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

DELIVERABLE WP9.2

Technical report on deep-learning model results and findings from 30 WGS samples from patients in Greek cohorts

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

1 Introduction

Intergenic non-coding variation remains a major interpretability bottleneck for precision oncology, because many variants fall outside protein-coding sequence yet can perturb gene regulation through effects on promoters, enhancers, and transcription factor binding. Regulatory elements operate in a context-dependent manner, integrating transcription factor occupancy, chromatin state, and three-dimensional genome organization to control transcription. Consequently, somatic variants embedded in promoter- and enhancer-like regions can rewire transcriptional programs without producing obvious amino-acid changes, complicating prioritization and clinical interpretation.

To move from generic genome annotations to biologically constrained regulatory hypotheses, this work constructs a layered framework that combines (i) ENCODE SCREEN candidate cis-regulatory elements (cCREs) [1] to define promoter-like and enhancer-like regulatory compartments, (ii) tissue-relevant chromatin interaction evidence to link distal regulatory elements to putative target genes, and (iii) orthogonal functional evidence layers to support mechanistic interpretation. The analysis is performed under a harmonized workflow for two TCGA cohorts—breast invasive carcinoma (BRCA) and bladder urothelial carcinoma (BLCA)—to enable direct comparison of regulatory architectures across tumour types. Regulatory variant burden is quantified both as absolute overlap counts and as length-normalized densities (variants per kilobase), ensuring that enrichment patterns are not driven by annotation footprint. In addition, functional support is evaluated via intersection with the ENCODE eQTL catalogue and via transcription factor binding site (TFBS) context to characterize potential disruption of regulatory grammar. Finally, outcome-oriented prioritization is performed using overall survival data through carrier-versus-noncarrier association testing.

Beyond the TCGA-focused regulatory landscape, this deliverable also reports findings from whole-genome sequencing of patients from Greek cohorts, where in-depth analysis aims to resolve unresolved hereditary cancer and rare disease cases that remain unsolved after routine diagnostic testing. Taken together, the report provides an evidence-weighted map of intergenic regulatory candidates and complementary real-world WGS findings, supporting both mechanistic hypothesis generation and downstream validation planning.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ
ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ

Με τη χρηματοδότηση
της Ευρωπαϊκής Ένωσης
NextGenerationEU

2 Materials and Methods

2.1 Data resources and regulatory annotation framework

Regulatory elements were obtained from ENCODE SCREEN candidate cis-regulatory elements (cCREs) on GRCh38. Analyses focused on promoter-like signatures (PLS) and enhancer-like signatures (pELS and dELS). In parallel, tissue-relevant chromatin interaction evidence was used to assign distal elements to putative target genes. ENCODE 3D chromatin interaction datasets (Hi-C and ChIA-PET) were aggregated and then filtered to breast-relevant sources (adult female breast epithelium, MCF_10A, and MCF-7), retaining RNAPII ChIA-PET, intact Hi-C, and CTCF ChIA-PET evidence. These interactions were converted into promoter–gene and enhancer–gene association layers to support downstream mapping of intergenic variants to candidate target genes.

2.2 Variant datasets, harmonization, and intersection strategy

Somatic variant datasets were analysed separately for TCGA BRCA and TCGA BLCA [2] under a consistent coordinate system (GRCh38) and harmonized preprocessing criteria to enable cross-cancer comparability. Variants were intersected with ENCODE cCRE annotations to classify overlaps into promoter (PLS) and enhancer (pELS/dELS) regulatory contexts, and were subsequently propagated to gene-level hypotheses via the tissue-filtered 3D interaction links. All overlaps and joins were performed using standardized keys and interval-based operations, with outputs exported as cohort-resolved tables summarizing variant counts, regulatory class membership, and mapped target genes.

2.3 Functional evidence layers and bias-aware quantification

To move beyond positional overlap, regulatory variants were evaluated under complementary functional layers. First, variants were intersected with the ENCODE eQTL catalogue using allele-matched identifiers (chr_pos_ref_alt), retaining tissue/source metadata and association counts per regulatory class and gene. Second, transcription factor binding site (TFBS) context was evaluated using JASPAR-supported TFBS annotations overlapping regulatory elements, enabling quantification of TFBS localization (promoter vs enhancer) and enumeration of TFBS-disrupting variants per TF and target gene. Because enhancers cover substantially larger genomic footprints than promoters, overlap results were interpreted both as absolute counts and as length-normalized densities (variants per kilobase) computed separately for PLS, pELS, and dELS, ensuring that enrichment patterns reflect biology rather than annotation size effects.

2.4 Population cross-referencing and clinical association testing

Where applicable, regulatory variants were cross-referenced against UK Biobank-derived [3, 4] sets to identify shared loci and provide population-scale context, noting that somatic and ultra-rare cohort-specific variants are not expected to systematically replicate in GWAS-driven resources. Finally, clinical relevance was assessed using overall survival (OS) data from TCGA clinical and follow-up tables.

