mechanistic links where the predicted splice-site perturbation can be localised to a specific expressed junction context. The resulting grouped tables constitute a compact functional atlas that can be mined to (i) prioritise transcript-supported splice-impact candidates for follow-up and (ii) separate candidates whose outcome associations may reflect mechanisms not readily captured as abundance shifts or junction-level reconstruction in short-read RNA-seq.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

*Table 5.* ***Integrated SpliceAI–transcriptome event classification landscape in TCGA–BRCA (significant isoform-linked events).*** *Splice-impact candidates were grouped by variant class (SNV/indel), dominant SpliceAI event type (AG/AL/DG/DL), and transcript-context category (predicted/other/none) based on whether the predicted splice-site change could be localised to reconstructed splice junction context in the assembled transcriptome. Counts are reported as unique variant loci (chr:position) and the number of significant isoform hits contributing to each class.*

| Variant class | Event class | Transcript context | # unique variant loci | # significant isoform hits |
|---|---|---|---|---|
| **SNV** | AG | predicted | 7 | 7 |
| **SNV** | AG | other | 13 | 24 |
| **SNV** | AG | none | 158 | 201 |
| **SNV** | AL | predicted | 4 | 7 |
| **SNV** | AL | other | 5 | 9 |
| **SNV** | AL | none | 10 | 13 |
| **SNV** | DG | predicted | 0 | 0 |
| **SNV** | DG | other | 0 | 0 |
| **SNV** | DG | none | 13 | 16 |
| **SNV** | DL | predicted | 2 | 4 |
| **SNV** | DL | other | 2 | 2 |
| **SNV** | DL | none | 5 | 5 |
| **Indel** | AG | predicted | 10 | 13 |
| **Indel** | AG | other | 19 | 26 |
| **Indel** | AG | none | 124 | 162 |
| **Indel** | AL | predicted | 7 | 9 |
| **Indel** | AL | other | 12 | 17 |
| **Indel** | AL | none | 19 | 28 |
| **Indel** | DG | predicted | 8 | 16 |
| **Indel** | DG | other | 5 | 19 |
| **Indel** | DG | none | 23 | 37 |
| **Indel** | DL | predicted | 4 | 11 |
| **Indel** | DL | other | 7 | 11 |
| **Indel** | DL | none | 17 | 26 |

### 3.1.6 Survival association highlights a small set of splice-impact candidates with prognostic signal

To link splice-impact candidates to clinical outcome, log-rank survival tests were performed across all predicted splice-altering variants (1,519 SNVs and 1,719 indels), comparing Mutated vs Wild-type groups and applying Benjamini–Hochberg FDR correction. While most variants did not retain significance after multiple testing adjustment, three indel loci reached FDR < 0.05 (Figure 1A–C): a 1 bp deletion at chr7:40014471 (TG→T; $p = 1.66\times10^{-5}$; FDR = 0.0194), a 3 bp insertion at chr4:170064704 (A→ACCT; $p = 2.26\times10^{-5}$; FDR = 0.0194), and a 1 bp deletion at chr7:55114899 (TG→T; $p = 3.60\times10^{-5}$; FDR = 0.0206). Notably, none of these loci was associated with significantly differentially expressed transcripts in the isoform-level analysis, suggesting that the observed prognostic associations may reflect mechanisms not captured as detectable transcript-level abundance shifts in short-read RNA-seq (e.g., subtle junction changes, NMD-sensitive products, stability/translation effects, or linkage with other alterations).

