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

DELIVERABLE WP9.1

Technical report on the identified Variants of Uncertain Significance (VUSs) — report and presentation at an international conference

Φορέας	Hellenic Pasteur Institute
Τύπος Παραδοτέου	Other
Ημερομηνία Υποβολής Παραδοτέου	15 February 2026
Ενότητα Εργασίας	Work Package 9 ***Characterization of polymorphisms in intergenic regulatory regions***

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

A major barrier to the routine clinical deployment of genome-scale sequencing is that a large fraction of observed variants cannot be interpreted with sufficient confidence to inform biological inference or patient management. This challenge is particularly acute for non-coding variants that lie outside gene bodies, within the intergenic regulatory landscape where functional effects are mediated through altered transcriptional control rather than direct changes to protein sequence. Promoters and enhancers act as the genome's regulatory logic layer: they integrate transcription factor binding, chromatin accessibility, and three-dimensional (3D) chromatin organization to orchestrate context-specific gene expression programs. Consequently, sequence variation in these elements can rewire transcriptional outputs, contribute to oncogenesis, and modulate disease progression or therapeutic response, even when coding regions remain unchanged.

Work Package 9 addresses this interpretation bottleneck by focusing on Variants of Uncertain Significance (VUSs) located in intergenic regulatory regions, with emphasis on promoter-like and enhancer-like regulatory elements. The objective is to move from large, unstructured lists of non-coding variants to cohort-resolved catalogues of regulatory candidates that are functionally anchored and clinically contextualized. This requires a framework that (i) defines regulatory search spaces using standardized reference maps, (ii) incorporates tissue-relevant regulatory architecture where available, and (iii) supports downstream prioritization using clinical endpoints such as overall survival.

In WP9.1, the analysis concentrates on the systematic identification and characterization of regulatory VUSs in cancer cohorts, using TCGA Breast Invasive Carcinoma (BRCA) and Bladder Urothelial Carcinoma (BLCA) as validation settings. Candidate regulatory elements were defined using ENCODE SCREEN candidate cis-regulatory elements (cCREs) [1], enabling consistent genome-wide classification of promoter-like signatures (PLS) and enhancer-like signatures (proximal and distal). Beyond interval-level annotation, regulatory interpretation was strengthened by integrating complementary functional layers: (i) gene–regulatory element associations supported by ENCODE 3D chromatin interaction evidence (e.g., Hi-C and ChIA-PET), which provides mechanistic linkage between distal enhancers and their putative target genes; (ii) transcription factor binding site (TFBS) annotations derived from JASPAR [2] , to capture motif-centric perturbations of transcriptional control; and (iii) expression quantitative trait loci (eQTL) evidence from ENCODE, to identify regulatory variants with documented associations to transcriptional output. Together, these layers enable construction of a regulatory VUS landscape that is not merely positional, but functionally structured and interpretable in a tissue-aware context.

A core deliverable requirement of WP9.1 is to assess whether identified regulatory VUSs carry prognostic value within each cohort. This is statistically challenging because regulatory variants are frequently rare and the number of candidate predictors can

greatly exceed the number of observed clinical events. Accordingly, the analysis strategy is designed to operate in a rare-variant, high-dimensional regime while maintaining interpretability. Survival analyses are performed by stratifying patients according to variant carrier status and evaluating associations with overall survival using Cox proportional hazards models, accompanied by multiple-testing control to limit false discoveries across large candidate sets. Where necessary to stabilize inference under sparse carrier counts, penalized modelling approaches can be employed to mitigate separation and overfitting, ensuring that effect estimates are finite and that selected candidates represent robust prioritization signals rather than artefacts of limited event counts.

In this deliverable, the primary objective is outcome-oriented prioritization: candidate regulatory VUSs are defined by overlap with standardized regulatory maps (ENCODE cCRE promoter/enhancer annotations) and then evaluated for prognostic association with overall survival using modelling strategies designed for sparse, rare-variant settings. Additional regulatory layers (e.g., TFBS and external cross-referencing) are evaluated as complementary annotation streams; however, the central deliverable output is a cohort-resolved set of regulatory loci with measurable survival association under penalized inference.

2 Materials and Methods

2.1 Cohorts, endpoints, variant inputs, and regulatory annotation framework

Analyses were conducted in TCGA cancer cohorts with available somatic variant callsets and clinical follow-up, focusing on BRCA and BLCA. Overall survival was derived from TCGA clinical fields using time-to-event (days to death) and censoring time (days to last follow-up), and demographic covariates (e.g., age at diagnosis) were retained to support multivariable modelling. Somatic variant callsets were processed through a standardized preprocessing layer to ensure consistent representation across samples and cohorts, including format validation, normalization of allele representations, and coordinate harmonization for interval-based overlap operations. Candidate regulatory VUSs were then defined by intersecting variants with a curated regulatory search space emphasizing promoters and enhancers. ENCODE SCREEN cCREs were used as the primary regulatory backbone, enabling classification of variant-overlapping regulatory elements into promoter-like and enhancer-like categories in a genome-wide and standardized manner. In parallel, a motif-centric annotation stream based on JASPAR TFBS resources was evaluated to provide a complementary representation of regulatory disruption focused on predicted TF binding sequence features rather than regulatory-element intervals. Together, these annotation streams provide a consistent and interpretable definition of intergenic regulatory space while supporting sensitivity to both element-level and motif-level regulatory hypotheses.

2.2 Cohort-wise feature matrix construction and filtering for modelling stability

For each cohort and each regulatory annotation stream, variants were encoded into a patient-by-feature matrix suitable for outcome-oriented modelling. The primary representation was a binary indicator per individual (carrier vs non-carrier) for each candidate regulatory feature, enabling direct stratification and interpretation in survival models. Because intergenic regulatory variation is typically sparse and many variants are cohort-rare, a frequency-based filtering step was applied prior to modelling to reduce instability driven by ultra-rare predictors (e.g., removing features observed in fewer than a minimum number of individuals). Additional harmonization checks were performed to prevent redundant predictors arising from duplicated locus encodings (for example, position-only identifiers alongside position-plus-allele identifiers), which can introduce exact collinearity and artefactual model selection; such duplicates were removed or unified using a consistent locus representation. This processing yields a tractable, cohort-resolved regulatory-variant feature space that preserves interpretability while meeting the practical requirements of penalized survival modelling.