Variants were tested under a carrier-versus-noncarrier stratification scheme, applying minimum carrier thresholds to avoid unstable estimates. Associations with OS were evaluated using Cox proportional hazards models (with stage adjustment when available) and visualized with Kaplan–Meier curves. Multiple testing was controlled using Benjamini–Hochberg false discovery rate correction, and significant loci were reported as prioritized regulatory candidates for downstream validation.

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

3 Results

3.1 Regulatory variant landscape and recurrently targeted loci (BRCA vs BLCA)

Intersection of TCGA somatic variants with ENCODE cCRE regulatory annotations revealed extensive regulatory mutational burden in both cancers, with clear differences in how variants distribute across promoters and enhancers.

In BRCA, 31,944 unique variants mapped to promoter-like regions (PLS), while 367,402 unique variants mapped to enhancer-like regions (ELS; proximal and distal). The resulting mapping space was enriched for protein-coding targets and lncRNAs, with promoter intersections involving protein-coding genes in 67.5% of mapping events, and enhancers showing similarly high protein-coding prioritization (over 900,000 mapping events). Recurrently targeted loci included the non-coding Y_RNA locus, histone gene clusters (e.g., H2BC8, H2AC8), and strong enhancer-linked hotspots in the MYC/CASC11 regulatory neighborhood.

In BLCA, 46,470 unique variants overlapped promoters and 23,532 overlapped enhancers, with promoter mapping producing a substantially larger number of mapping rows than enhancer mapping. As in BRCA, mapping events preferentially involved protein-coding genes (71.2% of promoter interactions; 62.7% of enhancer interactions), with broad representation of lncRNAs. BLCA promoter hotspots prominently included histone clusters (e.g., H2BC4**,** H2AC6) and the metabolic transporter SLC3A2, while enhancer-linked clusters included loci associated with CILK1, MYH9, and IRF2BP2, alongside signal at TOB1/TOB1-AS1.

3.2 Bias-aware enrichment: promoters show the highest mutational density after normalization

Because enhancer annotations span a much larger genomic footprint than promoters, absolute overlap counts were complemented with length-normalized mutational densities (variants per kilobase). In BRCA, distal enhancers dominated absolute overlaps, but after normalization, promoters exhibited the highest mutational density (4.27 variants/kb)**,** exceeding distal enhancers (2.97 variants/kb) and proximal enhancers (3.17 variants/kb).

n BLCA, promoters exceeded distal enhancers even in absolute counts, and normalization reinforced this trend: promoters reached 8.57 variants/kb, surpassing proximal enhancers (6.92 variants/kb) and distal enhancers (6.52 variants/kb). Collectively, these results indicate that promoter-proximal regulatory space is a consistent enrichment compartment for somatic variants across both cancers, with a particularly strong signal in BLCA.

3.3 Functional support via eQTL intersection reveals divergent regulatory architectures

Intersection of regulatory variants with the ENCODE eQTL catalogue provided functional evidence consistent with distinct regulatory mechanisms across cancers. In BRCA, eQTL support was strongly enhancer-dominated: among 9,696

eQTL associations, 9,350 (96.4%) localized to enhancers versus 346 (3.6%) to promoters. The enhancer associations arose from 202 unique variants targeting 501 unique genes, consistent with distributed, long-range enhancer-driven regulation.

In contrast, BLCA showed promoter centrality: among 10,960 eQTL associations, 10,381 (94.7%) localized to promoters, while 579 (5.3%) fell in enhancers. Notably, the promoter associations were generated by only 10 unique variants, implying very high association density per variant and suggesting strong promoter-proximal transcriptional effects in this cohort context.

3.4 TFBS disruption patterns differ sharply (enhancer-centric BRCA vs promoter-centric BLCA)

TFBS overlaps within regulatory elements revealed striking context differences. In BRCA, 190,805 TFBS overlapped regulatory elements, with 78.7% in enhancers and 21.3% in promoters. Intersection with somatic variants identified 4,606 unique TFBS-disrupting variants involving 373 TFs and 521 target genes predominantly enhancer-associated.

In BLCA, 393,441 TFBS overlapped regulatory elements, with 96.1% in promoters and only 3.9% in enhancers. Intersection with somatic variants identified 2,228 TFBS-disrupting variants involving 281 TFs and 479 target genes, overwhelmingly promoter-associated, consistent with direct disruption of transcriptional initiation control.

3.5 Clinical impact: survival-associated regulatory loci in BRCA and BLCA

Linking regulatory variants to overall survival identified a small number of high-confidence loci with significant associations after multiple-testing correction. In BRCA, regulatory variants in ENCODE PLS/ELS regions yielded multiple significant hits, including extremely strong effects for selected loci (e.g., chr1_174940449_A_T reported with HR 55.90, FDR 3.29×10^{-5}). Additional significant candidates included promoter-associated chr4_2934745_G_C (HR 61.01, FDR 1.54×10^{-6}) and chr1_210233993_TT_AA (4 carriers; HR 9.46; FDR 3.32×10^{-3}).

In BLCA, a promoter locus chr17_38428384_C_G was identified as a significant predictor of mortality (3 carriers; HR 10.82; FDR 0.0135), with carrier survival curves showing an early sharp decline relative to wild-type.