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

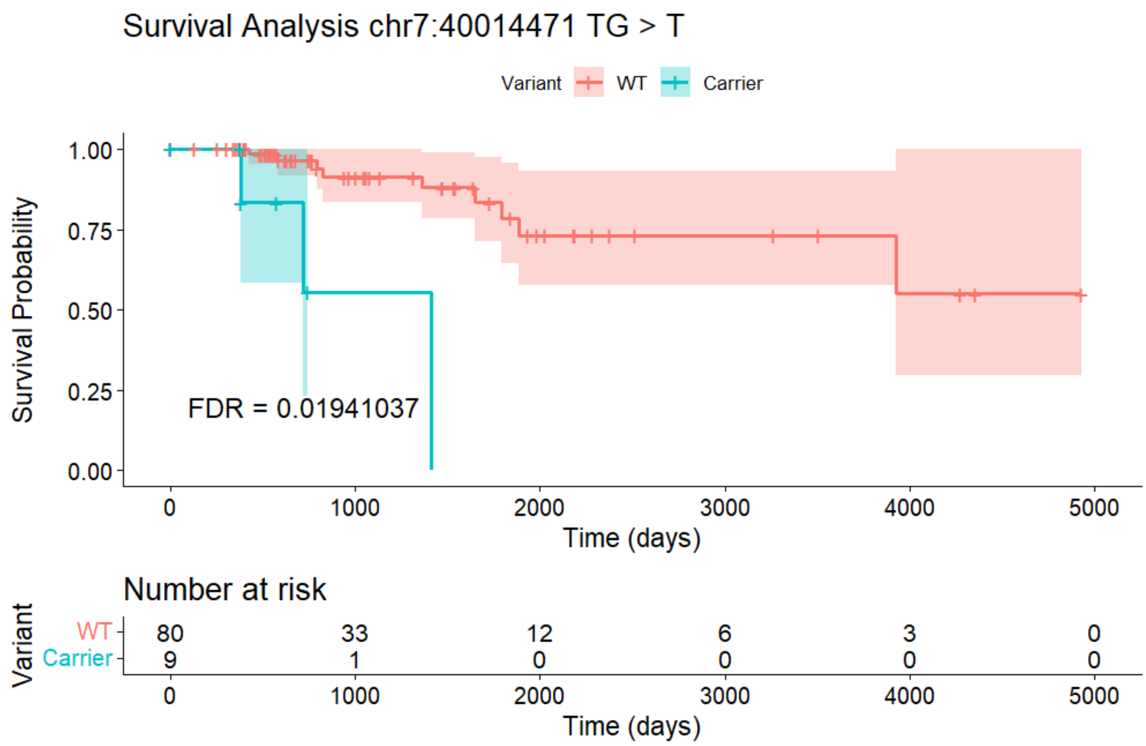


*Figure 1A.* ***Kaplan–Meier overall survival for a significant splice-impact indel at chr7:40014471 (TG>T; 1 bp deletion).*** *Overall survival is shown for TCGA–BRCA cases stratified by mutation status (WT vs carriers), with shaded confidence bands and the number-at-risk table below. This locus shows a pronounced early separation: carriers decline rapidly within the first ~1,000–1,500 days, reaching near-zero survival by ~1,500 days, whereas WT cases retain substantially higher survival over the same interval and continue to show long-term survivors throughout follow-up. The association remains significant after multiple-testing correction (FDR = 0.01941037). The at-risk table highlights the rarity and rapid attrition of the carrier group (WT n=80 vs carriers n=9 at baseline; 33 vs 1 at ~1,000 days; 12 vs 0 at ~2,000 days), indicating that most statistical evidence is concentrated in early follow-up where informative events occur. The widening carrier confidence band and loss of carriers beyond ~2,000 days underscore that effect interpretation is constrained by small carrier counts despite the strong early divergence.*


Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης NextGenerationEU

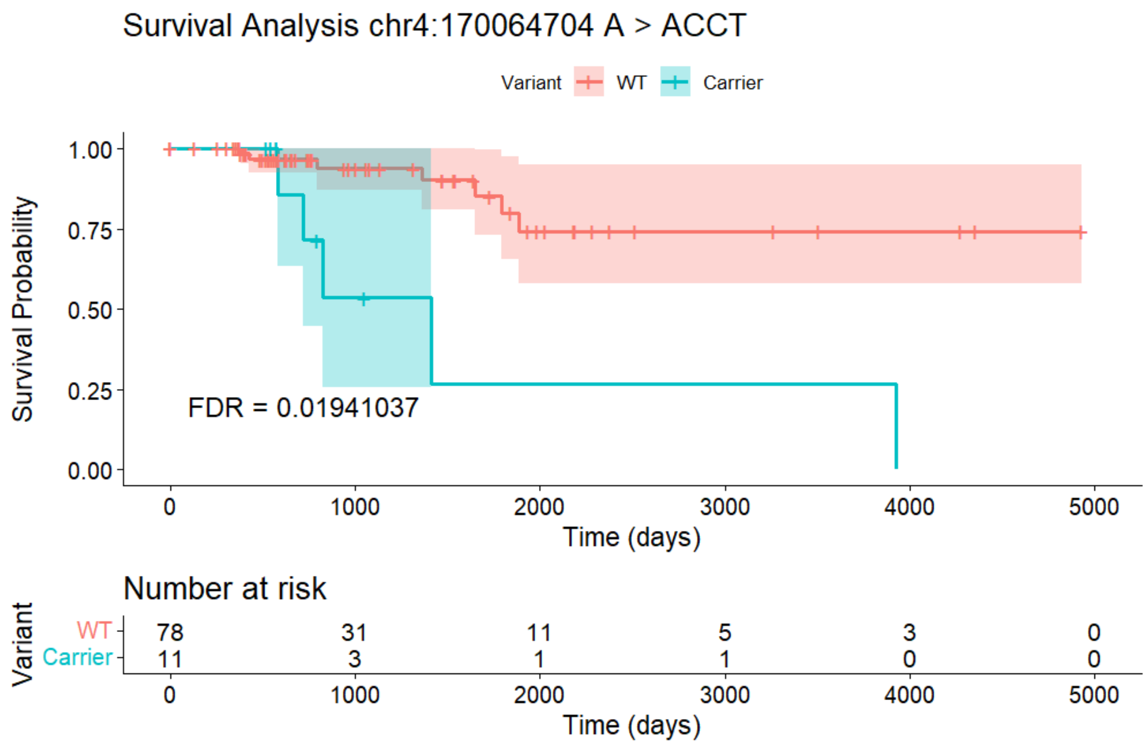


*Figure 1B.* ***Kaplan–Meier overall survival for a significant splice-impact indel at chr7:55114899 (TG>T; 1 bp deletion).*** *Kaplan–Meier curves compare WT and carrier groups with confidence bands and numbers at risk. Carriers exhibit an earlier and steeper decline in survival probability than WT cases, with separation emerging in the first ~1,000–2,000 days and persisting through mid follow-up. The association remains significant under Benjamini–Hochberg correction (FDR = 0.02064394). Group sizes again indicate a rare-event setting (WT n=77 vs carriers n=12 at baseline; 30 vs 4 at ~1,000 days; 11 vs 1 at ~2,000 days; 5 vs 1 at ~3,000 days; 3 vs 0 at ~4,000 days). As a result, the apparent long-term gap is supported primarily by early-to-mid follow-up, while late time points contribute limited information due to sparse carriers and broad confidence intervals.*