2.3 Penalized survival modelling, inference, and reporting artefacts

To prioritise regulatory VUSs for prognostic relevance in the high-dimensional, low-carrier regime, a two-stage modelling strategy was applied. First, LASSO-penalised Cox proportional hazards regression was used for feature selection. Prior to modelling, candidate loci were filtered by minimum carrier frequency to reduce instability from ultra-rare predictors and to ensure sufficient support for survival estimation. Cross-validation was then used to determine the regularisation strength, and the lambda.min solution was retained to define a sparse candidate predictor set for downstream evaluation. This stage reduces model complexity and mitigates overfitting when the number of candidate regulatory predictors is large relative to the number of observed survival events. Second, predictors selected by LASSO were refit using Firth's penalised Cox regression to obtain stable hazard ratio estimates, confidence intervals, and p-values under rare-feature and separation-prone conditions where standard Cox estimation can diverge. Multivariate models included age as a baseline covariate, and global model significance was assessed using a penalised likelihood ratio test alongside per-predictor inference.

To contextualise sparsity and support cautious interpretation, mutation-frequency distributions were computed and visualised, and cohort-level summary statistics (sample size and event rate) were reported. Finally, selected loci were optionally cross-referenced against population resources (e.g., a UK Biobank subset of the EBI GWAS Catalog [4]) to identify overlaps with previously reported associations, recognising that ultra-rare prognostic signals in cancer cohorts may not be represented in population GWAS resources. The workflow generated standardised outputs per cohort, including LASSO cross-validation curves, coefficient summaries, multivariate Firth Cox result tables (with carrier counts), and mutation-frequency diagnostics, forming the technical basis for the WP9.1 reporting package.

3 Results

3.1 TCGA-BRCA: regulatory VUSs and survival association

3.1.1 Cohort scale and modelling regime

The BRCA analysis was performed in a setting typical for regulatory-variant prioritization in cancer cohorts: a large number of candidate predictors but comparatively few events. After cohort assembly, the BRCA dataset comprised 888 patients with 95 deaths (event rate 10.7%), and 1,389 candidate regulatory variants were retained for modelling after the regulatory intersection and preprocessing steps. This combination (high dimensionality relative to events) motivates the use of penalized modelling and requires cautious interpretation of rare-variant effects as prioritization signals rather than definitive causal evidence.

3.1.2 LASSO feature selection and stability of the selected solution

Given the high dimensionality of the regulatory-variant feature space relative to the number of observed events in BRCA (1,389 candidate predictors versus 95 deaths), penalised survival modelling was used to identify a compact set of candidate loci with prognostic signal. LASSO-penalised Cox regression was therefore fit using cross-validation to select the regularisation strength, and the lambda.min criterion was used to define the selected model (Figure 1). The cross-validation profile shows that stronger regularisation rapidly reduces model complexity, and the selected solution converges to a highly sparse predictor set, consistent with the expected sparsity and low carrier frequencies of intergenic regulatory variants in TCGA cohorts.

Importantly, the sparsity of the selected model is itself informative: it indicates that, under the available event rate and cohort size, only a very small subset of regulatory loci provide stable incremental prognostic information beyond baseline covariates. In practice, this behaviour is desirable in rare-feature regimes because it limits overfitting and reduces sensitivity to noise features that may appear associated due to chance co-occurrence in a small number of carriers. The selected model was carried forward to penalised refitting using Firth Cox regression to obtain robust effect-size estimates and confidence intervals under rare-variant conditions.

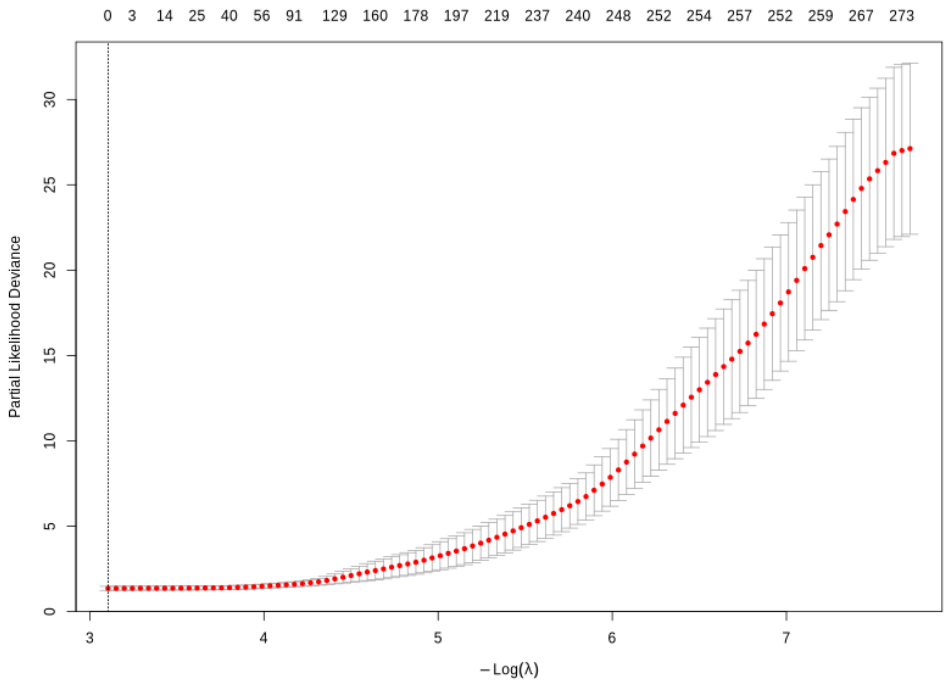


Figure 1. ***Cross-validation curve for LASSO-penalized Cox regression in TCGA–BRCA****. The selected penalty (lambda.min) defines the sparse model used for downstream refitting and inference.*

3.1.3 Multivariate survival association (Firth Cox refit) and prioritized variant

The predictors retained by LASSO were refit using Firth's penalised Cox regression to obtain stable multivariate effect-size estimates under rare-variant conditions. The resulting model, comprising age at diagnosis and a single ENCODE cCRE promoter-associated variant, was highly significant at the model level (penalised likelihood ratio test: $\chi^2 = 20.88$, df = 2, $p = 2.93\times10^{-5}$; Table 1). Within this multivariate context, the regulatory locus chr1:210,233,993 TT→AA (ENCODE promoter category) showed a strong adverse association with overall survival (HR = 10.60, 95% CI 2.927–26.964, p = 0.00154), despite being observed in only four carriers (Table 1). Age remained independently associated with outcome (HR = 1.029, 95% CI 1.011–1.047, p = 0.00106), indicating that the variant's effect is not explained by baseline age differences alone. The coefficient summary (Figure 2) highlights that the promoter-associated locus dominates the fitted risk signal relative to the modest but consistent age contribution.