3.6 Population cross-reference: UK Biobank overlap and splice-impact context

Cross-referencing with a UK Biobank-derived set identified **72 shared variants** for BRCA and **62 shared variants** for BLCA. In BRCA, three shared variants localized within enhancer regions (chr11:117217857 G>A; chr13:113838644 C>G; chr17:59784140 T>A). Importantly, none of the BRCA–UKB shared variants overlapped regions predicted to impact splicing by SpliceAI (delta score > 0.2), supporting a predominantly regulatory interpretation for these shared signals. In BLCA, one shared variant (chr3:12288912 G>A) localized within a promoter region.

4 Findings from 30 WGS samples (VUS) from patients in Greek cohorts

4.1 Hereditary cancer

4.1.1 Objective

Hereditary cancer accounts for approximately 5–10% of all malignancies and is caused by germline pathogenic variants in cancer-predisposition genes. The hereditary cancer spectrum also includes many rare inherited syndromes that predispose to the development of various malignancies, such as familial adenomatous polyposis and Lynch syndrome. Genetic analysis in hereditary cancer-predisposition syndromes is essential for appropriate clinical management, guidance and counselling of patients and their relatives, while simultaneously contributing to the understanding of the molecular mechanisms of carcinogenesis.

Multi-gene panel testing using next-generation sequencing (NGS) is now the main diagnostic tool for hereditary cancer syndromes. With this approach, multiple genes can be tested simultaneously in substantially less time and at lower cost compared with conventional sequencing. In addition, a complementary method—multiplex ligation-dependent probe amplification (MLPA)—is typically used to detect large genomic rearrangements in the genes under study. Nevertheless, routine diagnostic sequencing generally focuses on coding regions (exons) and/or exon–intron boundaries (splice regions), an approach that fails to identify the underlying genetic cause in a substantial proportion of patients.

The aim of the present work is the in-depth genetic investigation of oncology patients and families with strong suspicion of genetic predisposition who, despite diagnostic-level testing, remain without a clear genetic diagnosis. Whole-genome sequencing (WGS) enables the study of genomic regions within already known genes that escape standard diagnostic approaches, as well as the identification of new genes that may be associated with cancer predisposition.

4.1.2 Study cohort

The present cohort includes 15 individuals from 10 families with strong suspicion of a hereditary cancer syndrome (Table 1). Specifically, these are 10 patients who were assessed at the Laboratory of Human Molecular Genetics of NCSR “Demokritos”, in whom extensive diagnostic testing (sequencing, MLPA, and bioinformatic analysis for genomic rearrangements) did not reveal a pathogenic finding. In addition, in three families it was possible to collect samples from relatives of the index cases to perform duo/trio analysis.

In more detail, from the family of index patient 54VCA, two siblings were also included, both diagnosed with renal malignancy. From the family of index patient 2584CRC, the father (also diagnosed with colorectal cancer) was included. From

the family of index patient 2712CRC, both parents were included to investigate a possible recessive inheritance pattern. All participants were enrolled after detailed information and written consent, and all data were coded to ensure anonymity. Indicatively, pedigrees for patients 2753CRC and 2667CRC are shown in Figure 1.

Table 1. Patient cohort of the project

No.	Sample ID	Personal history (age at diagnosis)	Notes
1	54VCA	Kidney cancer (52); Pancreatic cancer (52); Pheochromocytoma (52)	Trio analysis
2	54α VCA	Lung cancer (57); Kidney cancer (60)	
3	54γ VCA	Kidney cancer (47)	
4	297 VCA	Brain cancer (11); Kidney cancer (25, 29)	
5	2753 CRC	Polyposis (54)	
6	2667 CRC	Colorectal cancer (59); Colorectal polyps (62)	
7	2712 CRC	Polyposis (27); Colorectal cancer (27)	Trio analysis
8	2712α CRC	Asymptomatic	
9	2712β CRC	Asymptomatic	
10	2571 CRC	Colorectal cancer (43); Ovarian cancer (45)	
11	2565 CRC	Colorectal cancer (22)	
12	2584 CRC	Colorectal cancer (31, 35)	Duo analysis
13	2584α CRC	Colorectal cancer (64)	
14	2663 CRC	Colorectal cancer (37, 66)	
15	333 GC	Gastric cancer (65)	