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

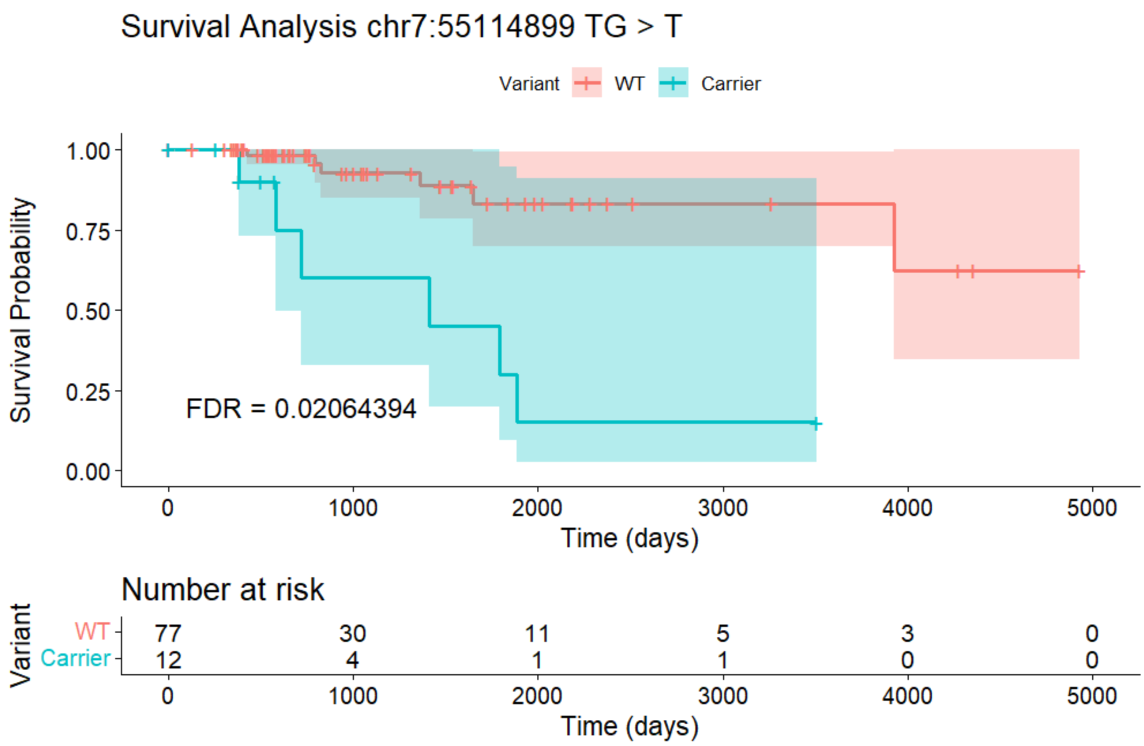


*Figure 1C.* ***Kaplan–Meier overall survival for a significant splice-impact indel at chr4:170064704 (A>ACCT; 3 bp insertion).*** *Overall survival is stratified by mutation status (WT vs carriers), with confidence intervals and numbers at risk shown. Carriers show a marked early drop in survival during the first ~1,000–1,500 days, while WT cases maintain higher survival over the same period and display sustained long-term survivors across follow-up. The log-rank association remains significant after multiple-testing correction (FDR = 0.01941037). The at-risk table indicates that inference is driven by a small carrier set (WT n=78 vs carriers n=11 at baseline; 31 vs 3 at ~1,000 days; 11 vs 1 at ~2,000 days; 5 vs 1 at ~3,000 days; 3 vs 0 at ~4,000 days), so late follow-up contributes little additional power and uncertainty increases as the carrier group collapses.*

### 3.1.7 miRNA candidate set evaluated for allele-aware post-transcriptional impact

In addition to splice-impact evaluation, the BRCA cohort was screened for intragenic variants overlapping expressed-miRNA binding windows to assess potential post-transcriptional regulatory consequences. Using binding-window intervals derived from highly expressed miRNAs (DIANA-miTED) and intersecting these coordinates with the harmonized BRCA somatic callset, a refined regulatory-context candidate set of 748 unique variants was obtained for downstream allele-aware rescoring and prioritization.

To support region-aware interpretation, candidates were mapped to curated MRE annotations (TarBase v9) to assign each variant to a gene-body regulatory context (3’UTR/5’UTR/CDS where supported by curated evidence). The resulting BRCA miRNA-overlap catalog was dominated by CDS-labeled candidates (650/748; 86.9%), with the remainder assigned to 3’UTR context (98/748; 13.1%) and no candidates assigned to 5’UTR under the applied mapping. These annotated candidate tables were advanced to allele-aware targeting prediction and disruption-score ranking to quantify putative regulatory impact.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

### 3.1.8 Allele-aware microT-CNN rescoring identifies gain- and loss-of-targeting events

To move beyond positional overlap and quantify functional directionality, candidate loci were evaluated using allele-aware microT-CNN rescoring under reference and alternative allele contexts. This enabled classification of candidates into gain-of-site**,** loss-of-site**,** or modulatory events based on the direction and magnitude of score change.

In BRCA, the score distributions indicate that most predicted events cluster in low-score ranges, consistent with predominantly modest quantitative shifts in predicted targeting strength, while a smaller subset occupies the upper tail and represents stronger candidate regulatory perturbations. In this reporting cycle, the gained-score histogram comprises 42 gain-of-site events, with the majority concentrated around a score of ~0.1 (Figure 2A), whereas the abolished-score histogram comprises 61 loss-of-site events, again concentrated in low-score ranges (Figure 2B). Collectively, these distributions support a prioritization strategy that focuses on the higher-score tails, where predicted perturbations of miRNA targeting are stronger and therefore more likely to be functionally informative.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