From a biological and translational perspective, the fact that the selected variant maps to a promoter-like regulatory element is consistent with the conceptual model of intergenic regulatory VUSs: even rare changes in promoter-proximal sequence may perturb transcriptional control and downstream gene-expression programs. However, the low carrier frequency implies that the hazard-ratio estimate primarily serves as a prioritisation signal rather than a definitive biomarker claim, motivating replication in additional cohorts and integration with complementary evidence (e.g., regulatory activity context and expression-level correlates where available).

Table 1. ***Multivariate Firth Cox regression results for TCGA–BRCA regulatory-variant predictors selected by LASSO.*** *The model includes age at diagnosis and the ENCODE promoter-associated locus chr1:210,233,993 (TT→AA). Reported are penalised hazard ratios (HR), 95% confidence intervals, p-values, and carrier counts, together with the global penalised likelihood ratio test for the fitted model.*

Predictor	Regulatory category	Region	HR (Firth)	95% CI	p-value	Carriers
chr1:210233993 TT>AA	ENCODE cCRE (PLS/ELS)	Promoter	10.60	2.927 - 26.964	0.00154	4.0
age_at_index	—	—	1.029	1.011 - 1.046	0.00106	—

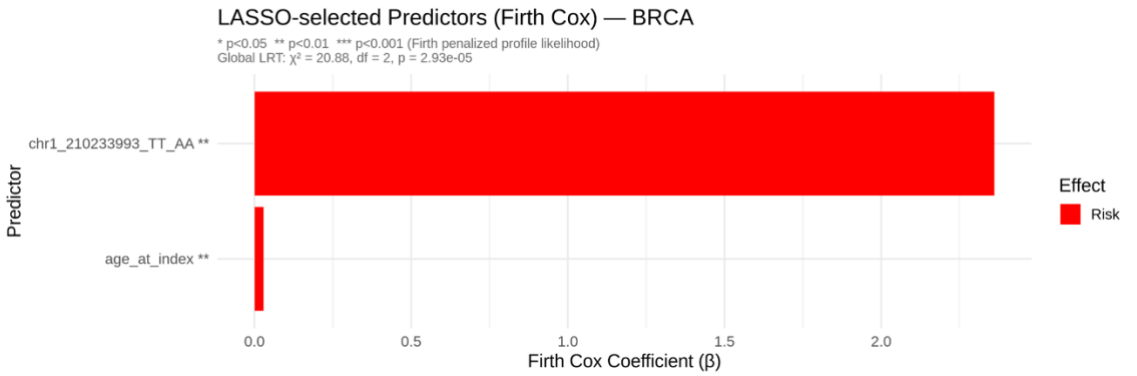


Figure 2. ***Coefficient summary for the multivariate Firth Cox model in TCGA–BRCA.*** *Regression coefficients are shown for age and the ENCODE promoter-associated locus chr1:210,233,993 (TT→AA), illustrating direction and relative magnitude of the fitted effects under penalized inference.*

3.1.4 Variant frequency structure and interpretability constraints

The carrier-frequency distribution of regulatory-variant predictors in BRCA demonstrates a strongly sparse and long-tailed feature space, in which the majority of candidate loci are observed in only a very small number of individuals (Figure 3). This pattern is expected for intergenic regulatory variation in somatic TCGA callsets after restricting to promoter/enhancer annotations, and it has direct implications for statistical modelling: ultra-rare predictors increase the risk of separation, inflate effect-size estimates under standard Cox regression, and amplify sensitivity to chance co-occurrence with events. The observed frequency structure therefore provides a practical justification for (i) applying minimum-carrier filtering prior to modelling and (ii) using penalised approaches (LASSO for selection and Firth Cox for stable refitting).

Within this frequency context, the ENCODE promoter-associated locus prioritised in the multivariate model is itself rare (4 carriers; ~0.44% of the cohort), reinforcing that the inferred hazard ratio should be interpreted cautiously. Rather than being treated as a definitive biomarker on its own, this locus is best viewed as a high-priority candidate emerging from an outcome-oriented prioritisation framework, motivating follow-up through replication in independent cohorts and integration with orthogonal evidence (e.g., regulatory

activity context or expression-level correlates where available). At the cohort level, Figure 3 therefore serves both as a diagnostic of modelling regime and as an interpretability anchor for understanding why a sparse penalised modelling strategy was required for regulatory VUS prioritisation in BRCA.