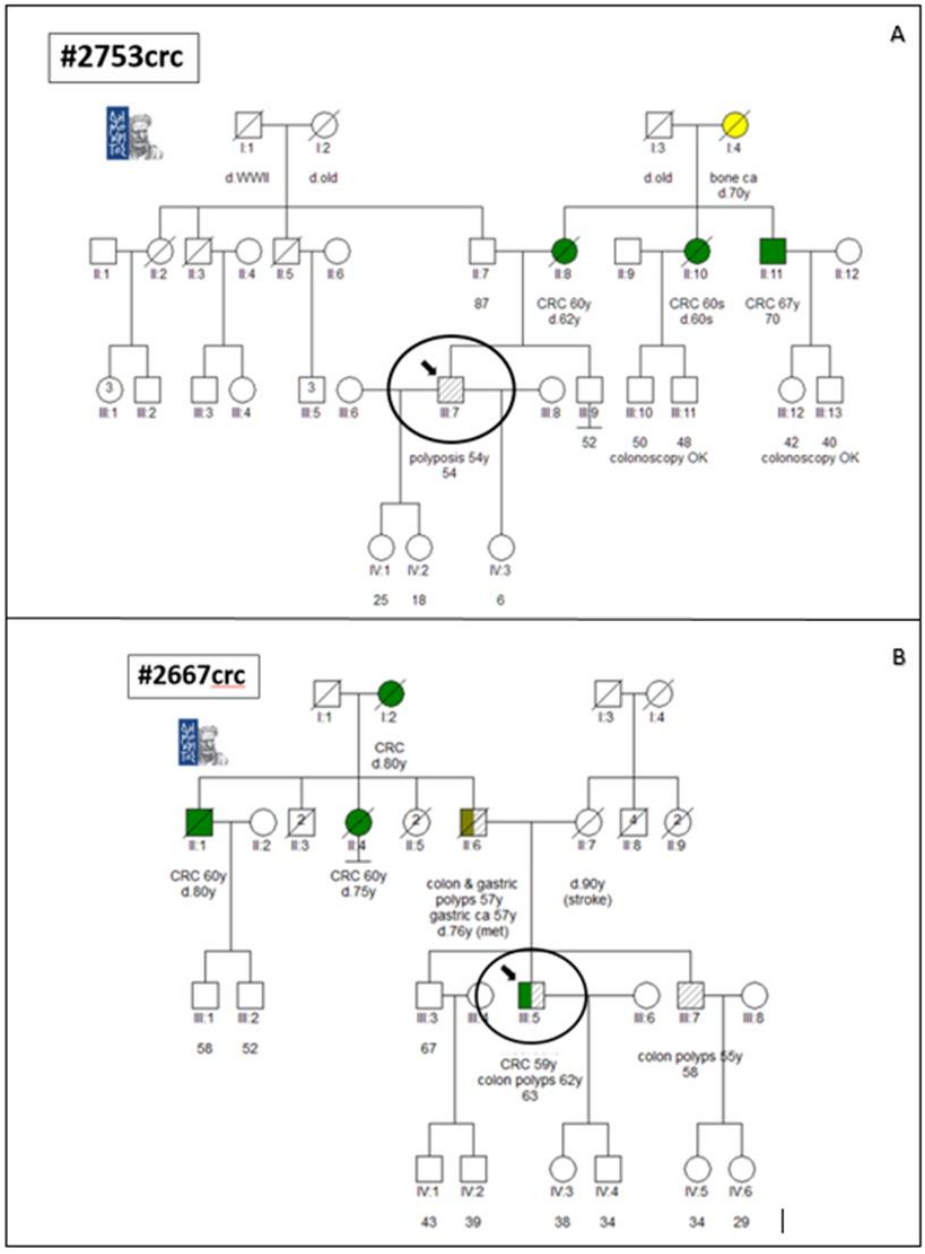

Figure 1. Pedigrees of families with strong suspicion of a hereditary cancer syndrome: (A) index patient 2753CRC and (B) index patient 2667CRC. [Arrow indicates the index patient; ca=cancer; CRC=colorectal cancer; d.=died; y=years.]

4.1.3 Sequencing methodology (WGS)

We constructed a total of 15 WGS libraries using either the MGIEasy FS PCR Free DNA Library Prep Set, v1.2 (MGI Tech Co., Ltd.), according to the MGIEasy FS PCR Free DNA Library Prep Set user manual, or the MGIEasy Fast PCR-FREE FS Library Prep Set, v2.0, according to the MGIEasy Fast PCR-FREE FS Library Prep Set user manual (Version 1.0). In brief, 900–1000 ng DNA was used for WGS library preparation; fragmentation time and size-selection strategy were determined based on the input DNA amount and according to the manufacturer's instructions. Library DNA concentration was assessed using the Invitrogen Qubit 4 fluorometer and the Invitrogen Qubit dsDNA HS Assay kit.

WGS libraries were batched in pools of 5 samples and combined equimolarly to a total of 300 ng. Denaturation of pooled DNA, conversion to single-stranded circles, enzymatic digestion, cleanup of digestion products, and QC steps were performed according to the MGIEasy Fast PCR-FREE FS Library Prep Set user manual (Version 1.0). Finally, 70 fmol of each pooled ssCirDNA was used for DNB preparation and sequencing on a DNBSEQ-G400 platform at the Genomics Unit of BSRC Alexander Fleming, using the G400 FCL PE150 high-throughput sequencing set (MGI Tech Co., Ltd.), according to the DNBSEQ-G400RS High-throughput (Rapid) Sequencing Set user manual (Version 8.0).

4.1.4 Quality control and mapping

For each sample, initial sequencing quality control (QC) was performed with FastQC. Low-quality reads were trimmed using TrimGalore. Only reads with length ≥50 bp that also retained correct read pairing were included in the final analysis.

Reads were mapped to the reference genome (hg38) using BWA (bwa-mem) with default parameters (per lane). Pre-processing for genomic variation analysis was performed with samtools (name sorting, fixmate, coordinate sorting, and markdup for duplicate marking prior to variant calling).