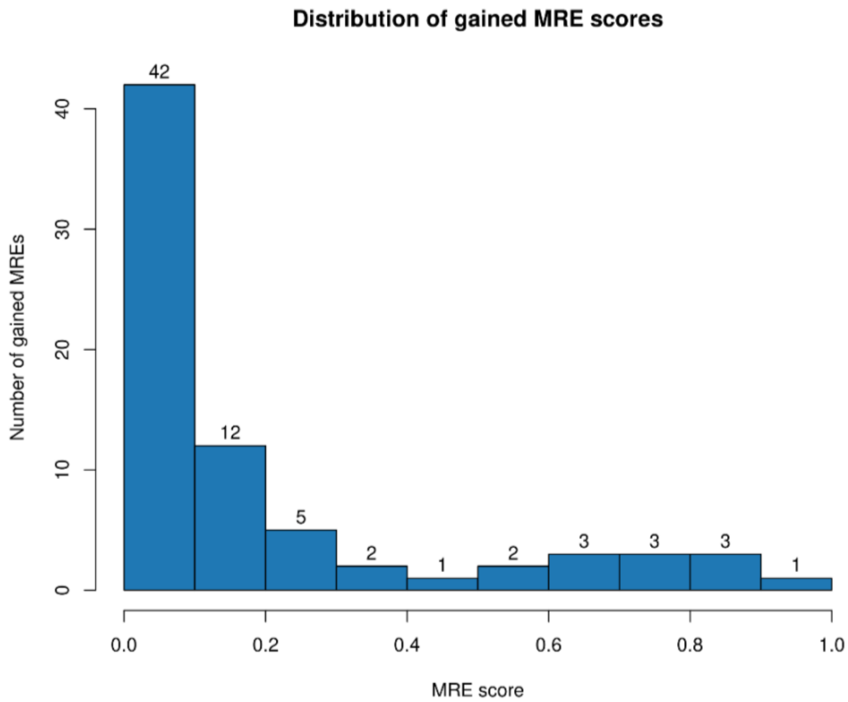


*Figure 2A.* ***Distribution of gained miRNA targeting scores in BRCA (microT-CNN).*** *Histogram of microT-CNN targeting scores for candidate interactions classified as gain-of-site under allele-aware rescoring. Scores were computed on the alternative-allele context and summarise the predicted strength of newly created/strengthened targeting.*

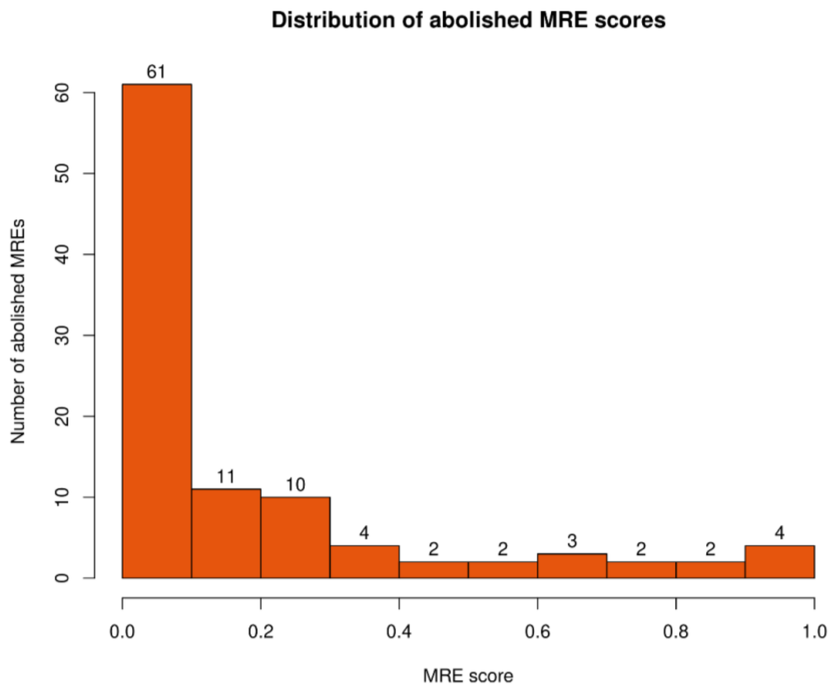


*Figure 2B.* ***Distribution of abolished miRNA targeting scores in BRCA.***
*Histogram of microT-CNN targeting scores for candidate interactions classified as loss-of-site under allele-aware rescoring. Scores correspond to the reference-allele context and summarise the predicted strength of targeting that is reduced or abolished by the variant.*


Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

### 3.1.9 Disruption-score landscape and prioritized high-impact candidates

To enable systematic ranking, a normalized disruption score was computed per miRNA–target interaction affected by a variant, integrating predicted targeting change between reference and alternative allele contexts into a single effect-size metric suitable for prioritization. The disruption-score distribution is strongly peaked near small-magnitude effects, with a narrower subset of higher-magnitude candidates forming the upper-tail prioritization landscape (Figure 2C). Ranking candidates by disruption score highlighted a compact set of high-impact gene-linked signals. The top-ranked genes in this cycle include **SECISBP2L, LAMC1, ANLN, ATM, EIF2B1, KLHL7, RYR2, ZBTB20, PIGY, and SLC17A5**, each associated with a specific miRNA context and a signed disruption direction (gain vs loss), providing an interpretable shortlist for mechanistic follow-up (Table 6).