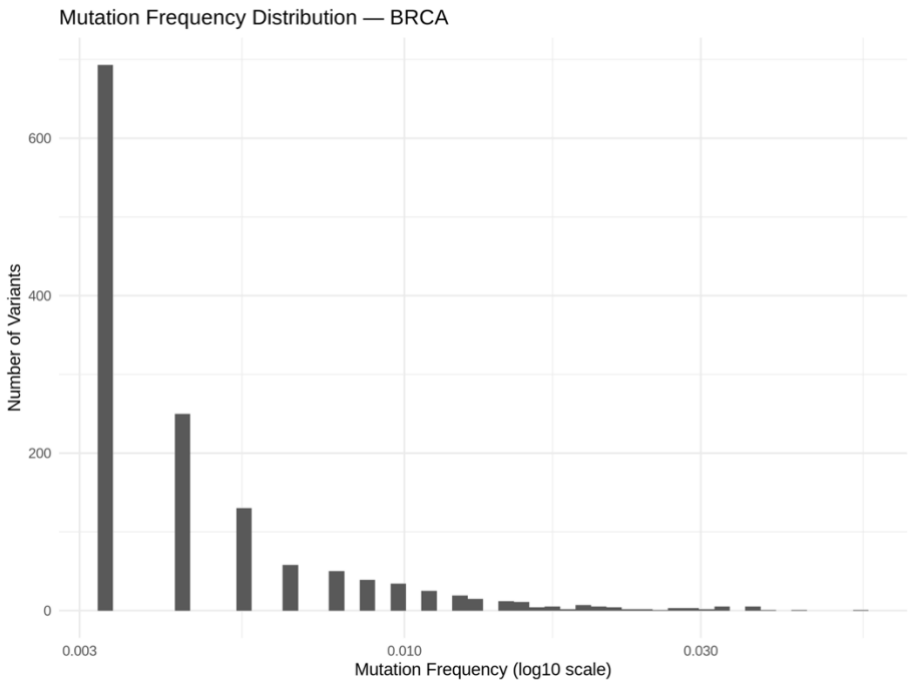


Figure 3. ***Carrier-frequency distribution of candidate regulatory-variant predictors in TCGA–BRCA (log10 scale).*** *The predominance of low-carrier features motivates penalised survival modelling and cautious interpretation of rare-locus effect sizes.*

3.2 TCGA-BLCA: regulatory VUSs and survival association

3.2.1 Cohort scale, event rate, and candidate regulatory feature space

The BLCA analysis was performed in a cohort with a substantially higher event rate than BRCA, providing improved statistical power for survival modelling under a rare-feature regulatory-variant regime. After cohort assembly and harmonisation of clinical endpoints, the BLCA dataset comprised 326 patients with 143 deaths (event rate 43.9%), enabling more stable estimation of survival associations relative to cohorts with sparse events. Following regulatory-interval intersection, preprocessing, and minimum-support filtering, 298 candidate regulatory variants were retained for modelling. In BLCA, an explicit deduplication step was additionally required to remove redundant locus encodings (e.g., alternative identifier formats describing the same genomic position, including allele-specific representations), ensuring that each candidate regulatory locus was represented consistently and only once in the modelling matrix and avoiding collinearity artefacts in downstream penalised selection.

This cohort-level configuration (moderate sample size, high event rate, and hundreds of sparse regulatory predictors) is well-suited to penalised Cox modelling: the event rate supports selection of a multivariate regulatory signature, while penalisation remains essential because individual regulatory variants are typically carried by only a small number of cases. The subsequent modelling therefore applies the same outcome-oriented framework as in BRCA (penalised feature selection followed by penalised refitting for inference), but with the expectation that BLCA can support a richer selected predictor set due to the increased number of observed events.

3.2.2 LASSO feature selection and composition of the selected regulatory signature

To prioritise intergenic regulatory VUSs with prognostic signal in BLCA while controlling overfitting in a sparse feature space, LASSO-penalised Cox regression was applied with cross-validation to select the regularisation strength. The cross-validation profile (Figure 4) supports a multivariate solution at lambda.min, indicating that the available number of events in BLCA is sufficient to retain a modest set of regulatory predictors without collapsing to a single-locus model. Compared with BRCA, the BLCA cohort's higher event rate (43.9%) enables a larger selected feature set and therefore a richer candidate regulatory "signature" for downstream inference.

Using the lambda.min solution, the selected predictor set comprised 12 ENCODE cCRE loci, dominated by promoter-like features (9 promoter loci) with a smaller enhancer-like component (3 enhancer loci), together with age at diagnosis as a baseline covariate. This category composition suggests that promoter-proximal regulatory space carries much of the cohort-level prognostic signal detectable under this framework, while still allowing a subset of enhancer-associated loci to contribute. The selected set was carried forward to penalised multivariate

refitting using Firth Cox regression to obtain stable effect estimates and to evaluate which loci remain individually significant when modelled jointly.

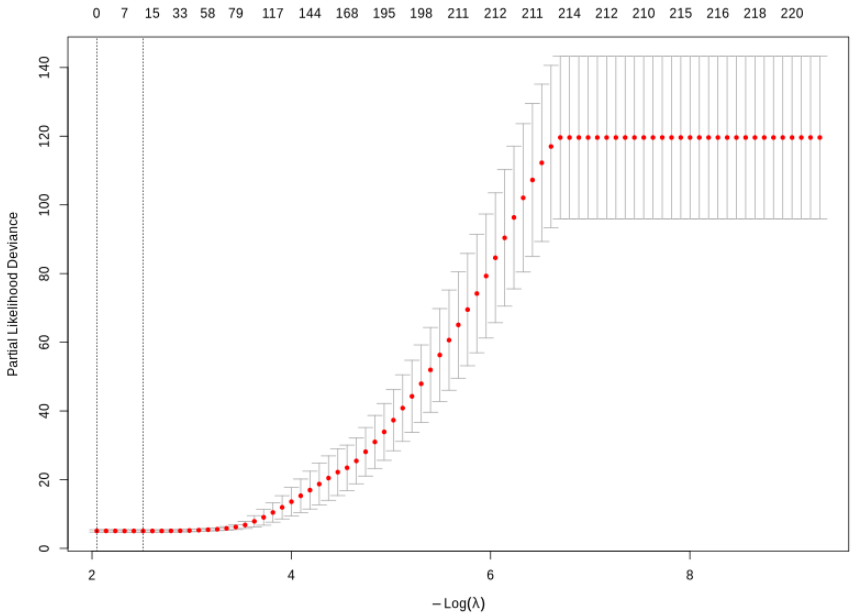


Figure 4. ***Cross-validation curve for LASSO-penalised Cox regression in TCGA–BLCA.*** *The lambda.min solution defines the multivariate regulatory predictor set (12 ENCODE cCRE loci) used for penalised refitting and inference.*