After mapping, genome coverage was quantified using Picard and a second QC was performed according to procedures available at https://github.com/moulos-lab/genomics-facility-processes.

4.1.5 Variant calling analysis

Variant calling followed an independent per-sample pipeline: each sample was processed separately and a gVCF file was generated per sample. Joint calling was not applied at this stage.

BAM validation and preparation

First, BAM validity was checked using the GATK ValidateSamFile tool to ensure SAM/BAM compliance, consistent coordinate information, flags and mate information, and that contig names correspond to the reference genome.

```
gatk ValidateSamFile \
  -I x.bam \
  -R Homo_sapiens_assembly38.fasta \
  -MODE SUMMARY
```

Base Quality Score Recalibration (BQSR)

Base Quality Score Recalibration (BQSR) was then applied to correct systematic biases (e.g., lower quality at read ends or in specific base contexts). Known variants from reliable resources (dbSNP and Mills indels) were used.

```
gatk BaseRecalibrator \
  -R Homo_sapiens_assembly38.fasta \
  -I x.bam \
```

```
  --known-sites Homo_sapiens_assembly38.dbsnp138.vcf \
  --known-sites Mills_and_1000G_gold_standard.indels.hg38.vcf.gz \
  -O x_recal_data.table
```

The resulting x_recal_data.table file contains the recalibration information, which was applied to the BAM file:

```
gatk ApplyBQSR \
  -R Homo_sapiens_assembly38.fasta \
  -I x.bam \
  --bqsr-recal-file x_recal_data.table \
  -O x_recal.bam
```

The result is a new BAM file with recalibrated base quality scores.

Header and read group fixes

Next, the BAM header was inspected (contigs @SQ, read groups @RG, and program records @PG). It was observed that sample information (SM tag) was missing, which is required for accurate variant calling.

```
samtools view -H x.bam > header.txt
sed '/^@RG/ s/$/\tSM:x/' header.txt > header_modified.txt
samtools reheader header_modified.txt x_recal.bam > x_reheadered.bam
```

Read group information was then added/replaced to ensure each read is correctly linked to its originating sample:

```
gatk AddOrReplaceReadGroups \
  -I x_reheadered.bam \
  -O x_fixed.bam \
  --RGLB WGS \
  --RGPL MGI \
  --RGPU unit1 \
  --RGSM x
```

Finally, an index was created for the final BAM file:

```
samtools index x_fixed.bam
```

Variant calling

Variant calling was performed with HaplotypeCaller using the reference genome Homo_sapiens_assembly38.fasta. A gVCF file was produced for each sample, containing information on variants as well as reference positions.

```
gatk HaplotypeCaller \
  -R Homo_sapiens_assembly38.fasta \
  -I x.bam \
```

```
  -O x.g.vcf.gz \
  -ERC GVCF
```

The -ERC GVCF option allows downstream joint genotyping across multiple samples using gatk CombineGVCFs. Alternatively, for direct single-sample VCF generation, this step can be skipped.

To generate final genotype calls from a gVCF:

```
gatk GenotypeGVCFs \
  -R Homo_sapiens_assembly38.fasta \
  -V x.g.vcf.gz \
  -O x.vcf.gz
```

VCF annotation and enrichment

To extract clinically useful information, called variants were annotated and enriched using ANNOVAR, taking as input the final VCF file and the human reference genome hg19.

This annotation includes information from RefSeq (refGene) and Ensembl (ensGene), as well as cytoband. Population frequency data were integrated from ExAC, gnomAD exome and gnomAD genome, as well as known variant identifiers from dbSNP. Pathogenicity assessment was supported using combined functional impact predictions from dbNSFP (versions 3.0 and 4.2), REVEL and dbscSNV, while clinical associations were retrieved from ClinVar.

The final enriched VCF includes genomic position, functional impact, population frequency and clinical significance, enabling downstream filtering and clinical/research evaluation.

Final processing of enriched VCF files

After completing all steps and obtaining the final files, it became clear that due to the large data volume, straightforward presentation and analysis were challenging. To address this, an R script was developed which, for each VCF, split the called variants by chromosome. This enabled downstream analysis with conventional tools, for example Microsoft Excel.

Variant filtering to identify potentially predisposing variants

Sample description

The distribution of the 15 patient samples was as follows: (i) 1 sample from a patient with gastric cancer; (ii) 1 sample with multisystem cancer (multiple primary tumours); (iii) 3 samples from one family with renal cancer and endocrine tumours; and (iv) 10 samples from seven families with colorectal cancer and/or polyposis syndrome. Specifically, in two families three members were analysed simultaneously (trio analysis), in one family two members (duo analysis), and in the remaining families one member per family, to focus on potentially predisposing variants in known or new candidate cancer-predisposition genes.

Analysis design

In the context of WGS data analysis from families with strong suspicion of hereditary cancer predisposition, a variant filtering stage was implemented to investigate the genetic aetiology in depth. Filtering was performed based on the following criteria:

- Population frequency: variants with allele frequency <1% in ExAC and gnomAD were selected to focus on rare variants that may be associated with cancer predisposition.
- Clinical significance: variants classified as benign/likely benign under ACMG/AMP and based on clinical repositories (e.g., ClinVar) were excluded.
- Predicted pathogenicity (in silico): the REVEL score was used to retain variants with high (>0.7) or moderate (0.5–0.7) risk scores.
- Phenotype consistency: variants compatible with patient phenotypic features, as described by Human Phenotype Ontology (HPO) terms, were selected.