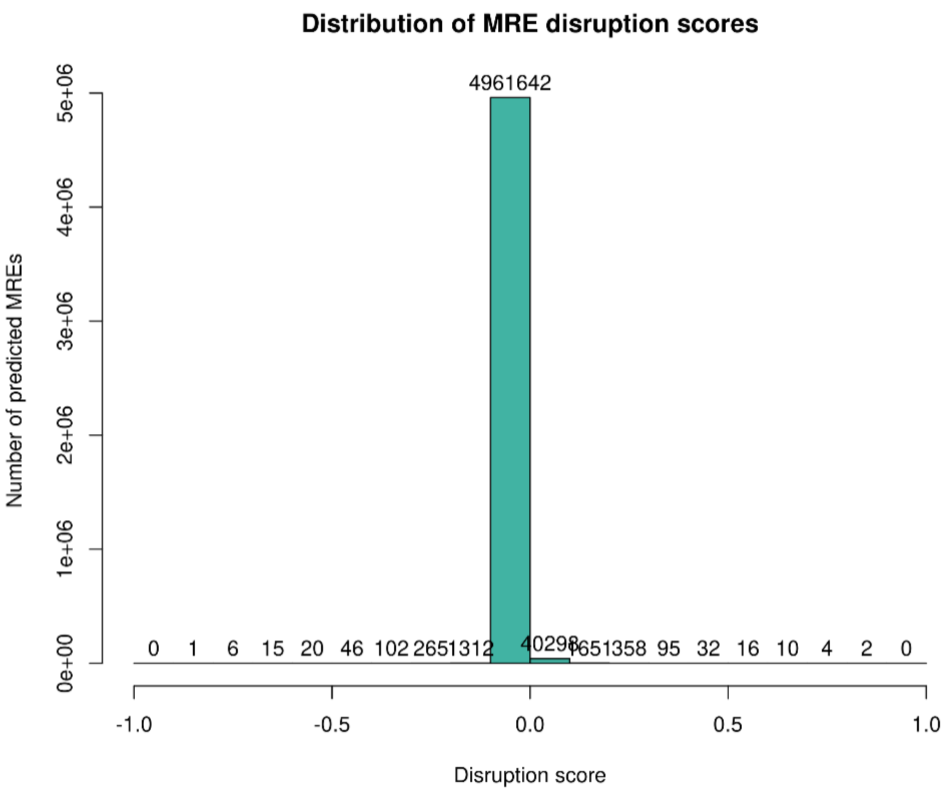


*Figure 2C. **Distribution of normalized miRNA disruption scores in BRCA.** Histogram of normalized disruption scores computed from paired reference vs alternative microT-CNN predictions, summarizing effect magnitude and direction across candidate interactions. The upper tail highlights higher-impact candidates used for prioritization.*

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης NextGenerationEU

*Table 6. **Top-ranked BRCA candidates by miRNA disruption score.***

| Gene | Disruption Score(s) | Expression Change | Associated miRNA | Proof of Expression Change |
|---|---|---|---|---|
| **SECISBP2L** | -0.8407 | Upregulated | hsa-miR-218-1-3p | Article Link |
| **LAMC1** | 0.8299 | Downregulated | hsa-miR-455-3p | Article Link |
| **ANLN** | 0.8186 | Upregulated | hsa-miR-141-5p | Article Link |
| **ATM** | -0.7878 | Downregulated | hsa-miR-30e-3p | Article Link |
| **EIF2B1** | -0.7793 | Not Associated | hsa-miR-424-3p | - |
| **KLHL7** | 0.7714 | Upregulation | hsa-miR-203b-5p | Article Link |
| **RYR2** | -0.7463 | Associated | hsa-miR-487a-5p | Article Link |
| **ZBTB20** | -0.7334 | Downregulated | hsa-miR-937-3p | Article Link |
| **PIGY** | 0.7314 | Not Associated | hsa-miR-142-3p | - |

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

| | | | | |
|---|---|---|---|---|
| **SLC17A5** | 0.7244 | Not Associated | hsa-miR-196a-5p | - |

### 3.1.10 Survival association analysis and multivariable prioritization

To evaluate whether a subset of miRNA-context candidate loci also carries prognostic signal, overall survival analyses were performed by stratifying patients into Carrier (mutated) versus Wild-type groups per locus. Clinical metadata were harmonised at the patient level (survival time derived from days-to-death or days-to-last-follow-up; age at diagnosis retained as a covariate), and survival curves were generated for each candidate to visually compare survival probability between groups. For locus-wise screening, Kaplan–Meier curves with log-rank testing were complemented by age-adjusted Cox proportional hazards models, yielding hazard-ratio estimates and confidence intervals for the carrier status effect.

In the BRCA cohort, mutation-status–stratified survival screening across eligible loci highlighted, within the microT-CNN–scored miRNA-context candidate set, PIK3CA-linked loci as notable candidates. Kaplan–Meier curves showed separation between carrier and wild-type groups for the strongest signal (Figures 3–4). In line with this, the age-adjusted Cox model indicated increased hazard for the top locus (Table 4), supporting its prioritization for downstream inspection in a broader regulatory context.

Because locus-wise screening is sensitive to multiple testing and correlation structure among predictors, a second stage applied penalised survival modelling (LASSO Cox) to jointly prioritise a compact set of predictors while controlling overfitting. Prior to penalisation, a minimal carrier-frequency filter was applied to exclude extremely rare loci (e.g., present in fewer than three patients), improving numerical stability. The LASSO penalty parameter was tuned by cross-validation (Figure 5), and the resulting sparse model retained a minimal predictor set that included age at diagnosis and the strongest variant-level signal. The direction and magnitude of the fitted coefficients (risk-increasing vs protective) are summarised in the coefficient barplot (Figure 6).