3.2.3 Multivariate survival association (Firth Cox refit) and prioritised regulatory loci

The LASSO-selected regulatory predictors were refit using multivariate Firth penalised Cox regression to obtain stable effect estimates in a sparse carrier regime and to mitigate separation artefacts that can arise in rare-variant survival modelling. In BLCA, the multivariate model comprised 12 ENCODE cCRE loci (9 promoter-like and 3 enhancer-like) together with age at diagnosis, and showed strong overall evidence of association with overall survival (global penalised likelihood ratio test reported in Table 2). Within this joint model, four loci remained individually significant, including one strong risk-associated promoter locus and three protective loci. Specifically, the promoter-associated variant chr17:38428384 C>G exhibited a large adverse effect (HR = 9.36, 95% CI 2.578–23.903, p = 0.00238; 3/326 carriers). In contrast, three loci were associated with reduced hazard: the promoter variant chr22:30694816 C>G (HR = 0.087, 95% CI 0.001–0.615, p = 0.00681; 3/326 carriers), the enhancer variant chr16:72996641 G>A (HR = 0.101, 95% CI 0.001–0.695, p = 0.0117; 5/326 carriers), and the promoter variant chr11:9758261 C>T (HR = 0.105, 95% CI 0.001–0.758, p = 0.0184; 4/326 carriers). Age at diagnosis remained independently associated with outcome (HR = 1.035, 95% CI 1.017–1.054, $p = 9.32\times10^{-5}$), indicating that the regulatory signals persist beyond baseline age effects.

The fitted coefficient profile across all retained predictors (Figure 5) provides a compact view of the directionality and relative contribution of each locus under penalised inference. Notably, although several additional loci were retained by LASSO, their individual multivariate p-values were non-significant (Table 2), consistent with (i) low carrier counts, (ii) partial redundancy among rare

predictors, and (iii) LASSO selection optimised for joint predictive structure rather than per-feature hypothesis testing. Overall, the BLCA results prioritise a small set of ENCODE promoter/enhancer–annotated regulatory VUS candidates with measurable prognostic association while highlighting the need for replication due to the very low carrier frequencies (3–6 carriers per locus).

Table 2. ***Multivariate Firth Cox regression results for TCGA–BLCA predictors retained after LASSO selection.*** *Hazard ratios (HR), 95% confidence intervals, p-values, and carrier counts are reported for 12 ENCODE cCRE loci (promoter/enhancer) and age at diagnosis.*

Predictor	Regulatory category	Region	HR (Firth)	95% CI	p-value (multi)	Carriers
chr17:38428384 C>G	ENCODE cCRE (PLS/ELS)	Promoter	9.36	2.578 - 23.903	0.00238	3/326
chr22:30694816 C>G	ENCODE cCRE (PLS/ELS)	Promoter	0.087	0.001 - 0.615	0.00681	3/326
chr16:72996641 G>A	ENCODE cCRE (PLS/ELS)	Enhancer	0.101	0.001 - 0.695	0.0117	5/326
chr11:9758261 C>T	ENCODE cCRE (PLS/ELS)	Promoter	0.105	0.001 - 0.758	0.0184	4/326
chr2:86440987 C>T	ENCODE cCRE (PLS/ELS)	Promoter	0.219	0.002 - 1.554	0.164	3/326
chr17:39197623 C>T	ENCODE cCRE (PLS/ELS)	Promoter	0.214	0.002 - 1.584	0.165	4/326
chr15:57591919 G>A	ENCODE cCRE (PLS/ELS)	Promoter	0.177	0 - 1.682	0.193	4/326
chr19:37980078 C>T	ENCODE cCRE (PLS/ELS)	Enhancer	0.295	0.002 - 1.875	0.262	3/326
chr2:28280674 C>T	ENCODE cCRE (PLS/ELS)	Enhancer	0.330	0.003 - 2.064	0.318	4/326
chr8:66712697 G>A	ENCODE cCRE (PLS/ELS)	Promoter	0.336	0.003 - 2.133	0.334	6/326
chr8:81112111 G>C	ENCODE cCRE (PLS/ELS)	Promoter	0.385	0.003 - 2.632	0.427	3/326
chr5:177600153 G>A	ENCODE cCRE (PLS/ELS)	Promoter	1.333	0.005 - 40986.618	0.928	3/326
age_at_index	—	—	1.035	1.017 - 1.054	9.32e-05	—

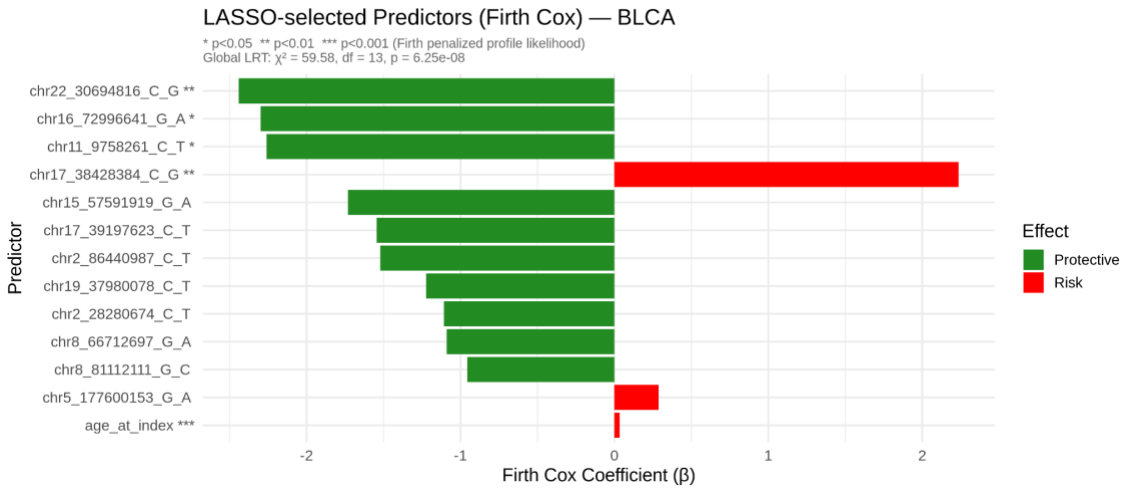


Figure 5. ***LASSO-selected predictors (Firth Cox) in TCGA–BLCA.*** *Barplot of Firth Cox coefficients for the 12 ENCODE cCRE loci retained after LASSO selection and for age at diagnosis. Positive coefficients (red) indicate increased hazard (risk), whereas negative coefficients (green) indicate decreased hazard (protective); significance markers (*, **, ***) correspond to the multivariate p-values reported in Table 2, and the global penalized likelihood ratio test summarizes overall model significance.*