This approach enabled prioritisation of variants likely to be cancer-predisposing, while substantially reducing the number of potential false-positive findings.

4.2 Analysis – Results

In the final per-individual filtered output, we initially focused on coding regions (exons) and splice regions—i.e., exon–intron boundaries—for the detection of (a) single nucleotide variants (SNVs) and (b) small insertions/deletions (indels).

Using this approach, in individual 2663CRC we identified the heterozygous variant c.942+3A>T in the MSH2 gene. The MSH2 c.942+3A>T variant has previously been reported in ClinVar and is classified as pathogenic. It is located in intron 5 of MSH2 and has been shown at the mRNA level to cause aberrant splicing by skipping exon 5 of MSH2. This variant was confirmed by Sanger sequencing (see Figure 2).

Secondarily, focusing on specific genes based on each individual's personal and family history, we targeted intronic regions of the genes under study to detect potential deep intronic predisposing variants. Using this approach, in two individuals from the same family (2584CRC, 2584αCRC) we identified the heterozygous variant c.2458+976A>G in MSH2. The MSH2 c.2458+976A>G variant has previously been reported in ClinVar and is classified as likely pathogenic. It lies deep within intron 14 and has been shown to create a novel aberrant splice site, ultimately leading to premature protein termination, p.(Gly820Glufs*47). This variant was confirmed by Sanger sequencing (see Figure 2).

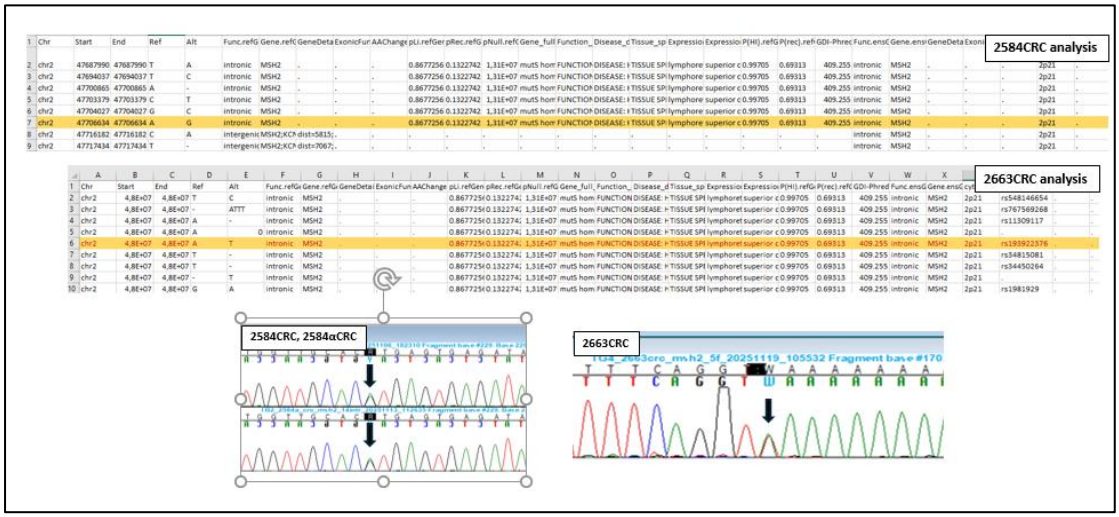

Figure 2. Images from WGS data analysis and Sanger sequencing chromatograms that led to the identification of the pathogenic MSH2 variants c.942+3A>T and c.2458+976A>G in the 2663CRC and 2584CRC families, respectively.

4.3 Conclusions

Identification of the above pathogenic MSH2 variants in the 2663CRC and 2584CRC families explains the phenotype of the tested individuals, as germline pathogenic variants in MSH2 predispose to Lynch syndrome, the most common hereditary colorectal cancer syndrome.

More specifically, patient 2663CRC was diagnosed with colorectal cancer at ages 37 and 66 years, and patient 2584CRC was diagnosed at ages 31 and 35 years. In the latter case, the family history includes multiple colorectal cancer cases, including the patient's father (2584αCRC) at age 64 years.

Individuals with Lynch syndrome also have increased risk for malignancies beyond colorectal cancer, including cancers of the endometrium, stomach, urothelial tract, ovaries, pancreas and brain. In addition, the autosomal dominant inheritance typical of pathogenic MSH2 variants impacts blood relatives of the tested individuals, who can now be identified as carriers through targeted genetic testing. In summary, WGS enabled clarification of the genetic basis in 2 of the 10 families under study; for the remaining cases, the underlying genetic cause has not yet been identified based on analyses performed so far. In both diagnosed families, the causal variants affect splicing and could not be detected by conventional routine diagnostic sequencing. These results highlight the contribution of WGS in revealing pathogenic variants that escape standard diagnostic approaches and support its utility in both research and clinical workflows. Given the broad tumour spectrum in Lynch syndrome, establishing a clear genetic diagnosis in these two families affects both clinical management of affected individuals and preventive care of asymptomatic relatives.

5 Rare diseases

5.1 Objective

Rare diseases are clinical entities that affect fewer than 1 in 2,000 individuals. To date, more than 7,000 rare diseases are known and new entities continue to emerge. Almost 400 million people worldwide are affected by rare diseases, while in Greece it is estimated that approximately 1,000,000 individuals are affected. Due to high phenotypic and genetic heterogeneity, accurate diagnosis is challenging and many patients remain undiagnosed. Because the majority of rare diseases have a genetic aetiology, the application of next-generation sequencing (NGS) technologies has proven particularly valuable for identifying underlying genetic alterations and achieving a definitive diagnosis. This approach enables simultaneous testing of multiple genes in substantially less time and at lower cost compared with conventional sequencing. Nevertheless, routine diagnostic sequencing focuses on coding regions (exons) and/or exon–intron boundaries (splice regions), which fails to identify the underlying genetic cause in a significant proportion of patients.

The aim of this work is the in-depth genetic investigation of patients affected by a rare disease with strong suspicion of a genetic basis, who, despite diagnostic-level testing, remain without a clear genetic diagnosis. WGS enables the study of genomic regions within already known genes that escape standard diagnostic approaches because they lie outside coding regions, as well as the identification of new genes that may be related to disease predisposition.

5.1.1 Study cohort

This cohort includes 15 individuals with a clear pattern of genetic disease and a strong phenotype. Specifically, 14 patients presented either to the University General Hospital of Patras, or to Karamandaneio, or to the Children's Hospital "Agia Sofia", but no variant associated with their phenotype was identified via WES or panel sequencing. Notably, in three cases first-degree relatives were also included in the study, as outlined in Table 2. More specifically, patients 1 and 2 are parent and child, respectively, with a similar phenotype. Patients 7, 8 and 9 are siblings with a similar phenotype and their unaffected mother, respectively. Finally, patients 14 and 15 are siblings with a similar phenotype.

Table 2. Patient cohort of the project (rare diseases)

No.	Sex	Phenotype	Prior testing
1	Male	Hypercholesterolaemia	Negative WES
2	Female	Hypercholesterolaemia	Negative WES
3	Male	Recurrent respiratory infections	Negative WES
4	Female	Gait disorder; broad-based gait;	Negative WES

		lower-limb hyperreflexia; limb dystonia; periventricular white-matter hypodensities	
5	Female	Developmental delay; myoclonic seizures; microcephaly; orofacial apraxia	Negative WES
6	Female	Liver failure; gait disorder; intellectual disability; dystonia; infantile onset	
7	Male	Severe intellectual developmental disorder; severe global developmental delay; generalized seizures	
8	Male	Severe intellectual developmental disorder; severe global developmental delay; generalized seizures	
9	Female	No phenotype	
10	Male	Dystonia; cerebral palsy; renal dysplasia; kidney transplant; corpus callosum anomaly; hypothyroidism	Negative WES
11	Male	Hypotonia; absent reflexes; apnoea; dysphagia	Negative WES
12	Male	Severe intellectual developmental disorder; delayed motor milestones; delayed speech/language; febrile seizures; generalized hypotonia; hyperthermia	
13	Male	Seizures; drug-resistant epilepsy	
14	Male	Severe intellectual developmental	

		disorder; learning difficulties; autism	
15	Male	Severe intellectual developmental disorder; severe global developmental delay; delayed motor milestones; delayed speech/language; learning difficulties; autism	

5.1.2 Sequencing methodology (WGS)

We constructed a total of 15 WGS libraries using either the MGIEasy FS PCR Free DNA Library Prep Set, v1.2 (MGI Tech Co., Ltd.), according to the MGIEasy FS PCR Free DNA Library Prep Set user manual, or the MGIEasy Fast PCR-FREE FS Library Prep Set, v2.0, according to the MGIEasy Fast PCR-FREE FS Library Prep Set user manual (Version 1.0). In brief, 900–1000 ng DNA was used for WGS library preparation; fragmentation time and size-selection strategy were determined based on the input DNA amount and according to the manufacturer's instructions. Library DNA concentration was assessed using the Invitrogen Qubit 4 fluorometer and the Invitrogen Qubit dsDNA HS Assay kit.

WGS libraries were batched in pools of 5 samples and combined equimolarly to a total of 300 ng. Denaturation, circularisation, enzymatic digestion, cleanup and QC were performed as described above. Sequencing was performed on a DNBSEQ-G400 platform at the Genomics Unit of BSRC Alexander Fleming, using the G400 FCL PE150 high-throughput sequencing set, following the manufacturer's user manual.

5.1.3 Quality control and mapping

Initial sequencing QC was performed with FastQC. Low-quality reads were trimmed using TrimGalore. Only reads with length ≥50 bp that retained correct read pairing were included in the final analysis.

Reads were mapped to the reference genome (hg38) using BWA (bwa-mem) with default parameters (per lane). Pre-processing was performed with samtools (name sorting, fixmate, coordinate sorting, markdup for duplicate marking prior to variant calling).