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

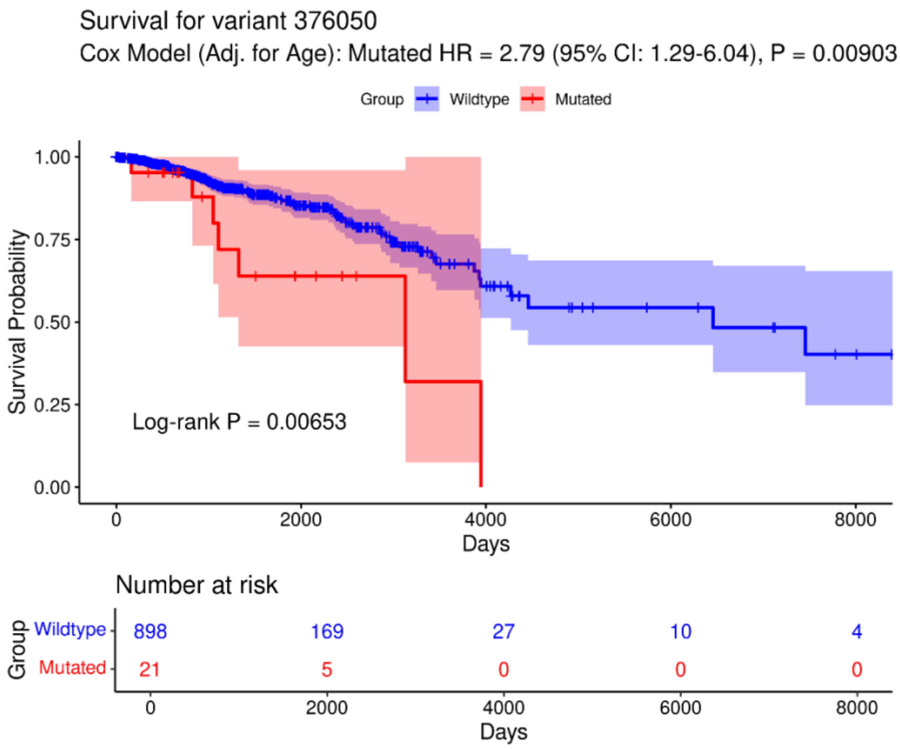


*Figure 3.* ***Kaplan–Meier overall survival for PIK3CA variant NM_006218.4:c.1035T>A.*** *Patients are stratified as Carrier (mutated) versus Wild-type. Age-adjusted Cox modelling indicates elevated hazard for carriers (HR = 2.79, 95% CI: 1.29–6.04; p = 0.009), with visible early separation of survival curves. Numbers at risk are shown below the plot.*


Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

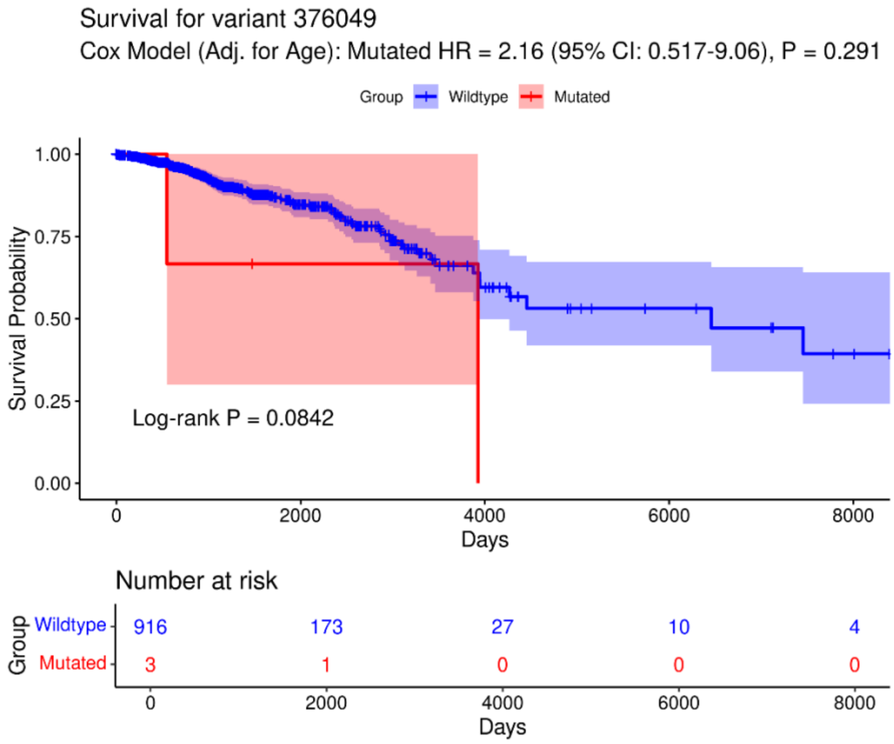


*Figure 4.* ***Kaplan–Meier overall survival for PIK3CA variant NM_006218.4:c.263G>A.*** *Kaplan–Meier curves comparing carriers and wild-type patients. Although separation is visible, the age-adjusted Cox model does not reach statistical significance (HR = 2.16, 95% CI: 0.52–9.06; p = 0.29), consistent with small carrier group size.*

*Table 7.* ***Univariable and age-adjusted Cox proportional hazards results for prioritized PIK3CA loci in TCGA–BRCA.***

| Predictor | HR_uni | CI_uni | P_value_uni | HR_multi | CI_multi | P_value_multi | Mutated_Patients | Wildtype_Patients | Mutation_Frecuency |
|---|---|---|---|---|---|---|---|---|---|
| **Age** | 1.037 | 1.012-1.062 | 0.00392 | 1.035 | 1.01-1.061 | 0.00647 | - | - | - |
| **376050** | 2.679 | 1.195-6.003 | 0.0167 | 2.818 | 1.252-6.344 | 0.0123 | 21 | 882 | 0.0233 |
| **376049** | 2.961 | 0.703-12.465 | 0.139 | 2.135 | 0.485-9.402 | 0.316 | 3 | 900 | 0.0033 |

Ελλάδα 2.0 ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης NextGenerationEU