3.2.4 Carrier-frequency structure and interpretability constraints

As in BRCA, the BLCA regulatory-variant feature space is dominated by rare loci, with most candidate ENCODE cCRE variants present in only a small number of samples. This is reflected both in the multivariate model inputs—where carrier counts per locus are typically 3–6, with some loci at the minimum-support threshold—and in the overall frequency landscape, which shows a strongly left-skewed distribution with the majority of variants occurring at very low cohort frequencies (Figure 6). This sparsity is expected after restricting somatic callsets to promoter/enhancer regulatory intervals and applying minimum-support filtering, and it has direct modelling implications: rare predictors can induce instability or quasi-complete separation under standard Cox regression, motivating penalised approaches (LASSO for selection and Firth Cox for refitting) to obtain finite and robust inference.

Within this interpretability boundary, the directionality and relative magnitude of fitted effects remain informative for prioritisation. The BLCA model contains one prominent risk-associated promoter locus (chr17:38428384 C>G; Figure 5) contrasted with multiple loci showing protective directionality, including three that remain individually significant in the multivariate Firth refit (Table 2). The combination of higher event rate in BLCA and penalised modelling enables detection of a sparse but multivariate regulatory signature; nevertheless, the low carrier frequencies indicated in Figure 6 and Table 2 mean that individual hazard-ratio estimates should be interpreted primarily as cohort-level prioritisation signals rather than stable population effect sizes, and therefore warrant replication and orthogonal functional support.

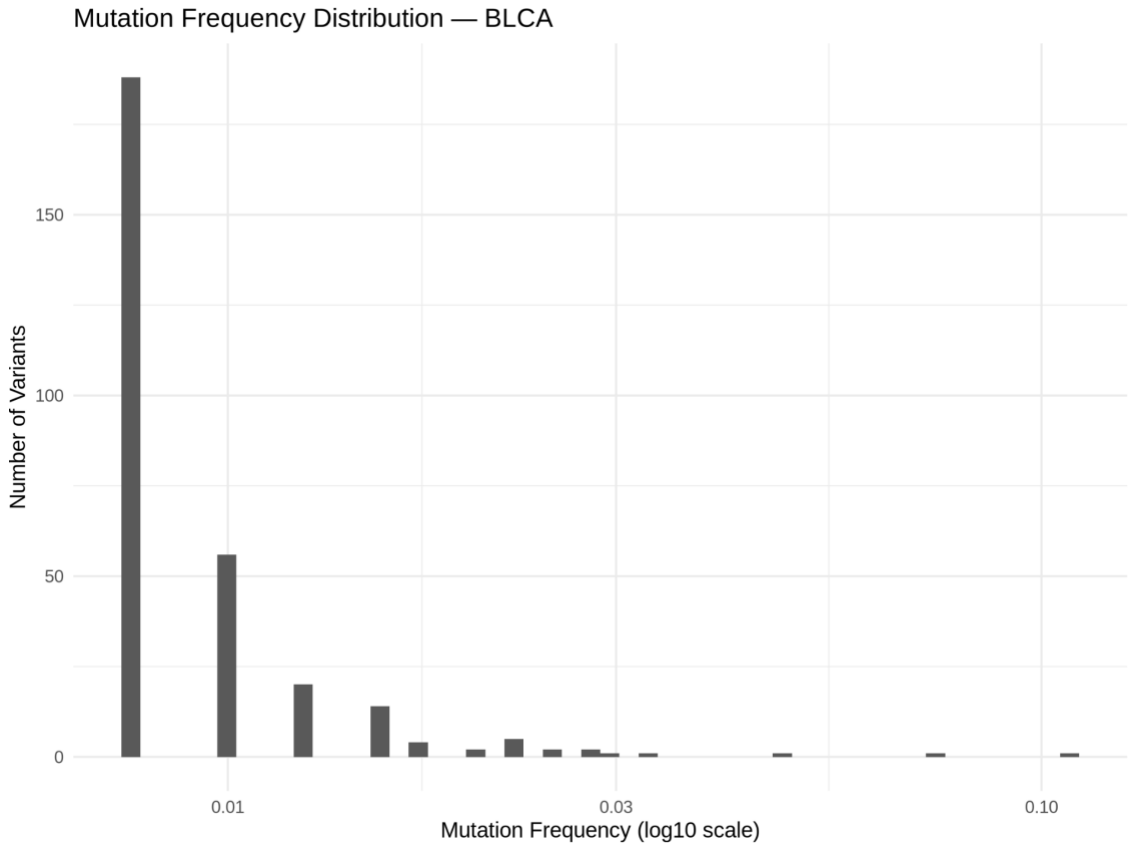


Figure 6. ***Mutation frequency distribution of candidate regulatory variants in TCGA–BLCA.*** *Histogram of per-variant cohort frequencies (x-axis on log10 scale) showing a strongly skewed distribution dominated by low-frequency loci, consistent with a rare-feature modelling regime and motivating penalised survival inference.*

3.3 Cross-referencing with population resources (UK Biobank / GWAS Catalog)

To provide population-scale context for the survival-prioritized regulatory loci, all variant positions retained in the final selected predictor sets (1 BRCA locus and 12 BLCA loci; total n = 13) were cross-referenced against a UK Biobank subset of the NHGRI–EBI GWAS Catalog. No overlaps were identified. This is consistent with the low carrier frequencies of the prioritized loci in the TCGA cohorts and the fact that GWAS catalog resources primarily capture common germline associations detectable in large population studies, whereas the present workflow targets rare, cohort-specific regulatory signals in cancer datasets.