After mapping, genome coverage was quantified using Picard and a second QC was performed according to procedures available at https://github.com/moulos-lab/genomics-facility-processes.

5.1.4 Genome variation analysis

For the 15 samples above, genome variation analysis was performed against the reference genome (hg38). First, BAM format validation was performed using ValidateSamFile and Base Quality Score Recalibration (BQSR) was applied using known variant sites (--known-sites Homo_sapiens_assembly38.dbsnp138.vcf and --known-sites Mills_and_1000G_gold_standard.indels.hg38.vcf.gz). Read group information was then adjusted, BAM headers were replaced (AddOrReplaceReadGroups), and BAM indices were created (indexing). Variant calling was performed with HaplotypeCaller using the reference Homo_sapiens_assembly38.fasta. Finally, filtering of low-confidence variants using bcftools is recommended, based on variant quality (QUAL) and read depth (DP), to remove low-reliability calls.

6 Discussion

This deliverable establishes a layered interpretation framework for intergenic regulatory variants by integrating cCRE-based regulatory compartment annotation with tissue-filtered chromatin interaction links and orthogonal functional evidence layers (eQTL support and TFBS context). Across BRCA and BLCA, the results show that a large fraction of somatic variation overlaps regulatory space, but that raw overlap counts alone are insufficient for biological interpretation due to the substantially larger genomic footprint of enhancer annotations relative to promoters. When this annotation-size bias is controlled by computing length-normalized mutational densities, promoters consistently emerge as the highest-density compartment, indicating that promoter-proximal disruption is a recurrent feature of intergenic regulatory burden and should remain a primary prioritization focus—particularly in BLCA, where the promoter density signal is strongest.

A central biological insight from the functional evidence layers is that BRCA and BLCA exhibit sharply divergent regulatory architectures. In BRCA, eQTL support is overwhelmingly enhancer-dominated, consistent with distributed long-range regulation where distal elements materially contribute to expression variability and regulatory wiring. In contrast, BLCA shows pronounced promoter centrality in eQTL intersections, driven by a very small number of promoter variants with disproportionately large numbers of eQTL associations. This implies a regulatory configuration in which promoter-proximal perturbations may produce more direct and detectable transcriptional consequences in this cohort context. TFBS localization and disruption analyses reinforce this contrast: BRCA shows enhancer-centric TFBS overlap and TFBS-disrupting variants predominantly associated with enhancers, whereas BLCA TFBS signal is almost entirely promoter-associated, consistent with direct perturbation of transcriptional initiation control. Together, these layers converge on a coherent model in which BRCA is comparatively enhancer-driven and BLCA comparatively promoter-driven, with distinct implications for how regulatory candidates should be prioritized and interpreted in each tumour type.

These findings also motivate a practical two-tier prioritization strategy for intergenic regulatory candidates. First, promoter-associated candidates represent high-density, high-priority hypotheses for direct transcriptional effects, particularly in BLCA. Second, enhancer-associated candidates require additional constraint through tissue-specific gene-linking (chromatin interactions) and functional support (eQTL evidence and TFBS disruption), especially in BRCA where enhancer mechanisms dominate functional support. Importantly, this layered approach reduces the risk of prioritizing candidates solely due to large enhancer footprints and instead emphasizes evidence-weighted mechanistic plausibility.

Population cross-referencing with UK Biobank provides complementary context but also clarifies expected limitations of replication for cohort-specific regulatory signals. The modest overlap observed is consistent with the ultra-rare and often somatic nature of many TCGA regulatory candidates, whereas GWAS-oriented

population resources are enriched for common germline variation. The additional observation that shared BRCA–UK Biobank variants do not overlap predicted splice-impact regions supports a predominantly regulatory interpretation for these shared signals and underscores the value of separating regulatory interpretation pipelines from coding/splice-impact workflows.

Finally, while survival association testing provides an outcome-oriented prioritization layer, regulatory interpretation should remain anchored to reproducible evidence. Effect sizes in rare-variant survival analyses can be sensitive to carrier counts and cohort event structure; therefore, survival results are best viewed as a prioritization signal rather than standalone proof of causality. The main outcome of this work is a structured map of intergenic regulatory candidates with mechanistic context (promoter vs enhancer class, TFBS localization/disruption, eQTL support, and tissue-filtered gene links) that can be advanced to downstream validation, including replication in independent datasets, integration with expression and chromatin activity profiles, and targeted functional assays to confirm regulatory effects on putative target genes.

In parallel, the Greek WGS cohort findings demonstrate the clinical value of extending analysis beyond conventional diagnostic targets. Identification of pathogenic and likely pathogenic deep intronic splice-altering variants in MSH2 in unresolved hereditary cancer families illustrates how whole-genome sequencing can reveal causal mechanisms missed by routine exon-focused testing. In this sense, the regulatory landscape analysis and the Greek cohort findings are complementary: both emphasize that clinically relevant genetic mechanisms frequently lie outside canonical coding changes and require integrative evidence and broader genomic coverage to resolve.

7 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] 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

[3] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, Bette, Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine, 12(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

[4] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature, 562(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z