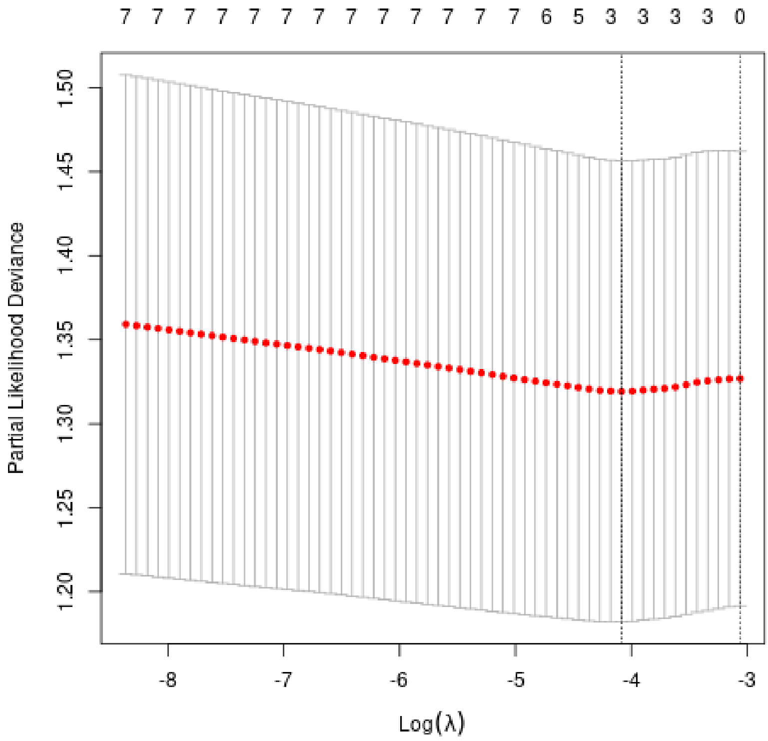


*Figure 5.* ***Cross-validation curve for penalized (LASSO) Cox regression.***
*Partial likelihood deviance plotted against log(λ). The optimal penalty parameter selected by cross-validation defines a sparse survival model retaining age and the strongest variant-level predictor.*

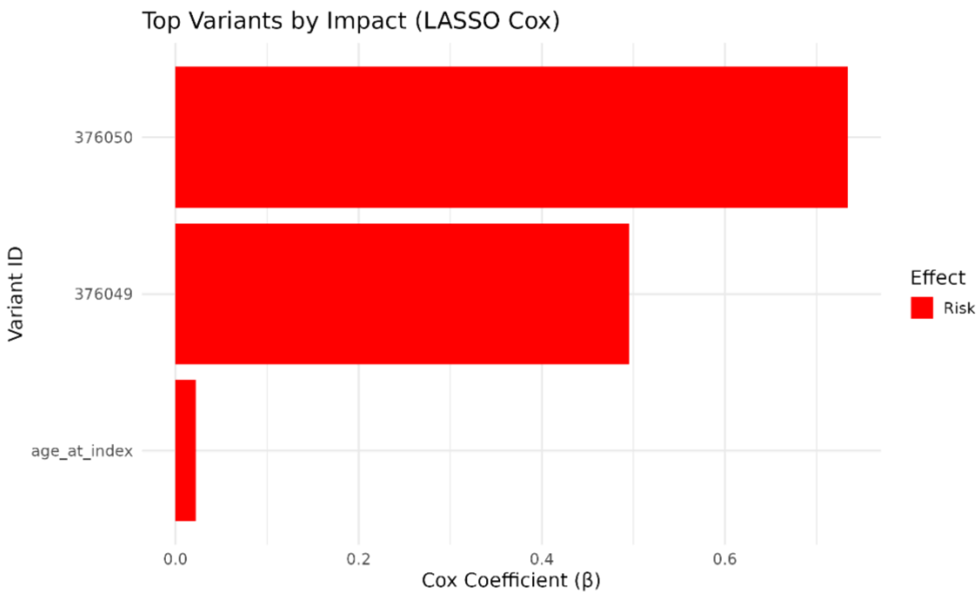


*Figure 6.* ***LASSO Cox coefficient estimates for retained predictors.***
*Barplot showing fitted coefficients (β) for variant-level predictors and age at diagnosis. Positive coefficients indicate risk-increasing effects.*

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

### 3.1.11 Feature selection and multivariate survival association

The analysis was conducted on the TCGA-BRCA cohort. From 903 clinical records, 443 patients had both variant and complete survival data (52 events, 11.7% event rate; follow-up range 5–8,605 days). A total of 4,255 candidate variants were assembled: 673 MRE variants from the microT-CNN–filtered analysis and 3,582 splice-site variants (1,520 SNPs, 2,062 indels). After applying the minimum carrier filter (≥ 5 patients), 3,534 variants entered the LASSO Cox model alongside age at diagnosis.

LASSO cross-validation selected 23 non-zero predictors (lambda.min = 0.0235): 22 variants and age at diagnosis. Of the 22 retained variants, 21 were splice-site loci and one was an MRE variant. Sixteen variants carried positive (risk-increasing) coefficients and six carried negative (protective) coefficients. Carrier frequencies among the selected variants ranged from 1.6% to 8.1% of the analysis cohort.

### 3.1.12 Multivariate survival association (Firth Cox refit) and prioritized variant

The 23 LASSO-selected predictors were refit using Firth's penalized Cox regression. The resulting multivariate model was highly significant at the global level (penalized likelihood ratio test: $\chi^2 = 80.01$, df = 23, $p = 3.16 \times 10^{-8}$; Table 1). Within this multivariate context, two predictors reached nominal significance ($p < 0.05$):

- **MRE variant at chr3:179,203,764 T>A** (affecting the hsa-miR-186-5p seed region): HR = 3.193, 95% CI 1.325–6.767, p = 0.012, observed in 21 carriers (4.7% of the cohort).
- **Age at diagnosis**: HR = 1.041, 95% CI 1.012–1.070, p = 0.005, confirming that the variant's prognostic effect is not confounded by baseline age differences.

No splice-site variant reached multivariate significance individually after Firth correction. This attenuation is expected when correlated rare predictors compete in a joint model with limited events, and it underscores the value of the LASSO pre-selection step in concentrating the signal.

From a biological perspective, the prioritized MRE variant is predicted by microT-CNN to disrupt a miR-186-5p binding site. miR-186-5p has been previously implicated in the regulation of proliferation and apoptosis pathways in breast cancer, lending functional plausibility to the observed adverse survival association. The carrier frequency of 4.7% places this variant in the low-frequency regime where Firth penalization is essential for unbiased estimation. Nevertheless, the moderate carrier count motivates replication in independent cohorts and integration with expression-level evidence where available. The co-selection of 21 splice-site variants alongside the MRE locus suggests that the combined regulatory burden, spanning both post-transcriptional (miRNA-mediated) and pre-mRNA processing (splicing) disruption, may jointly shape survival outcomes, consistent with the integrative rationale of WP10.