4 Discussion

This deliverable operationalizes the interpretation of intergenic regulatory VUSs by anchoring somatic variation to standardized promoter/enhancer regulatory maps (ENCODE cCREs) and quantifying prognostic relevance through outcome-oriented survival modelling. Across TCGA BRCA and BLCA, the analyses consistently indicate that a very small subset of regulatory loci can be prioritized under a rare-variant regime when penalized methods are used to control instability and separation. In BRCA (n = 888; 95 deaths), the framework converged to an extremely sparse model comprising age plus a single promoter-associated locus (chr1:210,233,993 TT→AA), with a strong adverse association with overall survival (HR = 10.60; 4 carriers). In BLCA (n = 326; 143 deaths), the higher event rate enabled selection of a multivariate regulatory signature (12 ENCODE cCRE loci plus age), with one risk-associated promoter locus and three protective loci remaining individually significant after multivariate Firth refitting, alongside a highly significant global model fit.

A key cross-cohort observation is that all retained survival-associated predictors map to ENCODE cCRE promoter/enhancer annotations, while no JASPAR TFBS-derived features were retained in the final selected predictor sets. This does not imply that TF binding perturbations are biologically irrelevant, but rather that—in the current cohort sizes, carrier frequencies, and modelling regime—the most stable, outcome-informative signals are captured at the level of regulatory element intervals (promoter/enhancer) rather than motif-centric abstractions. This aligns with the practical reality of rare-feature survival modelling: motif-level features can be highly redundant and correlated (many TFBS overlapping the same regulatory locus), which may reduce interpretability and stability under penalized selection when events are limited.

Interpretation must be framed by the carrier-frequency structure of regulatory VUSs. In both BRCA and BLCA, candidate regulatory predictors are dominated by rare loci, and the prioritized variants are typically present in only 3–6 individuals (BRCA prioritized locus: 4 carriers; BLCA significant loci: 3–5 carriers). Even with Firth penalization stabilizing coefficients and confidence intervals, effect sizes in such settings should be treated as prioritization signals rather than definitive biomarker claims. The strongest immediate value of this deliverable is therefore the creation of a ranked, cohort-resolved candidate list of regulatory loci with measurable association to survival, which can be escalated to orthogonal validation (independent cohorts, functional evidence layers, and mechanistic follow-up).

The results also highlight how cohort-level characteristics influence detectability. The low BRCA event rate (10.7%) strongly constrains power for rare-variant prognostic discovery and naturally drives selection toward extremely sparse solutions, whereas BLCA (43.9% events) supports a richer multivariate signature under the same modelling strategy. Consequently, differences in selected feature-set size between cohorts should be interpreted primarily as reflecting statistical regime (events and effective sample size), rather than necessarily indicating fundamentally different degrees of regulatory involvement in disease biology.

Finally, optional population cross-referencing did not identify overlaps between the selected TCGA loci and a UK Biobank subset of the GWAS Catalog, consistent with the ultra-rare carrier frequencies observed here and the focus of GWAS resources on common variation. This negative result is itself informative: the prioritized loci are plausible cohort-specific rare regulatory candidates, motivating validation strategies that do not rely on population GWAS replication (e.g., functional regulatory activity context, expression or chromatin correlates, and targeted sequencing in matched cohorts).

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

5 References

[1] Moore, J. E., Pratt, H. E., Fan, K., Phalke, N., Fisher, J., Elhajjajy, S. I., Andrews, G., Gao, M., Shedd, N., Fu, Y., Lacadie, M. C., Meza, J., Khandpekar, M., Ganna, M., Choudhury, E., Swofford, R., Phan, H., Ramirez, C. C., Campbell, M., … Weng, Z. (2026). An expanded registry of candidate cis-regulatory elements. Nature. https://doi.org/10.1038/s41586-025-09909-9

[2] Ovek Baydar, D., Rauluseviciute, I., Aronsen, D. R., Blanc-Mathieu, R., Bonthuis, I., de Beukelaer, H., Ferenc, K., Jegou, A., Kumar, V., Lemma, R. B., Lucas, J., Pochon, M., Yun, C. M., Ramalingam, V., Deshpande, S. S., Patel, A., Marinov, G. K., Wang, A. T., Aguirre, A., … Mathelier, A. (2025). JASPAR 2026: expansion of transcription factor binding profiles and integration of deep learning models. Nucleic Acids Research, 54(D1), D184–D193. https://doi.org/10.1093/nar/gkaf1209

[3] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), 1113–1120. https://doi.org/10.1038/ng.2764

[4] Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., … Harris, L. W. (2022). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Research, 51(D1), D977–D985. https://doi.org/10.1093/nar/gkac1010

Machine Learning Algorithm for Predicting Translation Initiation Start Positions

Stefanos Digenis[1,2] , Dimitris Grigoriadis[1,2], Marios Miliotis[1,2], Artemis G. Hatzigeorgiou[1,2]

1 DIANA-Lab, Department of Computer Science and Biomedical Informatics, Univ of Thessaly, Lamia, Greece,
2 Hellenic Pasteur Institute, Athens 11521, Greece

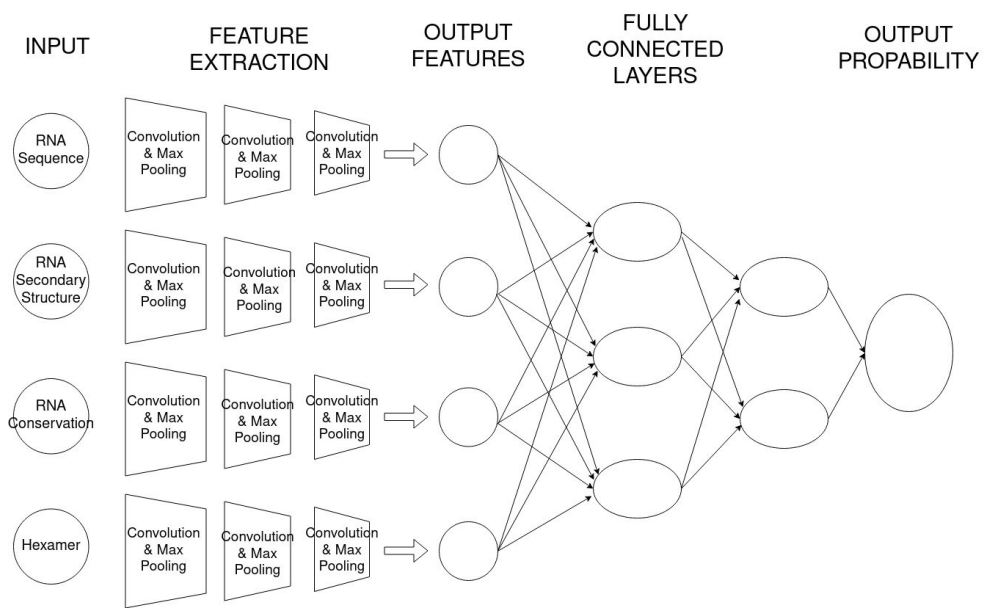
Abstract

Translation Initiation Start (TIS) sites are gene regions where protein synthesis begins. This study tackles RNA sequence classification using convolutional neural networks (CNNs) to extract genomic sequence features. Accurate prediction of these sites is key for understanding translation mechanisms. A deep learning model was developed to predict TIS presence in RNA sequences, leveraging CNNs to identify important patterns. Trained on a dataset of annotated TIS, the model achieved 93.75% accuracy by incorporating sequence, conservation, and hexamer features. This research supports applications in gene annotation, drug target identification, and protein synthesis mechanisms, benefiting both computational and experimental biology.

Dataset

The positive dataset in this study consisted of verified RNA sequences from the SwissProt database, focusing on well-annotated start codon positions. Data preprocessing merged overlapping transcripts and added 12Kbp sequences around each TIS site, ensuring comprehensive annotation. A negative set of non-functional AUG regions was also included to improve the model's training and avoid duplicate information.

Architecture



Methodology

Conservation SEQUENCE

15 nucleotides	A	T	G	5 nucleotides

Conservation

15 nucleotide cons score	0.35	0.21	0.56	5 nucleotide cons score

RAW SEQUENCE

15 nucleotides	A	T	G	5 nucleotides

One Hot Encoding

15 nucleotides encoded with [1, 0, 1, 0]	[1, 0, 0, 0]	[0, 0, 0, 1]	[0, 0, 1, 0]	5 nucleotides encoded [1, 0, 1, 0]

HEXAMER MATRIX

148 nucleotides	A	T	G	148 nucleotides

Hexame Matrix

AAAAAA	3
AAAAAT	1
...	
CCCCCA	0

Normalization

AAAAAA	0.72
AAAAAT	0.24
...	
CCCCCA	0.0

Contrafold SEQUENCE

150 nucleotides	A	T	G	150 nucleotides

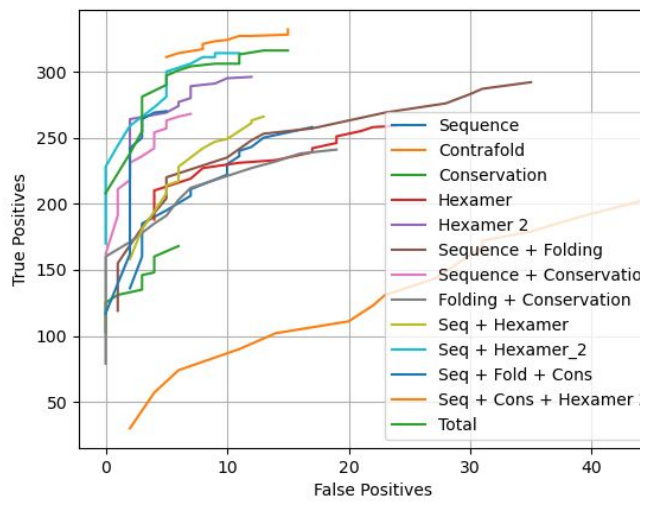
Contrafold

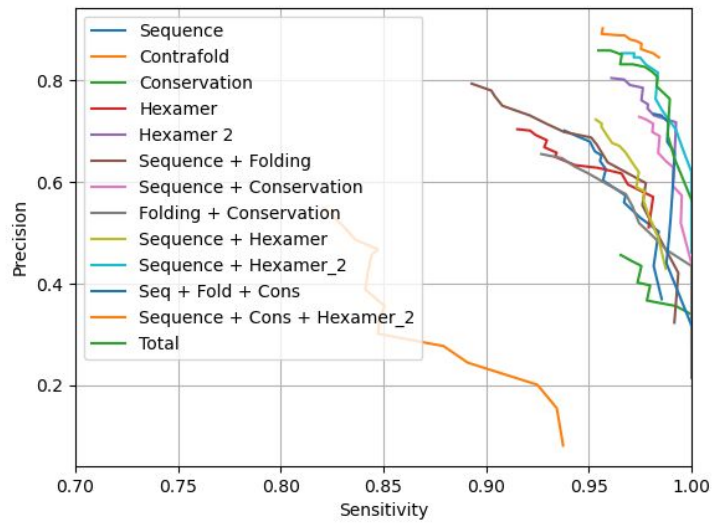
150 characters of ('(', '.', ')')	')'	'.'	'('	150 characters of ('(', '.', ')')

One Hot Encoding

150 characters encoded with [1, 0, 1]	[0, 0, 1]	[0, 1, 0]	[1, 0, 0]	150 characters encoded with [1, 0, 1]

Results





Comparison

	Accuracy	Precision	Recall
Best Model combination	0.9375	0.9497	0.9239
TITER	0.7282	0.7900	0.6168
TIS-Predictor	0.7900	0.7700	0.84

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ

Funded by the European Union - NextGenerationEU through Greece 2.0—National Recovery and Resilience Plan, under the call ”Flagship actions in interdisciplinary scientific fields with a special focus on the productive fabric” (ID 16618), project name ”Bridging big omic, genetic and medical data for

21st IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology

Certificate of Presentation

This certifies that the full paper entitled **A Machine Learning Algorithm for Predicting Translation Initiation Start Positions**, authored by **Stefanos Digenis, Dimitris Grigoriadis, Marios Miliotis and Artemis Hatzigeorgiou** was presented by **Stefanos Digenis** during the **21st IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology** held in Natal – Brazil, August 27-29, 2024.

Renan Cipriano Moioli

General chair – Renan C. Moioli
Federal University of Rio Grande do Norte, Brazil
Organizing committee of IEEE CIBCB 2024

HOSTED BY

SPONSORED AND SUPPORTED BY

CORPORATE SPONSORSHIP