## 3.2 TCGA-BLCA results

### 3.2.1 Cohort assembly and analysis scale

A matched TCGA–BLCA cohort was assembled by harmonizing somatic WGS variant callsets, RNA-seq alignments, and clinical metadata using stable identifiers, following the same multi-omic integration logic applied to TCGA–BLCA. The resulting cohort definition ensured that splice-impact candidates—if present—could be evaluated against transcript-level evidence and clinical outcome within a consistent per-patient framework.

### 3.2.2 Downstream isoform testing not applicable in the absence of prioritized splice-impact events

Variant-aware differential isoform expression testing and integrated variant–transcript event classification were designed to evaluate transcriptomic consequences of SpliceAI-prioritized splice-impact candidates. Because no BLCA variants passed the splice-impact prioritization threshold (max $\Delta \geq 0.20$), no candidate set was available for downstream transcriptome reconstruction–linked interrogation, and the remainder of the splicing/isoform analysis workflow was not executed for BLCA in this reporting cycle. Accordingly, the BLCA branch is reported as a negative result with respect to detectable splice-impact candidates under the applied selection criterion.

## 4 Discussion

This report presents a cohort-repeatable framework for functional interpretation of intragenic variants of uncertain significance by integrating splice-impact prediction, isoform-resolved RNA-seq evidence, allele-aware miRNA targeting perturbation scoring, and outcome-oriented modelling in matched multi-omic cancer cohorts. In the BRCA cohort, the workflow connects sequence-level disruption hypotheses to transcript-level consequences and then to clinical association, providing a structured path from prioritization to interpretable candidate selection.

On the splicing/isoform axis, splice-impact prioritization produced a large catalogue of candidate splice-altering events that could be interrogated at the transcript level using reference-guided transcriptome reconstruction and isoform quantification. A notable practical outcome is that many variant-associated isoform signals could not be localized precisely to a reconstructed splice junction at the predicted splice-impact position, highlighting limitations of short-read transcript reconstruction for junction-level attribution in complex loci and rare-variant settings. Nevertheless, the subset of transcript-supported events provides higher-confidence mechanistic candidates for follow-up, and the event-classified summary tables help separate structurally supported signals from those likely reflecting subtler processing or stability effects. Survival screening identified a small number of splice-impact indel loci with prognostic signal after multiple-testing correction, although these were not accompanied by significant isoform-level abundance changes, consistent with mechanisms not captured by abundance shifts alone and emphasizing the need for replication and orthogonal validation.

In parallel, the miRNA module evaluates post-transcriptional regulatory perturbation by rescoring candidate sites under reference and alternative allele contexts and summarizing effects with a normalized disruption score that captures both directionality and baseline interaction strength. This converts overlap-based candidate sets into an effect-size–ordered landscape and yields a tractable high-priority subset for interpretation. Outcome-oriented evaluation combines locus-wise survival screening with multivariable modelling strategies tailored to sparse, correlated predictors, using penalized selection followed by bias-reduced refitting to obtain more stable effect estimates. An integrated analysis of miRNA-impact and splice-site variant classes further supports the view that prognostic signal can be distributed across regulatory layers and may be more informative when evaluated jointly.

In the BLCA cohort, no variants met the splice-impact prioritization threshold, and downstream isoform and event analyses were not applicable in this cycle, illustrating cohort dependence of candidate yield under fixed gating criteria. Overall, the report delivers an interpretable, reproducible template for prioritizing intragenic variants through complementary splicing and post-transcriptional mechanisms and provides compact candidate sets suitable for downstream validation and replication.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

# 5 References

[1] Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, Wenwu, Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., & Farh, K. K.-H. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell, 176(3), 535-548.e24. https://doi.org/10.1016/j.cell.2018.12.015

[2] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

[3] Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology, 33(3), 290–295. https://doi.org/10.1038/nbt.3122

[4] Mudge, J. M., Carbonell-Sala, S., Diekhans, M., Martinez, J. G., Hunt, T., Jungreis, I., Loveland, J. E., Arnan, C., Barnes, I., Bennett, R., Berry, A., Bignell, A., Cerdán-Vélez, D., Cochran, K., Cortés, L. T., Davidson, C., Donaldson, S., Dursun, C., Fatima, R., … Frankish, A. (2024). GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Research, 53(D1), D966–D975. https://doi.org/10.1093/nar/gkae1078

[5] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12). https://doi.org/10.1186/s13059-014-0550-8

[6] Kavakiotis, I., Alexiou, A., Tastsoglou, S., Vlachos, I. S., & Hatzigeorgiou, A. G. (2021). DIANA-miTED: a microRNA tissue expression database. Nucleic Acids Research, 50(D1), D1055–D1061. https://doi.org/10.1093/nar/gkab733

[7] Skoufos, G., Kakoulidis, P., Tastsoglou, S., Zacharopoulou, E., Kotsira, V., Miliotis, M., Mavromati, G., Grigoriadis, D., Zioga, M., Velli, A., Koutou, I., Karagkouni, D., Stavropoulos, S., Kardaras, F. S., Lifousi, A., Vavalou, E., Ovsepian, A., Skoulakis, A., Tasoulis, S. K., … Hatzigeorgiou, A. G. (2023). TarBase-v9.0 extends experimentally supported miRNA–gene interactions to cell-types and virally encoded miRNAs. Nucleic Acids Research, 52(D1), D304–D310. https://doi.org/10.1093/nar/gkad1071

[8] Zacharopoulou, E., Paraskevopoulou, M. D., Tastsoglou, S., Alexiou, A., Karavangeli, A., Pierros, V., Digenis, S., Mavromati, G., Hatzigeorgiou, A. G., & Karagkouni, D. (2024). microT-CNN: an avant-garde deep convolutional neural network unravels functional miRNA targets beyond canonical sites. Briefings in Bioinformatics, 26(1). https://doi.org/10.1093/bib/bbae678

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU