High-throughput Functional Screening of *ATM* Gene with Saturation Genome Editing Using Prime Editing

Lee, Kwangseob

Department of Medicine
Graduate School
Yonsei University

High-throughput Functional Screening of ATM Gene with Saturation Genome Editing Using Prime Editing

Advisor Hyongbum Kim

**A Dissertation Submitted
to the Department of Medicine
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy in Medical Science**

Lee, Kwangseob

June 2025

High-throughput Functional Screening of ATM Gene with Saturation Genome Editing Using Prime Editing

This Certifies that the Dissertation of Lee, Kwangseob is Approved

Committee Chair	**Lee, Seung-Tae**
Committee Member	**Kim, Hyongbum**
Committee Member	**Rha, Sun Young**
Committee Member	**Nam, Eun Ji**
Committee Member	**Kim, Younggwang**

Department of Medicine
Graduate School
Yonsei University
June 2025

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude for the excellent environment and support provided by my institution, which enabled me to carry out this research. First and foremost, I would like to extend my sincere appreciation to my parents, Byungrip Lee and Bongsoon Park, who have raised me with unwavering dedication, and to my brother, who has always supported me throughout my journey. I also want to express my gratitude and love to my wife, Hayn Moon, who has always stood by my side even in hard times, to our adorable son, Jeongwon Lee, and to "Podo," whom we will meet this summer.

I would also like to express my gratitude to Professor Hyongbum Kim for his guidance and support throughout my journey. His passion for science and curiosity have taught me countless lessons that go far beyond what can be learned from books.

I would also like to thank Joon-Goo Min for his dedication and hard work in carrying out this research with me. I am truly delighted that we were able to achieve such great results together. Without him, completing this study would not have been possible. Additionally, I want to extend my appreciation to all the researchers in our lab, from whom I have learned so much. It has been a privilege and a great fortune to conduct research alongside such talented colleagues. I am also deeply grateful to Goosang Yu, Yusang Jung, Hyeong-Cheol Oh, and MinYoung Lee for helping me get off to a strong start in my research journey. Thanks to Young-hye Kim and Seonmi Park for providing an excellent research environment so that I could concentrate on my work. This research was sponsored by the Hur Jiyoung Foundation.

Lastly, I would like to express my heartfelt gratitude to Professors Seung-Tae Lee, Jong Rak Choi, Sun Young Rha, Sang-Guk Lee, and John Hoon Rim for guiding me on my path as a physician-scientist. Their support and mentorship have been truly invaluable.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

High-throughput Functional Screening of *ATM* Gene with Saturation Genome Editing Using Prime Editing

ATM, a large gene with 63 exons, plays a critical role in the DNA damage response, and its loss-of-function increases cancer risk and affects the prognosis of cancer patients. However, interpreting the functional impact of *ATM* variants remains challenging, because most are variants of uncertain significance (VUSs). Here, we used prime editing and deep learning to assess the functions of all 27,513 possible single nucleotide variants (SNVs) in *ATM*. By leveraging haploidization and olaparib, a PARP inhibitor, we experimentally evaluated 23,092 SNVs, thereby identifying critical residues. Using cancer genetics data and UK Biobank data, we found that our results are useful for estimating both cancer risk and prognosis. We also developed a deep learning model, DeepATM, which predicted the functional effects of the remaining 4,421 SNVs with unprecedentedly high accuracy. This complete evaluation of *ATM* variants supports precision medicine and provides a framework for addressing VUSs in other genes.

Key words : VUS, prime editing, functional screening, saturation genome editing

1. INTRODUCTION

1.1. The hurdles of clinical genetics

The rapid advancement of massively parallel sequencing technologies has revolutionized genetic diagnosis for hereditary diseases and cancers, significantly influencing clinical practice. However, a major challenge remains: the growing number of variants of uncertain significance (VUSs), which complicates genetic interpretation. Among different variant types, missense mutations are particularly difficult to assess functionally without direct experimental validation, unlike synonymous, nonsense, and indel variants. Additionally, the number of observed variants in a gene tends to increase with gene length, making large genes like *ATM* and *BRCA1/2* especially challenging to interpret at a saturation level [1]. Notably, *ATM* variants have frequently been classified inconsistently across different clinical laboratories [2].

1.2. Significance of *ATM* gene

The *ATM* (Ataxia-telangiectasia mutated) gene encodes a key regulator of the DNA damage response and serves as a tumor suppressor, ensuring cellular stability under stress conditions [3]. Biallelic loss-of-function in *ATM* result in ataxia-telangiectasia (MIM# 208900), a recessive disorder characterized by progressive cerebellar ataxia, immune dysfunction, insulin resistance, infertility, increased sensitivity to ionizing radiation, and a heightened risk of malignancies [4, 5]. Furthermore, heterozygous pathogenic *ATM* variants have been linked to an increased risk of various cancers, including breast, colorectal, pancreatic, and prostate cancers [6-10]. As a result, *ATM* is included in most hereditary cancer gene panels. The NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines®) advise genetic panel testing that includes *ATM* for individuals at high risk, along with regular cancer screenings for carriers of pathogenic *ATM* variants [11]. Consequently, global research efforts are focused on clarifying *ATM*'s role in cancer and standardizing variant classification [12-14]. Despite these initiatives, interpreting *ATM* variants remains a significant challenge.

In addition to its role in cancer predisposition, *ATM* is an important target for cancer therapies, with loss of function serving as a biomarker for treatment selection. For instance, olaparib, a poly ADP-ribose polymerase (PARP) inhibitor, is approved for metastatic castration-resistant prostate cancer patients with loss-of-function mutations in homologous recombination repair genes, including *ATM* [15-17]. Therefore, systematically assessing the functional consequences of all possible single nucleotide variants (SNVs) in *ATM* could improve treatment strategies for affected patients. Additionally, prognosis varies among cancer patients with *ATM* loss-of-function mutations depending on cancer type—those with breast and hematologic malignancies generally have worse outcomes, whereas bladder cancer patients may experience better prognoses [18-20]. Thus, comprehensive functional evaluation of *ATM* variants would be highly valuable for predicting cancer risk, diagnosing ataxia-telangiectasia, informing cancer treatment strategies, and estimating patient prognosis.

1.3. Content and significance of this study

In this study, we employed prime editing [21] and deep learning to systematically analyze the functional impact of *ATM* variants across the entire coding sequence, spanning 62 exons. Using prime editing, we generated and analyzed 23,092 *ATM* SNVs, covering 84% of the 27,513 theoretically possible SNVs. Our findings revealed that *ATM* haploidization—caused by a large deletion in one allele—along with the selective pressure of olaparib to deplete cells harboring loss-of-function *ATM* variants in the remaining allele, significantly improved signal-to-noise ratios in high-throughput functional assessments. Furthermore, we identified a specific region in *ATM* that is particularly intolerant to missense mutations, located within the kinase domain responsible for interactions with p53. Most notably, by reevaluating previously published clinical datasets, including UK Biobank (UKB) data, using our functional findings, we demonstrated that our variant assessments enhance predictions of cancer risk and patient prognosis. Additionally, we developed DeepATM, a deep learning model that predicts *ATM* variant functionality with exceptional accuracy. By integrating DeepATM with our experimental data, we determined the functional effects of all 27,513 potential *ATM* SNVs. This comprehensive approach not only provides critical insights into *ATM* but also offers a scalable framework for evaluating variants in other genes, further advancing precision medicine for individuals with *ATM* mutations.

2. MATERIALS AND METHODS

2.1. General cell culture conditions

HEK293T (ATCC) cells and all HCT116-derived cell lines were maintained in high-glucose DMEM (Sigma-Aldrich, D6429) supplemented with 10% fetal bovine serum (RDT) and 1% penicillin/streptomycin (Gibco) at 37°C in a 5% CO_2 atmosphere. Antibiotics were excluded during transfection. Cells were passaged every three to four days.

2.2. Generation of the *ATM*-haploid HCT116 cell line

To create *ATM*-haploid cells, wild-type HCT116 cells were plated at a density of 4×10^6 cells in a 100-mm dish 24 hours before transfection. The transfection utilized PEI Prime™ linear polyethylenimine (Sigma-Aldrich) with a plasmid cocktail that included a SpCas9-encoding plasmid (pRGEN-Cas9-CMV/T7-Puro-RFP; sourced from ToolGen, Republic of Korea), and two sgRNA-encoding plasmids (pRG2; Addgene #104174) targeting approximately 30 bp upstream of *ATM* exon 1 and 80 bp downstream of *ATM* exon 63. Additionally, a single-stranded oligodeoxynucleotide (ssODN; synthesized by Bionics, Republic of Korea) was included to facilitate the formation of large deletions spanning 146,380 bp. The standard transfection protocol involved mixing 60 μL of PEI with 500 μL of Opti-MEM (Gibco) and combining this with 20 μg of the DNA mixture in a 3.5:1:1 mass ratio for SpCas9, sgRNA1, and sgRNA2, respectively, in another 500 μL of Opti-MEM, resulting in a final volume of 1 mL. After a 15-minute incubation at room temperature, the mixture was added to the HCT116 cells. The sequences for sgRNA targets, with the PAM shown in parentheses, and the ssODN were:
Large deletion sgRNA1 (5'-*ATM*): 5'-AGGGCGGGGAGGACGACGA(GGG)
Large deletion sgRNA2 (3'-*ATM*): 5'-AAGGAGAAAGCAGTGAGCA(AGG)
ssODN:
5'-TTCCGTCCTCAGACTTGGAGGGGCGGGGATGAGGAGGGCGGGGAGGACGA
GCAAGGCAGGCATAGTCTGCCTATATAAAGCTCCCAATCTGAGGAGGATA-3'

Following transfection, fresh culture medium was provided after 24 hours, and puromycin (1 μg/mL, Gibco) was added at 48 hours for selection, which continued for two days. After selection, cells were cultured in puromycin-free medium for expansion. Between days 7 and 10 post-transfection, single cells were sorted by flow cytometry into 96-well plates. After two weeks, individual clones were expanded in duplicate, one set for growth and the other for PCR verification of the large deletion. Genomic DNA (gDNA) was extracted using a lysis buffer (50 mM Tris-HCl, 1 mM EDTA, 0.05% SDS, 0.2 mg/mL Proteinase K from Enzynomics, Republic of Korea) at 56°C for 1 hour, followed by enzyme inactivation at 80°C for 15 minutes. Clones with *ATM* deletions were identified by PCR using primers flanking the cut sites (FP1 and RP3 in **Figure 3A**; **Table 1**), resulting in a ~300 bp product in the presence of the deletion. PCR conditions were: 95°C for 3 minutes; 35 cycles of 95°C for 30 seconds, 58°C for 30 seconds, and 72°C for 30 seconds; with a final extension at 72°C for 3 minutes. PCR products were analyzed on a 2% agarose gel. Clones with one copy of *ATM* deleted and lacking the c.3380C>T variant in the other copy were confirmed by sequencing exon 23. Final validation of *ATM*-haploid clones involved deep sequencing of four regions: (i) the newly formed deletion junction, (ii) the upstream sgRNA1 region, (iii) the

downstream sgRNA2 region, and (iv) the c.3380 region in *ATM*. The PCR amplification conditions for these regions matched those described above, with annealing at 54°C for regions (ii), (iii), and (iv).

2.3. Generation of *ATM*-knockout HCT116 cell lines

The *ATM* semi-KO HCT116 cell line was created using a transfection protocol similar to that used for generating *ATM*-haploid cells. Wild-type HCT116 cells were transfected with a plasmid encoding SpCas9 and two sgRNA-expressing plasmids in a 3.5:1:1 mass ratio. One of the sgRNA plasmids contained a standard 20 bp guide RNA designed to target the c.3380T allele, introducing a frameshift mutation in that allele. The second plasmid carried a truncated 15 bp catalytically inactive dead-guide RNA (dgRNA), specifically targeting the c.3380C allele, preventing its modification by SpCas9.

To create the *ATM*-haploid-KO cell line, *ATM*-haploid HCT116 cells were transfected using the same approach. A plasmid encoding SpCas9 and an sgRNA-expressing plasmid were co-transfected in a 3.5:1 mass ratio. The sgRNA targeted the c.3380C allele to introduce a frameshift mutation. Following transfection, deep sequencing was used to validate the modifications in both cell lines.

ATM-c.3380C-dgRNA: 5’-CTTGAAAGCTCAGGA(AGG)
ATM-c.3380T-sgRNA: 5’-CATACTTGAAAGTTCAGGA(AGG)
ATM-c.3380C-sgRNA: 5’-CATACTTGAAAGCTCAGGA(AGG)

2.4. Generation of prime editor-expressing cell lines

To introduce PE2max into *ATM* semi-KO and *ATM*-haploid HCT116 cells, lentiviral particles containing pLenti-PE2max-P2A-BSD (Addgene #191102) were produced using the standard PEI transfection protocol. Lentivirus was generated by co-transfecting HEK293T cells in four 150-mm culture dishes with psPAX2 (Addgene #12260), pMD2.G (Addgene #12259), and pLenti-PE2max at a mass ratio of 3:1:4. The culture medium was refreshed 24 hours after transfection. After 72 hours post-transfection, the viral supernatant was collected, filtered through a 0.45 μm bottle-top filter, and concentrated using Vivaspin Turbo 15 (Sartorius) to obtain approximately 2 mL of lentiviral concentrate (Lenti-Conc). A small fraction (2 mL) of the initial supernatant was set aside for titration (Lenti-Titer).

For transduction, cells were plated at a density of 5×10^5 cells per well in a 6-well plate 24 hours prior to infection. Lentiviral particles (Lenti-Conc, 2 mL) and various volumes of Lenti-Titer (50–1,000 μL) were combined with polybrene (Sigma-Aldrich) in a total of 4 mL of medium, maintaining a final polybrene concentration of 8 μg/mL. Following a 24-hour incubation, the medium was replaced with fresh culture medium, and selection with 8 μg/mL blasticidin (Invivogen) was initiated 48 hours after transduction. Cells were maintained under selection for at least two weeks, with passaging every three to four days. The lentiviral titer was determined after confirming that all control cells (subjected to blasticidin selection without transduction) had died. To further enhance prime editing efficiency, transduction and selection were repeated in cells that had already integrated the PE2max cassettes.

2.5. Construction of plasmids expressing sgRNAs

The pRG2 vector was digested with BsaI-HF®v2 (NEB) for four hours and subsequently purified following electrophoresis on a 2% agarose gel. The gel-extracted DNA fragment was isolated using the MEGAquick-Spin Total Fragment DNA Purification Kit (iNtRON, Republic of Korea). Oliogonucleotides containing spacer sequences (5'-G+N19 or N15) with BsaI overhangs were designed (synthesized by Bionics, Republic of Korea). These oligonucleotide strands were phosphorylated using T4 Polynucleotide Kinase (Enzynomics) according to the manufacturer's instructions. The prepared linearized vector was then ligated with the inserts using T4 DNA Ligase (NEB) following the manufacturer's protocol.

2.6. Preparation of epegRNA libraries and electroporation

To perform saturation editing of all *ATM* exonic coding sequences—including splicing regions within 5 bp of exon-intron boundaries—epegRNA libraries were generated in eight subsets, each covering 5 to 10 exons. The pooled oligonucleotides required for library construction were synthesized using array synthesis (Twist Bioscience). The synthesized oligonucleotides, ranging from 268 to 276 bp in length depending on exon size, incorporated the following elements in common:

- A 17 bp homology sequence at the 3' terminus of the human U6 promoter.
- A 19 bp guide RNA (gRNA) sequence with a 'G' at the 5' end.
- An optimized SpCas9 sgRNA scaffold for enhanced performance.
- A reverse transcriptase template (RTT) and primer binding site (PBS) designed for precise genome editing.
- An 8 bp linker sequence generated using pegLIT tools to enhance prime editing.
- A 37 bp tevopreQ1 sequence followed by a 6 bp poly-T sequence.
- An 18 bp barcode with a random buffer sequence to equalize the overall oligonucleotide length.
- A 19 bp sequence for exon-specific amplification within the subset library.

For each subset, 4.8 ng of oligonucleotides were amplified via PCR using Q5 High-Fidelity DNA Polymerase (NEB). A common forward primer containing a 38 bp human U6 promoter overhang and a 17 bp homology sequence was used, along with a reverse primer that included the 19 bp exon-specific sequence and a 37 bp 3' overhang sequence (**Table 2**). Each exon library was amplified across six PCR reactions (50 µL per reaction) containing 800 pg of template DNA, 25 pmol of each primer, and 1 µL of Q5 polymerase. The PCR cycling conditions were as follows:

- Initial denaturation: 98°C for 3 minutes
- 17 cycles of:
  98°C for 30 seconds
  61°C for 30 seconds
  72°C for 2 minutes
- Final extension: 72°C for 3 minutes

The PCR products were purified, and correctly sized amplicons were extracted using 2% agarose gel electrophoresis. Meanwhile, pLenti-gRNA-Puro (Addgene #84752) was digested with BsmBI (Enzynomics) at 55°C for 6 hours, followed by purification via gel electrophoresis.

To assemble the epegRNA libraries, the amplified oligonucleotide pool was combined with the linearized pLenti-gRNA-Puro plasmid using NEBuilder® HiFi DNA Assembly Master Mix (NEB). The assembled constructs were concentrated through isopropanol precipitation with GlycoBlue™

Coprecipitant (Invitrogen) and subsequently electroporated into EC100 electrocompetent cells (Lucigen) using a MicroPulser (Bio-Rad). After a one-hour recovery in SOC medium (Welgene), transformed cells were plated on Luria-Bertani agar square plates (Dulbecco) supplemented with 75 μg/mL carbenicillin (Sigma-Aldrich) and incubated for 12 to 16 hours. Plasmid DNA was extracted from the resulting bacterial colonies using the Nucleobond Xtra Midi EF Kit (Macherey-Nagel).

2.7. Lentivirus production of epegRNA libraries

HEK293T cells were plated in 150-mm culture dishes at a density of 10×10^6 cells per dish. After incubating for 18 to 24 hours, transfection was carried out following our standard PEI protocol. In brief, 15 μg of psPAX2, 5 μg of pMD2.G, and 20 μg of plasmids containing the exon-specific libraries were combined with 120 μL of PEI in a total volume of 2 mL Opti-MEM. Following a 15-minute incubation, the mixture was added to the cells. Between 20 and 24 hours post-transfection, the culture medium was replaced with 30 mL of fresh, antibiotic-free medium. The lentivirus-containing supernatant was then collected 72 hours after transfection. To remove cellular debris, the supernatant was centrifuged at approximately 350 g for 3 minutes, passed through a 0.45-μm Sartolab RF 50 PES vacuum filter (Sartorius), and stored in aliquots at -80°C.

2.8. Lentiviral transduction

A modified colony formation titration assay was used to measure the titer of lentivirus aliquots for each exon-specific library. To ensure that each cell incorporated only a single copy of the epegRNA cassette into its genome, lentivirus transduction was carried out at a multiplicity of infection (MOI) below 1, allowing for a single intended edit per cell. PE2max-expressing cells were plated in multiple 150-mm culture dishes at a density of 4×10^6 cells per dish, 24 hours prior to transduction. Lentivirus aliquots were diluted in a series of concentrations and treated with polybrene (Sigma-Aldrich) at 8 μg/mL before being added to the PE2max-expressing cells. The culture medium was replaced 24 hours post-transduction, and selection of transduced cells began 48 hours later using puromycin at a concentration of 1 μg/mL. The experiment included both positive controls (PEmax-expressing HCT116 cells not transduced with epegRNA libraries and untreated with puromycin) and negative controls (puromycin-treated cells without transduction). Once all negative control cells had died, the appropriate volume of lentivirus aliquot needed to maintain an MOI below 1 was determined by calculating the percentage of viable cells compared to the total cell count in the mock-treated group.

2.9. High-throughput functional assay for *ATM* variants

Each exon-specific library was transduced and analyzed through separate experiments, with each experiment targeting a single exon of the *ATM* gene using a distinct library of epegRNAs. In total, we performed 62 independent experiments, each focusing on one exon of *ATM*. For the pooled experiments, we plated 1.6×10^7 to 2.4×10^7 cells in several 150-mm culture dishes at a density of 4×10^6 cells per dish to ensure adequate coverage of the library (typically more than 5,000 × the size of each library). Forty-eight hours post-transduction, the cells were cultured for an additional 11 days in the presence of 1 μg/mL puromycin to allow for prime editing. On the 13th day after transduction (day 0), half of the cell population was harvested for gDNA extraction, while the remaining cells were reseeded and divided into two groups: one treated with 800 nM olaparib and

the other with DMSO as a control. Equal amounts of olaparib (Selleckchem) and DMSO (Sigma-Aldrich) were applied to each plate. After an additional 10 days of culture, all remaining cells were harvested for gDNA extraction. Olaparib was dissolved in DMSO and stored in aliquots at -80°C. Additionally, non-transduced cells were cultured, collected, and used as unedited samples to control for PCR or sequencing errors for each exon.

2.10. Genomic DNA extraction and deep sequencing

Genomic DNA was extracted using the Wizard® Genomic DNA Purification Kit (Promega), following the instructions provided by the manufacturer. The gDNA was then amplified through two rounds of PCR. In the first round, 80 to 160 μg of purified gDNA (providing over 5,000× coverage of each library, assuming 6.6 μg of gDNA per 10^6 cells) was amplified using exon-specific primers with PrimeSTAR® GXL polymerase (Takara Bio) (**Table 3**). For each exon, the PCR conditions, including the annealing temperature and cycle number, were optimized within the ranges of 52 to 60°C and 29 to 30 cycles. The PCR reactions, with a total volume of 50 μL, contained 25 pmol of each primer and 4 μg of gDNA, and were cycled as follows: an initial denaturation at 98°C for 3 minutes; 29-30 cycles of 98°C for 30 seconds, 52-60°C for 30 seconds, and 68°C for 1 minute; followed by a final extension at 68°C for 3 minutes. After amplification, the PCR products were pooled and purified using the QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's protocol. The amplicons were then purified using 2% agarose gel electrophoresis and gel extraction. For the indexing PCR, 40 to 60 ng of purified PCR product from the first round was used with Pfu polymerase (Solgent) and Illumina indexing primers, following the cycling conditions: 95°C for 3 minutes; 8 cycles of 95°C for 30 seconds, 57°C for 30 seconds, and 72°C for 1 minute; followed by a final extension at 72°C for 3 minutes. The products from the second PCR round were also purified with the same kit. Finally, the amplicons were sequenced on a NovaSeq 6000 (Illumina).

2.11. Design of the epegRNA libraries

Using the human reference genome (hg38) and whole exome sequencing data from *ATM*-haploid cells, we developed libraries designed to introduce all possible single nucleotide variants (SNVs) within the *ATM* coding region. Additionally, we included a synonymous substitution near the target editing site to help distinguish between PCR/sequencing errors and true SNV edits. To create highly efficient saturation prime editing libraries, we employed the DeepPrime-FT model, which predicts the efficiency of PE2max combined with epegRNA and an optimized scaffold in HEK293T cells, to calculate DeepPrime scores for potential epegRNAs that could induce SNVs. The length of the RTT was limited to 40 bp, while the PBS length was restricted to 17 bp. For regions with rare target NGG PAM sites, we chose NGA or NAG PAMs instead, taking advantage of the prime editor's ability to recognize non-canonical PAMs. To ensure precise editing at the intended site, we excluded epegRNAs with a right homology arm (RHA) shorter than 4 bp. We selected the three top-scoring epegRNAs for each SNV edit, ensuring that at least two spacers were included. For SNV edits with fewer than three epegRNAs linked to an NGG PAM, the highest-scoring epegRNAs targeting NGA or NAG PAMs were used to achieve three epegRNAs.

The position for the additional synonymous substitution edit was determined based on the following criteria: (i) it must be located in a different codon than the intended edit, (ii) preference was given to substitutions within exonic regions, excluding areas within 2 nucleotides of the exon

boundary or within 5 nucleotides of the exon-intron or intron-exon junctions, as these could affect splicing, (iii) substitutions that disrupt the PAM sequence (GG) were prioritized, (iv) when PAM disruption was not feasible, substitutions in the left homology arm closest to the PAM site were favored, and (v) priority was given to substitutions in the RHA nearest to the intended edit.

2.12. Individual functional evaluation of single variants

The epegRNA sequences with the highest DeepPrime scores were designed to induce SNVs that produce the intended edit, without introducing any concurrent synonymous mutations. These sequences consisted of three annealed components: (i) a spacer sequence with overhangs for cloning, (ii) an optimized SpCas9 sgRNA scaffold sequence with overhangs, and (iii) an annealed epegRNA RTT-PBS with poly-T sequences and overhangs. The annealed sequence was then cloned into a BsmBI-linearized pLenti-crRNA-Puro vector.

ATM-haploid cells were transduced with lentivirus carrying the epegRNA sequences. Two days after transduction, the medium was replaced with fresh puromycin-containing medium, and the cells were cultured for an additional seven days to allow for editing. The transduced cells were seeded into 6-well plates at a density of 2.0×10^5 cells per well, with each well treated with either DMSO or olaparib at a concentration of 800 nM, similar to the conditions used in the high-throughput assay. Cells were incubated until they reached 80% confluence. Genomic DNA was collected for deep sequencing at two or more time points during cell passaging, and the relative fold change of the SNVs was calculated by comparing the data to Day 0.

2.13. High-throughput functional assay for BRCA1 variants

The epegRNA libraries targeting BRCA1 exons 4 and 19, as outlined in a prior study [22], were employed in the experiments. All procedures, including preparations and assays conducted in the *ATM*-haploid cell line, were carried out using the same approach as for the *ATM* exons.

2.14. Off-target effect analysis

DNA sequences at both the on-target and possible off-target locations were analyzed using deep sequencing in two clones with non-functional variants. These variants were created using the epegRNAs utilized in the high-throughput functional assessments of the SNVs. To locate potential off-target sites for the epegRNAs, we used Cas-OFFinder [23], allowing for up to two mismatches or one mismatch combined with either an insertion or deletion in the guide sequence relative to the target sequence. gDNA was amplified with custom primers specific to the potential off-target sites (**Table 4**), followed by deep sequencing.

2.15. Western blots

To evaluate ATM protein expression and the phosphorylation levels of ATM and CHK2, 4 million cells were plated in 100-mm dishes. The next day, the cells were treated with 3 μM etoposide (Selleckchem) for 1 hour, then immediately collected using scrapers. Protein extraction was performed using PRO-PREP™ Protein Extraction Solution (iNtRON) with added phosphatase inhibitors (1 mM Na3VO4 and 1 mM NaF), and protein concentration was determined using the

Bio-Rad Protein Assay Dye Reagent Concentrate (Bio-Rad). A total of 20 μg of protein was loaded into each well of a 6-13% PAGE gel for electrophoresis. After electrophoresis, proteins were transferred to a nitrocellulose membrane (Cytiva). The membrane was cut into strips, each targeting a specific protein. To block non-specific binding, membranes were incubated for 1 hour with 3% bovine serum albumin (GenDEPOT) in Tris-buffered saline with 0.1% Tween-20 (TBS-T). The membranes were incubated overnight at 4°C with the primary antibody (diluted 1:1,000 in the blocking solution), followed by washing and a 1-hour incubation at room temperature with the peroxidase-conjugated secondary antibody (diluted 1:2,000 in 0.1% TBS-T). The target protein bands were detected using an ECL detection system (Cytiva) and imaged with the Amersham Imager 600 (Cytiva). Image processing was done with ImageJ software, version 1.53 h (National Institutes of Health).

2.16. Raw sequencing data filtering and analysis

To identify SNVs in deep sequencing data from cells transduced with libraries targeting exon 2 through exon 63, an SNV reference sequence sheet was created. This reference sheet, based on the NM_000051.4 transcript, was derived from the coding sequence and included 5 nucleotides of adjacent intronic sequence. The sheet contained only the intended SNV and the additional synonymous variant, with no mismatches. Processed reads from exon-targeted deep sequencing were aligned to these SNV reference sequences, and read counts were recorded when the reads perfectly matched the SNV reference sequence. Reads from unedited wild-type cells were also recorded if they perfectly matched the reference transcript sequence.

To differentiate true prime-edited reads from errors introduced during sequencing or library preparation, we calculated the odds ratio (OR) and P-value using Fisher's exact test, comparing sequencing reads from Day 0 (D0) with those from unedited cells, as follows:

$$\mathrm{OR} = \frac{(SNV\ read\ count\ at\ D0\ +\ 1)/(Wild-type\ read\ count\ at\ D0\ +\ 1)}{(SNV\ read\ count\ in\ unedited\ cells\ +\ 1)/(Wild-type\ read\ count\ in\ unedited\ cells\ +\ 1)}$$

For each exon library experiment, true-edited reads were identified based on an OR of ≥ 3 and a false discovery rate (FDR) of < 0.05, with multiple testing correction performed using the Benjamini-Hochberg method.

2.17. Calculation of the function score

We calculated the $\log_2$-fold change (LFC) for each SNV by comparing allele frequencies at Day 10 (D10) with those at Day 0 (D0). Given that editing efficiency can vary depending on sequence context and positional biases, leading to variable editing from D0 through D10, we standardized the LFC of each SNV using the LFCs of synonymous SNVs, which were assumed to have a neutral LFC due to their lack of amino acid changes. The regressed LFC for synonymous SNVs at each position within an exon was obtained through LOWESS (Locally Weighted Scatterplot Smoothing) regression. LFC standardization involved subtracting the regressed LFC of synonymous SNVs at each position and dividing by the interquartile range of the synonymous SNV LFCs within each exon. This process allowed for the generation of function scores, enabling direct comparisons between exons.

After standardization, we calculated the weighted average of the standardized LFC for each

SNV across different co-occurring synonymous mutations (internal replicates), accounting for statistical confidence and sequencing read depth.

Weighted LFC =

$$\frac{\sum allele\ frequency\ at\ D0 \times LFC\ of\ each\ (SNV+additional\ synonymous\ mutation)}{\sum allele\ frequency\ of\ each\ (SNV+additional\ synonymous\ mutation)\ at\ D0}$$

The function score was calculated by averaging the weighted LFCs across biological replicates. To determine the function score for a specific amino acid substitution, we averaged the function scores of all SNVs that induce the same amino acid change.

Variants were classified into three functional categories: 'Non-functional', 'Intermediate' (both categories also referred to as depleting variants), and 'Functional' (variants with stable frequency). The cutoff values for each category were set at -1.360 (the 5th percentile of the function score for synonymous variants) and -0.912 (Youden's index used for classifying nonsense vs. synonymous variants).

To support real-world variant interpretation, we included a column in the final datasheet (https://github.com/Labmed-Lee/Lee_et_al) indicating the confidence level of an SNV's functional classification based on its frequency at D0. The confidence levels are defined as follows: high (SNV frequency at D0 $\geq$ 0.001%), medium-high (0.0001% – 0.001%), and medium (< 0.0001%). Variants with a medium confidence level may require additional context, such as family history or de novo status, for more accurate interpretation.

2.18. Visualization of protein structure

To visualize the function scores in relation to protein structure, we calculated the average function scores of missense SNVs at each residue and mapped these scores to the ATM protein structure (PDB: 8OXO, 7SID) using PyMol v2.5.5.

2.19. ClinVar database and population sequencing data analysis

ATM variant entries with at least a one-star rating in ClinVar [24] were downloaded on 9 April 2024. Variants classified as 'pathogenic/likely pathogenic' were labeled as 'likely pathogenic,' while those classified as 'benign/likely benign' were labeled as 'likely benign.' Variants not reported in ClinVar were labeled as 'uncertain significance.'

Tumor sequencing data from the AACR GENIE Cohort v16.0 [25] was obtained through cBioPortal on 21 September 2024. To calculate the odds ratio of variant occurrence in tumors, allele counts from gnomAD v.3.1.2 (non-cancer population data) were used as the control set [26].

2.20. Comparisons between computational predictions and function scores

Computational prediction scores (SIFT, REVEL, CADD, ESM1b, EVE, AlphaMissense) were obtained from dbNSFP v.4.8 [27, 28]. The BoostDM score was calculated using the 'Prostate Adenocarcinoma' model [29]. For score comparisons, only variants for which all metrics and function scores were available were included.

2.21. Survival analysis

Genomic and clinical data from patients with chronic lymphocytic leukemia and bladder cancer were obtained through cBioPortal [30, 31]. Participants were grouped according to the functional classifications of their *ATM* variants. Kaplan-Meier survival curves were generated using the R packages 'survminer' and 'survival'. Log-rank tests were conducted to compare survival curves for patients with different *ATM* carrier statuses.

2.22. UKB data analysis

Whole-exome sequencing data from 424,909 participants were stored as population-level VCF files aligned to GRCh38 and accessed via the UKB research analysis platform. We identified participants with variants in the *ATM* coding region, including 5 nucleotides adjacent to exon-intron junctions, excluding those with indel variants. Participants were categorized into three groups (non-functional, intermediate, and functional) based on the function score or eDA score. The intact *ATM* group was defined as the absence of *ATM* variants in this region.

To investigate the relationship between cancer susceptibility phenotypes and *ATM* variants, we performed Cox multivariate regression, adjusting for sex and baseline age, and generated Kaplan-Meier survival curves. Cancer diagnosis information was retrieved from cancer registry data. The time from enrollment to cancer diagnosis was simplified into years, and participants' ages at the time of diagnosis were used to analyze the lifelong risk of cancer incidence.

Phenotypic variables, with corresponding ICD10 codes, included: all cancers combined; breast cancer (C50); oropharyngeal (C00-C14); esophago-gastric (C14-C15); small intestine (C17); colorectal (C18-C20); anal (C21); pancreato-biliary (C24-C25); bronchopulmonary (C34); melanoma (C43); other malignancy of skin (C44); cervix (C53); utero-ovarian (C54-C56); prostate (C61); testicular (C62); uretero-renal (C64-C66); bladder (C67); brain (C71); thyroid (C73); Hodgkin's disease (C81); Non-Hodgkin lymphoma (C82-C85); plasma cell neoplasm (C90); lymphoid leukemia (C91); myeloid leukemia (C92); melanoma in situ (D03); carcinoma in situ, breast (D05); benign neoplasm (D10-D36).

For regression models incorporating computational scores (AlphaMissense, CADD, REVEL, and EVE) and our function or eDA scores, we adjusted for age at enrollment and sex. Given the different output ranges of the computational tools, we normalized the values to the range [0,1] using the 'rescale' function in R.

2.23. Deep learning dataset and feature engineering

To predict the functional effects of unevaluated SNVs, we developed DeepATM, a deep learning model trained on experimentally determined function scores of *ATM* variants from this study, along with mutation information and 16 scores derived from various tools, including SIFT [32], FATHMM [33], MutationTaster [34], LRT [35], DANN [36], PolyPhen-2 HVAR [37], PROVEAN [38], REVEL [39], CADD [40], phyloP100, GERP [41], ESM1b [42], EVE [43], AlphaMissense [44], BoostDM [29], and SpliceAI [45]. These features were used as additional inputs for model training and evaluation.

The test dataset consisted of all missense variants classified as pathogenic, likely pathogenic,

benign, or likely benign with a one-star rating or higher in ClinVar (n = 116), irrespective of whether they had been experimentally evaluated. The training dataset was constructed by excluding all evaluated variants that shared amino acid positions with the test set. The remainder of the training data included evaluated missense (n = 16,275), synonymous (n = 4,395), and nonsense variants (n = 1,183). Mutations at stop codon positions were excluded.

The pathogenicity target variable was transformed using an arcsinh transformation ($y = \sinh^{-1}(\text{function score} + 0.912)/2$) to reduce skewness in the function score distribution. This transformation was applied to improve model stability and predictive performance during training.

2.24. Model architecture

DeepATM, a Transformer-based regression model, was composed of the following components:
- Amino acid embedding: The amino acid sequence encoded by the *ATM* gene was represented using a 64-dimensional embedding vector for each amino acid. The embeddings were initialized randomly, and the model processed the sequence in a continuous vector space.
- Domain embedding: An additional embedding layer encoded domain annotations for each amino acid in the sequence. Domains and their positions were annotated as shown below:

Domain	Start position (a.a)	End position (a.a)
TAN	1	166
FAT	1940	2566
PI3/4 Kinase	2686	2998
FATC	3024	3056

- Coordinate embedding: To incorporate structural information of the ATM protein, a multi-layer perceptron (MLP) processed the 3D coordinates of the alpha-carbon atoms. Coordinates were obtained from AlphaFold 3 to avoid gaps caused by missing residues in experimentally determined structures [46]. The output from the MLP was integrated with the amino acid and domain embeddings.
- Transformer encoder: The combined embeddings were passed through two Transformer encoder layers, each consisting of 8 attention heads. This structure was designed to capture long-range interactions and dependencies between amino acids in the sequence.
- Fully connected layers: The Transformer output at the mutation location was concatenated with the 16 precomputed scores to provide the model with additional functional information. This concatenated vector was passed through a fully connected network with 128 hidden units and ReLU activation, followed by a single output neuron to predict the pathogenicity score.

2.25. Model training

The model was trained using the AdamW optimizer with an initial learning rate of 1e-3 and a weight decay of 10^{-2} to prevent overfitting. A cosine annealing schedule with periodic restarts was employed to adjust the learning rate dynamically, with the initial cycle length set to 10 epochs. After each restart, the learning rate was reduced by 20%, and the cycle length was doubled, helping the model escape local minima. The model was trained for a maximum of 150 epochs, with early stopping triggered if the validation loss did not improve for 20 consecutive epochs.

Training data were dynamically sampled in each batch with a batch size of 20, consisting of 90% missense variants, 5% synonymous variants, and 5% nonsense variants. The mean squared

error was used as the primary loss function. Automatic mixed precision was implemented to accelerate training and reduce memory usage by utilizing both 16-bit and 32-bit precision. Gradient clipping was applied to stabilize training by limiting the magnitude of gradients and preventing gradient explosion.

2.26. Performance evaluation

The model's performance was evaluated using 5-fold cross-validation. In each evaluation, the training and validation sets were randomly split. The predictions made by the model were compared to the actual function scores, and performance was assessed using Spearman's correlation and Pearson's correlation.

During training, model checkpoints were saved whenever the validation loss improved. The best-performing models from all five cross-validations were ensembled to provide final predictions for the unevaluated variants.

The ensemble model's ability to classify variants was assessed by calculating the area under the receiver operating characteristic curve (auROC). This analysis was performed on two groups: the first consisting of variants with a ClinVar one-star or higher status (n = 116), and the second group with at least a two-star ClinVar status (n = 68). DeepATM's auROC was compared against other pathogenicity prediction tools, including AlphaMissense, ESM1b, phyloP, and PROVEAN. Performance was evaluated based on 1,000 bootstrap resampling of the test set.

2.27. Predicting the effects of unevaluated *ATM* variants

DeepATM was used to predict the effects of 4,421 unevaluated SNVs in the *ATM* gene. To generate eDA scores, raw prediction values for 23,092 SNVs were aligned to their function scores using a rank-based approach. The relationship between the eDA scores and the function scores was modeled using generalized additive regression. eDA scores for the 4,421 unevaluated SNVs were then derived from this model. Based on the eDA scores and predefined function score cutoffs, the predicted SNVs were classified into pathogenic, intermediate, or benign categories.

2.28. Statistical analysis

Basic statistical analysis was performed in R (v4.2.1) using RStudio. All tests were two-sided. Exact P-values were calculated using the 'pnorm' function in R, and multiple testing correction was applied using the 'p.adjust' function.

Table 1. Sequences of primers used for screening *ATM*-haploid cells

Name (referred to as)	Direction	Sequence
ATM_3' downstream_RP (RP3)	reverse	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGATGCAGCATTATCAGACTG
ATM_3' downstream_FP (FP3)	forward	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAATGTATTACTTTACTGTTACCTG
ATM_c3380_NGS_L1 (RP2)	reverse	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCAAGCATATGATAACAGCAAA
ATM_c3380_NGS_U1 (FP2)	forward	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGAAAGCACTTCCTTTGAAGC
ATM_5' upstream_RP (RP1)	reverse	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACTGCCCCAAAACATTCGG
ATM_5' upstream_FP (FP1)	forward	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAATCGCTTCCGCCAGAG

Table 2. Sequences of primers used for molecular cloning

Name (referred to as)	Direction	Sequence
Oligo_Amp_RP12	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGAGCCTCGTCGGCATACGGT
Oligo_Amp_RP11	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGCGATCCATCGCGCGCTCTT
Oligo_Amp_RP10	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGTATTGTGCAGGCACGCCCG
Oligo_Amp_RP9	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGTGCGTGGCGCAATAGGTGG
Oligo_Amp_RP8	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGTGGCGAGCTCAATGTGCCG
Oligo_Amp_RP7	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGTGAACTGCGCGCCTAGCAC
Oligo_Amp_RP6	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGTAGGCGCATGCCCTCGTCT
Oligo_Amp_RP5	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGTCCACTTATGCCGCGCCAC
Oligo_Amp_RP4	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGACACTACCGCGTGCAGTGC
Oligo_Amp_RP3	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGACCCACCATGCGCGAACAG
Oligo_Amp_RP2	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGCGGCAGTAACGCCCTTGCA
Oligo_Amp_RP1	reverse	GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGTTGCGGGCCTGCACTAGGA
Oligo_Amp_FP1	forward	TTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACC

Table 3. Sequences of primers used for PCR amplification of endogenous sites

Name	Binding	Combined (NGS adaptor + binding)
ATM_ex_02_NGS_F1	CACCTCTTTCTCTCTATATATGC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACACCTCTTTCTCTCTATATATGC
ATM_ex_02_NGS_R1	GGGTTACTAATCACACTTATTTC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGTTACTAATCACACTTATTTC
ATM_ex_03_NGS_F1	GAAAATAAGTGTGATTAGTAACCC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGAAAATAAGTGTGATTAGTAACCC
ATM_ex_03_NGS_R1	GAAGCAAAGATAAATGTTAAGAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAAGCAAAGATAAATGTTAAGAC
ATM_ex_04_NGS_F1	AAGTATTCAACGAGTTTCTGAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAAGTATTCAACGAGTTTCTGAA
ATM_ex_04_NGS_R1	AAACTCACGCGACAGTAATC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAACTCACGCGACAGTAATC
ATM_ex_05_NGS_F1	CCAAGTGTCTTATTTTTGTTCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACCAAGTGTCTTATTTTTGTTCA
ATM_ex_05_NGS_R1	GTGAAGTTTCATTTCATGAGGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGAAGTTTCATTTCATGAGGA
ATM_ex_06_NGS_F1	GTGCAGTTTTAAAATCCTTTTTC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTGCAGTTTTAAAATCCTTTTTC
ATM_ex_06_NGS_R1	CTGAGTCTAAAACATGGTCTTG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTGAGTCTAAAACATGGTCTTG
ATM_ex_07_NGS_F1	GTTATACCCAGTTGAGCTTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTATACCCAGTTGAGCTTG
ATM_ex_07_NGS_R1	CTTCTATGTTTGAATGAAGAAGC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTCTATGTTTGAATGAAGAAGC
ATM_ex_08_NGS_F1	GGAGCTAGCAGTGTAAACAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGAGCTAGCAGTGTAAACAG
ATM_ex_08_NGS_R1	AACAGGAAATTTCTAAATGTGAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAACAGGAAATTTCTAAATGTGAC
ATM_ex_09_NGS_F1	AACAACAGCGAAACTCTGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAACAACAGCGAAACTCTGG
ATM_ex_09_NGS_R1	CAAGAGATTAAAATGACACTGAA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAAGAGATTAAAATGACACTGAA
ATM_ex_10_NGS_F1	CCTTTAGTTTGTTAATGTGATGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACCTTTTAGTTTGTTAATGTGATGG
ATM_ex_10_NGS_R1	CTGTGTGTGTTTATCTGTAAGTC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTGTGTGTGTTTATCTGTAAGTC
ATM_ex_11_NGS_F1	GTCTTTGCCCCTCCAATAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTCTTTGCCCCTCCAATAG
ATM_ex_11_NGS_R1	AATAAGTGGAGAGAGCCTGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAATAAGTGGAGAGAGCCTGA
ATM_ex_12_NGS_F1	AGAAGTCAAGATTTATAGCTAAAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGAAGTCAAGATTTATAGCTAAAC
ATM_ex_12_NGS_R1	CCCAGCTAAAATTATCATCTTTG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCCAGCTAAAATTATCATCTTTG
ATM_ex_13_NGS_F1	GCTAATACATATAAGGCAAAGC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCTAATACATATAAGGCAAAGC
ATM_ex_13_NGS_R1	CCTAACAGTTTACCAAAGTTGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTAACAGTTTACCAAAGTTGA
ATM_ex_14_NGS_F1	ATGTATGTAGAATTTGTTCTTACA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATGTATGTAGAATTTGTTCTTACA
ATM_ex_14_NGS_R1	CATTCAAATTTATCCGAAACTTTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCATTCAAATTTATCCGAAACTTTA
ATM_ex_15_NGS_F2	GTCCAAGATCAAAGTACACTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTCCAAGATCAAAGTACACTG
ATM_ex_15_NGS_R2	GTGACAGAGAAAGATCCTATC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGACAGAGAAAGATCCTATC

Name	Binding	Combined (NGS adaptor + binding)
ATM_ex_16_NGS_F1	AGAAAACACTGTCTGCCAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGAAAACACTGTCTGCCAA
ATM_ex_16_NGS_R1	GCTATATGTTGTGAGATGCATC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTATATGTTGTGAGATGCATC
ATM_ex_17_NGS_F1	AAGCCATCTTGAACATCTTTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAAGCCATCTTGAACATCTTTG
ATM_ex_17_NGS_R1	GCCTCTTATACTGCCAAATCA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCTCTTATACTGCCAAATCA
ATM_ex_18_NGS_F1	GCCCTTCTCTTAGTGTTAATG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCCCTTCTCTTAGTGTTAATG
ATM_ex_18_NGS_R1	TCAGATAAAATCCAAGAGCTTC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCAGATAAAATCCAAGAGCTTC
ATM_ex_19_NGS_F1	AATGATTTGTGGATAAACCTGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAATGATTTGTGGATAAACCTGA
ATM_ex_19_NGS_R1	CAACTTTATAAGCTTAACAGAACA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAACTTTATAAGCTTAACAGAACA
ATM_ex_20_NGS_F1	TGTTCTGTTAAGCTTATAAAGTTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATGTTCTGTTAAGCTTATAAAGTTG
ATM_ex_20_NGS_R1	GATACAAAACTTGCATTCGTATC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGATACAAAACTTGCATTCGTATC
ATM_ex_21_NGS_F2	ACTTACAATAACCTTTCAGTGAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACTTACAATAACCTTTCAGTGAG
ATM_ex_21_NGS_R2	CTGTGGTTAAATATGAAATAGAGAA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTGTGGTTAAATATGAAATAGAGAA
ATM_ex_22_NGS_F1	GCAGTCTTTGTTTGTTAATGAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCAGTCTTTGTTTGTTAATGAG
ATM_ex_22_NGS_R1	TGTAAGACATTCTACTGCCAT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGTAAGACATTCTACTGCCAT
ATM_ex_23_NGS_F1	GTTCTGGAATATGCTTTGGAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTCTGGAATATGCTTTGGAA
ATM_ex_23_NGS_R1	AGCAAGCATATGATAACAGC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCAAGCATATGATAACAGC
ATM_ex_24_NGS_F1	GGGATTTTATAATTGATTGTTAAAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGGATTTTATAATTGATTGTTAAAC
ATM_ex_24_NGS_R1	CTAAGGAAGCTTCTAATAAAATAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAAGGAAGCTTCTAATAAAATAC
ATM_ex_25_NGS_F1	TTCATTTTTCTTAACACATTGAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATTCATTTTTCTTAACACATTGAC
ATM_ex_25_NGS_R1	GGGACTTGCTAAGTATTGTTAAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGACTTGCTAAGTATTGTTAAC
ATM_ex_26_NGS_F1	GTATGATACTTTAATGCTGATGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTATGATACTTTAATGCTGATGG
ATM_ex_26_NGS_R1	GGTTATATCTCATATCATTCAGGG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGTTATATCTCATATCATTCAGGG
ATM_ex_27_NGS_F1	GAGCTGTCTTGACGTTCAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGAGCTGTCTTGACGTTCAC
ATM_ex_27_NGS_R1	AATTGAAATAGACATTGAAGGTG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAATTGAAATAGACATTGAAGGTG
ATM_ex_28_NGS_F1	CATTTTGGAAGTTCACTGGTC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATTTTGGAAGTTCACTGGTC
ATM_ex_28_NGS_R1	TTAGCTAAAAAAGAAGGAATGTTC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTAGCTAAAAAAGAAGGAATGTTC
ATM_ex_29_NGS_F1	GCCGAGTATCTAATTAAACAAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCCGAGTATCTAATTAAACAAG
ATM_ex_29_NGS_R1	AAGACTGCTTATATATTGGTCT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGACTGCTTATATATTGGTCT

Name	Binding	Combined (NGS adaptor + binding)
ATM_ex_30_NGS_F1	AACTTACTGGTTGTTGTTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAACTTACTGGTTGTTGTTG
ATM_ex_30_NGS_R1	CAAATCCTTCTAACAATACTTTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAAATCCTTCTAACAATACTTTA
ATM_ex_31_NGS_F1	GGCTTACTTTAAAATTATTTCTCTC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGCTTACTTTAAAATTATTTCTCTC
ATM_ex_31_NGS_R1	TTGAAAAGTACTACTATGTTCTCTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTGAAAAGTACTACTATGTTCTCTA
ATM_ex_32_NGS_F1	AACCAATACGTGTTAAAAGC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAACCAATACGTGTTAAAAGC
ATM_ex_32_NGS_R1	CAGGTAGAAATAGCCCATG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAGGTAGAAATAGCCCATG
ATM_ex_33_NGS_F1	GTGTTGTCTTCATGCTAGTTTA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTGTTGTCTTCATGCTAGTTTA
ATM_ex_33_NGS_R1	CTATATGTGATCCGCAGTTG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTATATGTGATCCGCAGTTG
ATM_ex_34_NGS_F1	ATGATCTCTTACCTATGACTCTA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATGATCTCTTACCTATGACTCTA
ATM_ex_34_NGS_R1	CTCCATGAATGTCATATTGAGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTCCATGAATGTCATATTGAGA
ATM_ex_35_NGS_F1	GTGGAGGTTAACATTCATCAAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTGGAGGTTAACATTCATCAAG
ATM_ex_35_NGS_R1	GACCCACAGCAAACAGAAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGACCCACAGCAAACAGAAC
ATM_ex_36_NGS_F1	GGTACAATGATTTCCACTTCTC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGTACAATGATTTCCACTTCTC
ATM_ex_36_NGS_R1	CAGGTCATAAACAAGGAATTATATC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAGGTCATAAACAAGGAATTATATC
ATM_ex_37_NGS_F1	ACTCATTTTTACTCAAACTATTGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACTCATTTTTACTCAAACTATTGG
ATM_ex_37_NGS_R1	CTTTCTCTAGAACTGAGTTTACA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTTCTCTAGAACTGAGTTTACA
ATM_ex_38_NGS_F2	GGAAGAAGGTGTGTAAGCAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGAAGAAGGTGTGTAAGCAA
ATM_ex_38_NGS_R1	CAGCCGATAGTTAACAAGTTAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAGCCGATAGTTAACAAGTTAC
ATM_ex_39_NGS_F1	ACATGCTTTTATTTTGATATTGAAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACATGCTTTTATTTTGATATTGAAG
ATM_ex_39_NGS_R1	CCTTATTGAGACAATGCCAAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCTTATTGAGACAATGCCAAC
ATM_ex_40_NGS_F1	GAGCTTCCAAATAGTATGTTCTC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGAGCTTCCAAATAGTATGTTCTC
ATM_ex_40_NGS_R1	GCATCTGTACAGTGTCTATAAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCATCTGTACAGTGTCTATAAC
ATM_ex_41_NGS_F1	AGAGTTGGGAGTTACATATTGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGAGTTGGGAGTTACATATTGG
ATM_ex_41_NGS_R1	ACACATAACTCCTTCATAAACAG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACACATAACTCCTTCATAAACAG
ATM_ex_42_NGS_F1	CTGTTTATGAAGGAGTTATGTGT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGTTTATGAAGGAGTTATGTGT
ATM_ex_42_NGS_R1	GGCTGTGTAAATATCCACCAA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGCTGTGTAAATATCCACCAA
ATM_ex_43_NGS_F1	CTGGTTTTCTGTTGATATCTTTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGGTTTTCTGTTGATATCTTTG
ATM_ex_43_NGS_R1	GAATGAGGAGAGAGGCAAAA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAATGAGGAGAGAGGCAAAA

Name	Binding	Combined (NGS adaptor + binding)
ATM_ex_44_NGS_	ATACATGTATATCTTAGGGTTCTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATACATGTATATCTTAGGGTTCTG
ATM_ex_44_NGS_	CTTCATCAATGCAAATCCTTAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTCATCAATGCAAATCCTTAC
ATM_ex_45_NGS_	GCAAAGCCTATGATGAGAAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCAAAGCCTATGATGAGAAC
ATM_ex_45_NGS_	GCTGCACTTTAGGATAACAA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTGCACTTTAGGATAACAA
ATM_ex_46_NGS_	CATTTCTCTTGCTTACATGAAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACATTTCTCTTGCTTACATGAAC
ATM_ex_46_NGS_	AGGAAAGTCAAGAGGTAAGATG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGAAAGTCAAGAGGTAAGATG
ATM_ex_47_NGS_	ATGGTAGTAGTATCAGTAGTAAAAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATGGTAGTAGTATCAGTAGTAAAAG
ATM_ex_47_NGS_	CAGTAAAACACTAATCCAGCC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAGTAAAACACTAATCCAGCC
ATM_ex_48_NGS_	GTTGGGTACAGTCATGGTAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTGGGTACAGTCATGGTAA
ATM_ex_48_NGS_	GCTTTGGAAATATATTGATCTTGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTTTGGAAATATATTGATCTTGA
ATM_ex_49_NGS_	CCTTAATTTGAGTGATTCTTTAGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACCTTAATTTGAGTGATTCTTTAGA
ATM_ex_49_NGS_	GCCGACCTTTAGAGCTCAA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCGACCTTTAGAGCTCAA
ATM_ex_50_NGS_	GTTCATGGCTTTTGTGTTTTAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTCATGGCTTTTGTGTTTTAC
ATM_ex_50_NGS_	CACAGGGTAGAATATTGGGC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCACAGGGTAGAATATTGGGC
ATM_ex_51_NGS_	GCTTAGATGTGAGAATATTTGAAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCTTAGATGTGAGAATATTTGAAA
ATM_ex_51_NGS_	GTATTTCCATTTCTTAGAGGGA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTATTTCCATTTCTTAGAGGGA
ATM_ex_52_NGS_	GTTAAGCAAAATGAAAAATATGGATTA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTAAGCAAAATGAAAAATATGGATTA
ATM_ex_52_NGS_	AAAGACTGAATATCACACTTCTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAAGACTGAATATCACACTTCTA
ATM_ex_53_NGS_	CTCTGAGAAGTTTAAATGTTGGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTCTGAGAAGTTTAAATGTTGGG
ATM_ex_53_NGS_	CTACAGAGAGTAACACAGCAAG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTACAGAGAGTAACACAGCAAG
ATM_ex_54_NGS_	GACCTTCAATGCTGTTCCTC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGACCTTCAATGCTGTTCCTC
ATM_ex_54_NGS_	GGTTGAAACATATGAAATTTGCC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGTTGAAACATATGAAATTTGCC
ATM_ex_55_NGS_	GTGCAAATAGTGTATCTGACCTA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTGCAAATAGTGTATCTGACCTA
ATM_ex_55_NGS_	TTCATCACTAAAACTCTAAGGGCT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTCATCACTAAAACTCTAAGGGCT
ATM_ex_56_NGS_	AACTGTACTTGTTTATTCATGCT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAACTGTACTTGTTTATTCATGCT
ATM_ex_56_NGS_	CCCAACCAAATGGCATCTTTTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCCAACCAAATGGCATCTTTTA
ATM_ex_57_NGS_	ATCAAATGCTCTTTAATGGCC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATCAAATGCTCTTTAATGGCC
ATM_ex_57_NGS_	AGCCATTAAATAATCTTACAATAACC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCCATTAAATAATCTTACAATAACC

Name	Binding	Combined (NGS adaptor + binding)
ATM_ex_58_NGS_F1	GTGTATATTAGTTTAATTGAACACAA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTGTATATTAGTTTAATTGAACACAA
ATM_ex_58_NGS_R1	AAACAACAAAGTGCTCAATCTAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAACAACAAAGTGCTCAATCTAC
ATM_ex_59_NGS_F1	ACTTAAAGATTATACCAAGTCAGTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAACTTAAAGATTATACCAAGTCAGTG
ATM_ex_59_NGS_R1	GTAGGCAAACAACATTCCATG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTAGGCAAACAACATTCCATG
ATM_ex_60_NGS_F1	GTAAATTAGCTGTCAAACCTCC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTAAATTAGCTGTCAAACCTCC
ATM_ex_60_NGS_R1	GCCCAGCCCATGTAATTTTG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCCAGCCCATGTAATTTTG
ATM_ex_61_NGS_F1	GCTCAGCATACTACACATGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCTCAGCATACTACACATGA
ATM_ex_61_NGS_R1	GTGACTTCCTGATGAGATACAC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGACTTCCTGATGAGATACAC
ATM_ex_62_NGS_F1	GGTTCTACTGTTTCTAAGTATGTG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGTTCTACTGTTTCTAAGTATGTG
ATM_ex_62_NGS_R1	GTGAACAGTTTAAAGGCCTTG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGAACAGTTTAAAGGCCTTG
ATM_ex_63_NGS_F1	CAAGGCCTTTAAACTGTTCAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACAAGGCCTTTAAACTGTTCAC
ATM_ex_63_NGS_R1	TTCTAAAGGCTGAATGAAAGGG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTCTAAAGGCTGAATGAAAGGG
BRCA1_ex_4_NGS_F1	GGCTCTTAAGGGCAGTTGTGAG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGCTCTTAAGGGCAGTTGTGAG
BRCA1_ex_4_NGS_R1	CTTTTTCCTACTGTGGTTGCTTCC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTTTTTCCTACTGTGGTTGCTTCC
BRCA1_ex_19_NGS_F1	CTGCTCCACTTCCATTGAAGGA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGCTCCACTTCCATTGAAGGA
BRCA1_ex_19_NGS_R1	GTGGAATACAGAGTGGTGGGG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGGAATACAGAGTGGTGGGG

Table 4. Sequences of primers used for off-target analysis

Name	NGS_Adaptor	Binding	Combined
K331E_off_1_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTA	TTCCTCAAATGATTCAGAATTTC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATTCCTCAAATGATTCAGAATTTC
K331E_off_1_R1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	ATGACATAATATAGCACCTAGCA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATGACATAATATAGCACCTAGCA
K331E_off_2_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTA	AGATGAAGGTGAGGCTGACA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGATGAAGGTGAGGCTGACA
K331E_off_2_R1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	GTGTGTTTCGGGGAATGG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGTGTTTCGGGGAATGG
K331E_off_3_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTA	AGTTTTGGATTAACTTGAATACATT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGTTTTGGATTAACTTGAATACATT
K331E_off_3_R1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	TTGCAATGGACAGATATGTACTT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTGCAATGGACAGATATGTACTT
K331E_off_4_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTA	TACATGCTAAGTCCCTCAAGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATACATGCTAAGTCCCTCAAGG
K331E_off_4_R1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	TTTTCCTCAAGTGAACAAATACATG	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTCCTCAAGTGAACAAATACATG
L969P_off_1_F1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTA	ATAGGAGAGCACTTTGGGTT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAATAGGAGAGCACTTTGGGTT
L969P_off_1_R1	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	CAAACAAAGCCTGATGAGATAAT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCAAACAAAGCCTGATGAGATAAT

3. RESULTS

3.1. Cell line generation for the functional evaluation of *ATM* variants

The *ATM* gene consists of 63 exons, with its coding region extending from exon 2 to exon 63. It encodes a full-length protein comprising 3,056 amino acids, which includes three recognized functional domains (**Figure 1A**). Importantly, both pathogenic variants and variants of uncertain significance (VUSs) are dispersed across the entire coding sequence (**Figure 1B**). The absence of distinct hotspot regions makes it more challenging to assess the functional impact of these variants.

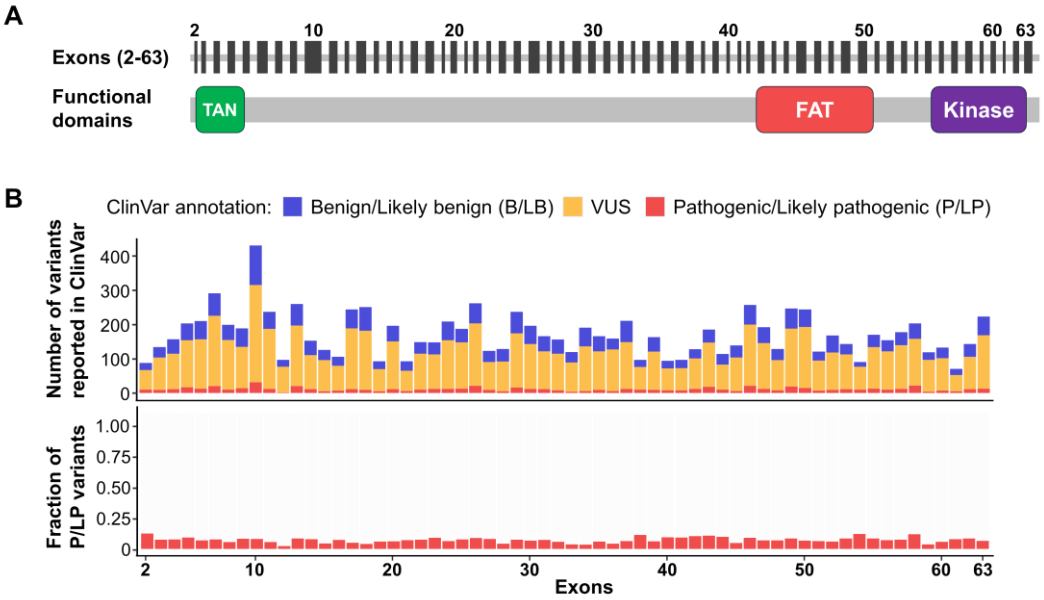


Figure 1. The structure of the *ATM* gene and the distribution of variants. (A) The structure of the *ATM* gene. Exons 2-63 encode the 3,056 amino acid-long *ATM* protein. Gray boxes represent exons; numbers indicating the exon positions are intermittently shown above the boxes. Three functional regions, which include TAN (Tel1/*ATM* N-terminal motif), the FAT (FRAP-*ATM*-TRRAP) domain, and the kinase domain, are shown. **(B)** Variants in the coding exons in *ATM*. Numbers indicating exon positions are shown on the x-axis. The number of variants (top) and the fraction of pathogenic or likely pathogenic (P/LP) variants among the total number of variants reported in ClinVar (bottom) are shown.

To systematically analyze all possible SNVs in *ATM* using a high-throughput approach, we aimed to utilize a diploid cell line, as it provides a more physiologically relevant model compared to nearly haploid or triploid cells, such as HAP1 or HEK293T cells, respectively. For this reason,

we selected the HCT116 cell line based on several criteria: (i) it is a nearly diploid cancer cell line, (ii) it harbors wild-type *ATM* along with at least one functional copy of *BRCA1*, *BRCA2*, and *TP53*, (iii) it demonstrates relatively high prime editing efficiency, and (iv) its proliferation or survival is compromised by *ATM* loss, particularly in the presence of PARP inhibitors [47-49]. Additionally, HCT116 cells lack functional *MLH1* [50], which is anticipated to enhance prime editing efficiency [51].

Whole-exome sequencing of HCT116 cells confirmed the presence of two *ATM* copies: one wild-type and the other carrying an SNV (c.3380C>T) (**Figure 2**). Since *BRCA2*-haploid cells facilitated the more sensitive detection of hypomorphic variants in *BRCA2* screening compared to *BRCA2*-diploid cells [52], we applied a similar approach for *ATM* variant screening. We established an *ATM*-haploid HCT116 clone by deleting the entire *ATM* gene copy (~146,000 bp) containing the c.3380C>T SNV using SpCas9 and two single-guide RNAs (sgRNAs) (**Figure 3A**)

To delete the entire copy of the *ATM* gene (~146,000 bp) containing the SNV (c.3380C>T), we transfected plasmids encoding SpCas9 and two single-guide RNAs (sgRNAs) targeting the regions ~30 bp upstream of the 5' transcriptional start site and ~80 bp downstream of the 3' transcriptional end site (**Figure 3A**). The transfected cell pool was sorted using flow cytometry into single cells, which were then expanded in culture (**Figure 3B**). Agarose gel electrophoresis and Sanger sequencing of PCR amplicons from these 200 single-cell-derived clones for the new junction sequence revealed that 14 clones (7%) had both an *ATM* gene-containing allele and an allele with the intended large deletion (representative images for five or four clones with the large deletion are shown in **Figures 4A and 4B**, respectively), suggesting they were *ATM*-haploid cells. Using PCR amplification of the region containing c.3380 and subsequent Sanger sequencing, we identified a clone that had both a large deletion that removed c.3380C>T (clone 4 in **Figure 4A**) and the wild-type *ATM* gene with only a single base pair insertion downstream of the 3' UTR (**Figure 4A and 4B**), suggesting this clone contained a single copy of intact *ATM*. Thus, we chose this *ATM*-haploid clone for subsequent studies.

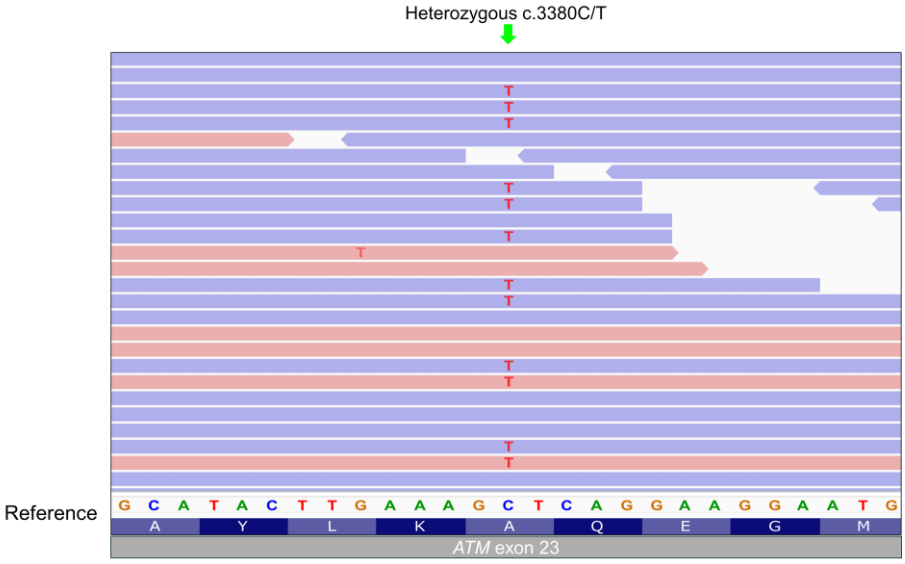


Figure 2. Integrative genomics viewer image of whole exome sequencing results from HCT116 cells. HCT116 cells contain the c.3380C>T mutation in one of the two *ATM* alleles.

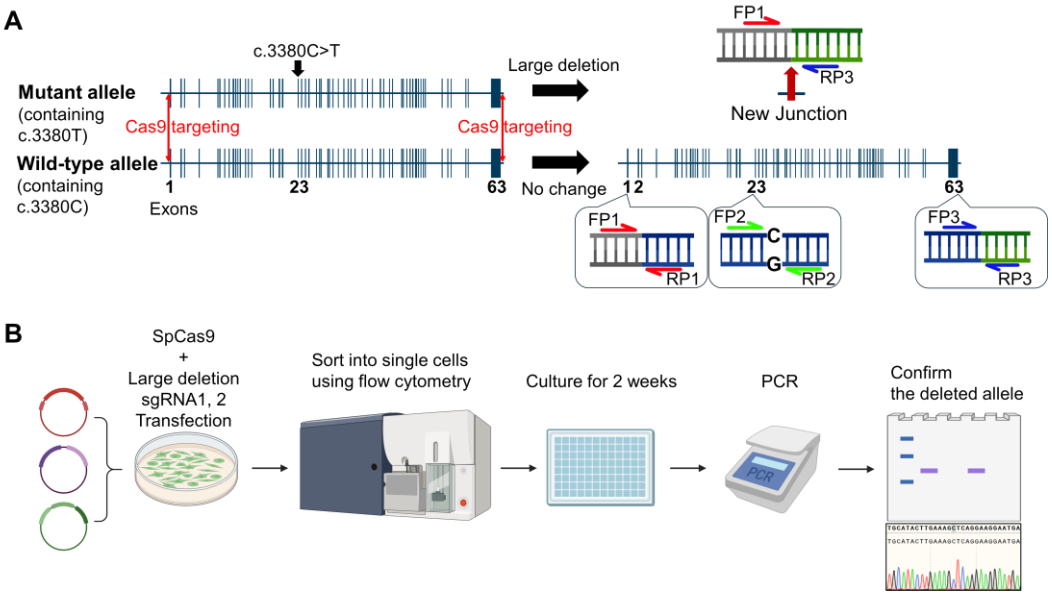


Figure 3. Cell line generation strategy for *ATM*-haploid HCT116 cells. (A) Haploidization of the *ATM*-coding region in HCT116 cells. A large deletion in the *ATM* allele containing c.3380C>T was induced using Cas9 and two sgRNAs. PCR primers used for the analyses

are shown (FP, forward primer; RP, reverse primer). **(B)** Cells transfected with plasmids encoding SpCas9 and two sgRNAs were sorted into single cells using flow cytometry. PCR was performed using the lysates of each single cell-derived clone. Positive clones identified using gel electrophoresis were further validated using Sanger and deep sequencing.

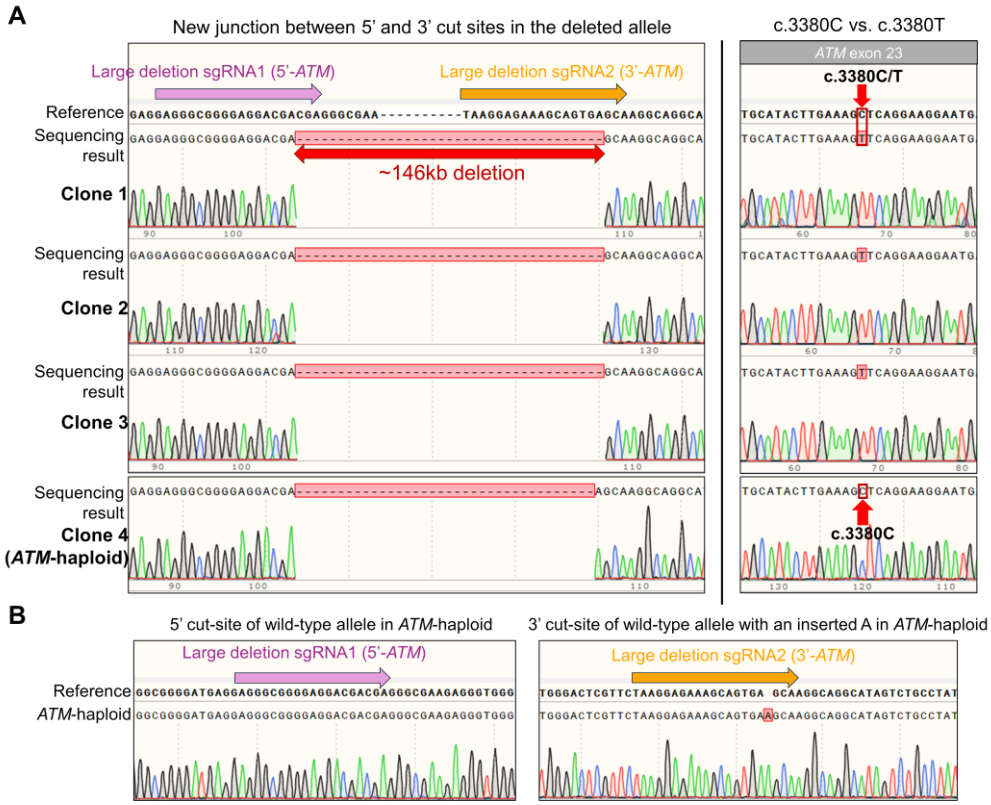


Figure 4. Confirmation of haploidization of *ATM* for *ATM*-haploid HCT116 cells. (A) Sanger sequencing results from four representative clones (clones 1, 2, 3, and the *ATM*-haploid clone, clone 4). The protospacer sequences that bind to the sgRNA guide sequences are shown. The deleted region is shown. **(B)** Sanger sequencing results from the two sgRNA binding sites in the *ATM*-haploid cells (clone 4).

3.2. *ATM* haploidization can increase the accuracy of variant evaluation

Previous studies have demonstrated that cells lacking *ATM* exhibit reduced proliferation and survival compared to *ATM*-proficient cells [49, 53, 54]. To verify that *ATM*-deficient cells become depleted when cultured alongside *ATM*-intact cells, we generated *ATM*-deficient cells using the *ATM*-haploid cells (clone 4, **Figure 4A**). By employing Cas9 nucleases, we introduced a frameshift mutation (c.3383dup) in *ATM*, resulting in the creation of an *ATM*-haploid-knockout (KO) clone (**Figure 5A**).

Co-culturing these *ATM*-haploid-KO cells with *ATM*-haploid cells revealed a decline in the relative fraction of *ATM*-haploid-KO cells over time (**Figure 6**), indicating that *ATM*-deficient cells are selectively depleted in the presence of *ATM*-proficient cells.

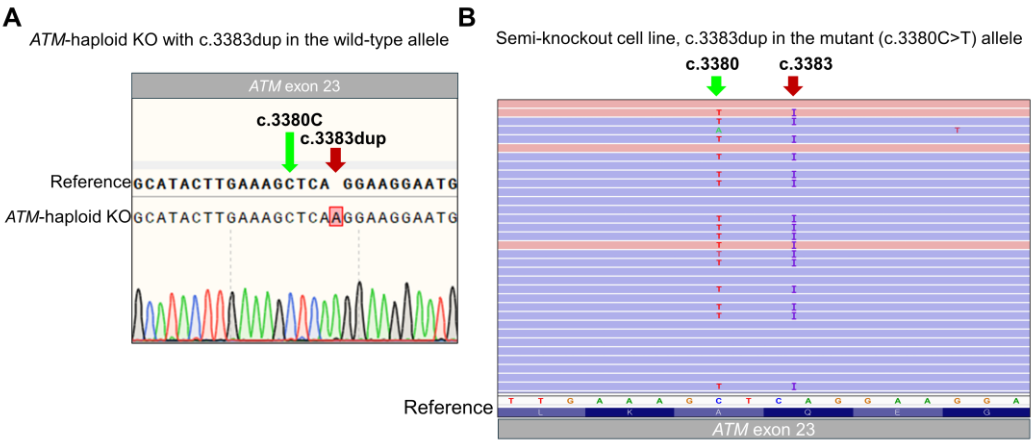


Figure 5. Sequence confirmation for *ATM*-haploid-KO and semi-KO HCT116 cells. **(A)** Sanger sequencing results showing the c.3383dup mutation in the *ATM*-haploid-KO cells. **(B)** Integrative genomics viewer image of sequencing reads from *ATM*-semi-KO cells. The c.3380 and c.3383 sites are indicated by the green and red arrows, respectively. This image reveals that the *ATM*-semi-KO cells have the c.3380C>T and c.3383dup mutations in a single allele in cis, whereas the other allele has neither of the mutations. I, insertion

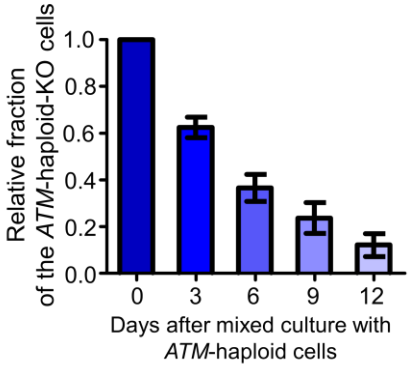


Figure 6. Relative fraction of *ATM*-haploid-KO cells after growth in a mixed culture with *ATM*-haploid cells. The *ATM*-haploid-KO cells contain a single copy of *ATM*, which has a frameshift mutation (c.3383dup). Error bars indicate standard errors. The number of independent culture n = 4.

Given this depletion of *ATM*-deficient cells in competition with *ATM*-intact cells, we hypothesized that *ATM* variant functionality could be assessed using cells containing one wild-type *ATM* allele and another with a KO mutation. To generate these semi-KO cells, we created a clone with the frameshift variant (c.3383dup) on the allele carrying the existing SNV (c.3380C>T), referred to as the semi-KO clone, using Cas9 and a single-guide RNA (sgRNA) targeting this site (Methods; **Figure 5B**).

To introduce a comprehensive set of SNVs into *ATM* semi-KO and *ATM*-haploid cells, we selected exons 55 and 56 as representative regions. We aimed to generate all possible SNVs within these exons and the adjacent intron regions (within 5 bp of exon boundaries) by constructing two engineered prime editing guide RNA (epegRNA) libraries—one per exon [55]. Using DeepPrime-FT, a deep-learning model optimized for predicting pegRNA efficiency [56], we designed a total of 2,396 epegRNAs: 1,350 for exon 55 (151 bp × 3 SNV/bp × 2-3 epegRNAs/SNV) and 1,046 for exon 56 (127 bp × 3 SNV/bp × 2-3 epegRNAs/SNV).

For accurate functional assessments, we directly sequenced the prime-edited regions using PEER-seq (Prime Editing and Endogenous Region sequencing) rather than relying solely on epegRNA abundance-based analysis [22]. Each epegRNA was designed to introduce one synonymous mutation in addition to the intended SNV, ensuring precise identification of the target mutation in sequencing reads [22, 57-60]. Synonymous mutations were intentionally placed outside exon-intron junctions to avoid disrupting splicing. Additionally, we incorporated internal replicates for the same intended edit by designing epegRNAs with varying synonymous edits.

Each of the two libraries was introduced into PE2max-expressing *ATM*-haploid and *ATM* semi-KO cells via lentiviral delivery at day-13 (D-13), and cells were cultured for 13 days to facilitate prime editing (**Figure 7**). The prime-edited cells were then divided and maintained under two conditions: DMSO (control) vs. olaparib for 10 days. Olaparib promotes depletion of *ATM*-deficient cells [49, 53, 54]. Deep sequencing was used to determine SNV frequencies, and the $\log_2$-fold change (LFC) of each SNV frequency at day 10 (D10) relative to day 0 (D0) was standardized based on synonymous SNVs (Methods). These standardized LFCs (sLFCs) were used to assess SNV functional effects.

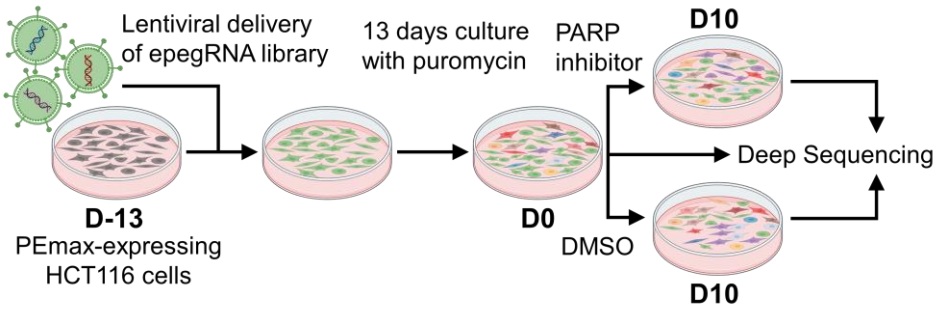


Figure 7. High-throughput functional evaluation of *ATM* variants. *ATM* variant-containing cells were generated by transducing epegRNA libraries into PEmax-expressing HCT116 cells. After 13 days of prime editing, the cell libraries were treated with a PARP inhibitor (olaparib) or solvent control (DMSO) for 10 days. Frequencies of variant-containing cells at day 0 (D0, 13 days after the transduction of epegRNAs) and day 10 (D10) were determined using deep sequencing.

To validate this approach, we compared sLFCs of 32 nonsense SNVs introduced via prime editing, as *ATM*-deficient cells carrying nonsense mutations were expected to be depleted. In DMSO-treated cells, the median sLFCs for *ATM* nonsense variants in *ATM*-haploid and *ATM* semi-KO cells were -2.1 and -0.71, respectively, while in olaparib-treated cells, the values were -3.1 and -1.1 (**Figure 8**). Notably, in DMSO-treated *ATM* semi-KO cells, the sLFCs of seven nonsense variants exceeded zero, indicating a relatively low signal-to-noise ratio. Conversely, in olaparib-treated *ATM*-haploid cells, the sLFCs of all 32 nonsense variants were below -1.5, reflecting a higher signal-to-noise ratio.

These findings suggest: (i) *ATM*-haploid cells provide more accurate functional evaluations than *ATM* semi-KO cells, and (ii) the addition of olaparib enhances the signal-to-noise ratio in functional assessments of *ATM* variants. Based on these observations, we proceeded with *ATM*-haploid cells in the presence of olaparib for further functional evaluations of *ATM* variants.

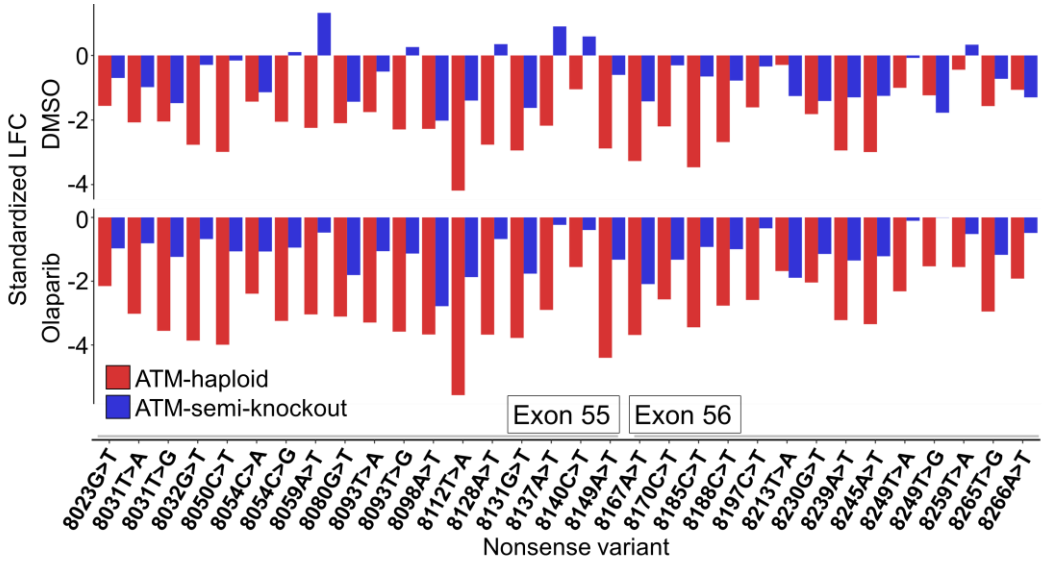


Figure 8. Comparison of standardized log2-fold changes of nonsense SNVs in *ATM*-haploid and *ATM*-semi-KO cells. Standardized log2-fold changes (sLFCs) in *ATM*-haploid cells (red), which contain only a single copy of wild-type *ATM*, and in *ATM*-semi-KO cells (blue), which contain both a single copy of wild-type *ATM* and another gene copy containing the c.3383dup frameshift mutation, that also contain the indicated mutations in exons 55 and 56, in the presence or absence of olaparib. The means of sLFCs in two replicates are shown for simplicity. DMSO represents the solvent control for olaparib. The x-axis shows a total of 32 nonsense mutations.

3.3. Olaparib increases the accuracy of functional evaluation of *ATM*

variants

To investigate the functional impact of all potential SNVs in *ATM*-haploid cells, we generated 62 distinct lentiviral libraries, each targeting a specific exon, containing epegRNAs. We then evaluated the effects of these SNVs both with and without olaparib treatment. The sLFCs of cells containing variants at day 10 (D10) were determined relative to day 0 (D0), which was 13 days post-transduction with the lentiviral epegRNA libraries (**Figure 7**). Our analysis revealed a stronger correlation between biological replicates in the olaparib-treated condition (Pearson correlation coefficient r = 0.76) compared to the DMSO-treated condition (r = 0.52) (**Figure 9A**), indicating a higher signal-to-noise ratio in the olaparib-treated group. The correlation between sLFCs from the DMSO- and olaparib-treated groups was 0.73 (**Figure 9B**).

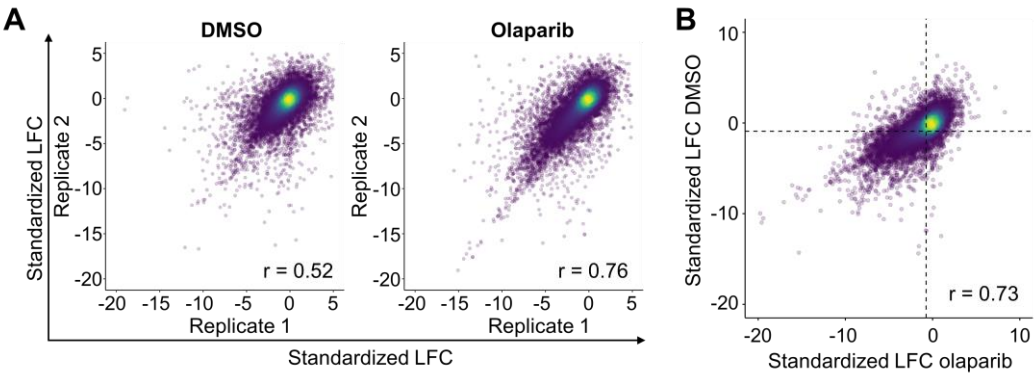


Figure 9. Correlation of standardized log2-fold changes between replicates. (A) Correlation between sLFCs of replicates in the solvent control (DMSO) (left) or olaparib (right) groups. **(B)** Correlation between sLFCs of the solvent control (DMSO) and olaparib groups. The Pearson correlation coefficient (r) is shown.

To assess the accuracy and sensitivity of the analyses under both DMSO and olaparib conditions, we conducted receiver operating characteristic (ROC) curve analyses. We assumed that 1,141 nonsense SNVs would disrupt *ATM* function, leading to reduced sLFCs, while 4,837 synonymous variants would preserve *ATM* function. The area under the curve (AUC) was higher in the olaparib-treated group (0.95) compared to the DMSO-treated group (0.89) (DeLong's test, $P = 3.6 \times 10^{-24}$) (**Figure 10A**). Based on Youden's J statistic, the optimal sLFC thresholds for distinguishing nonsense and synonymous variants were -0.912 (sensitivity = 93.0%, specificity = 91.4%) in the olaparib group and -0.745 (sensitivity = 81.2%, specificity = 89.3%) in the DMSO group.

We further performed ROC analyses using 1,603 variants previously classified in ClinVar, including 440 pathogenic or likely pathogenic (P/LP) variants and 1,163 benign or likely benign (B/LB) variants from multiple submitters. The AUC values for the olaparib and DMSO groups were 0.94 and 0.88, respectively (**Figure 10A**). When we restricted the analysis to 17 variants (11 P/LP and 6 B/LB) that excluded nonsense variants and were annotated by an expert panel, the AUCs were 1.00 for the olaparib group and 0.97 for the DMSO group (**Figure 10B**).

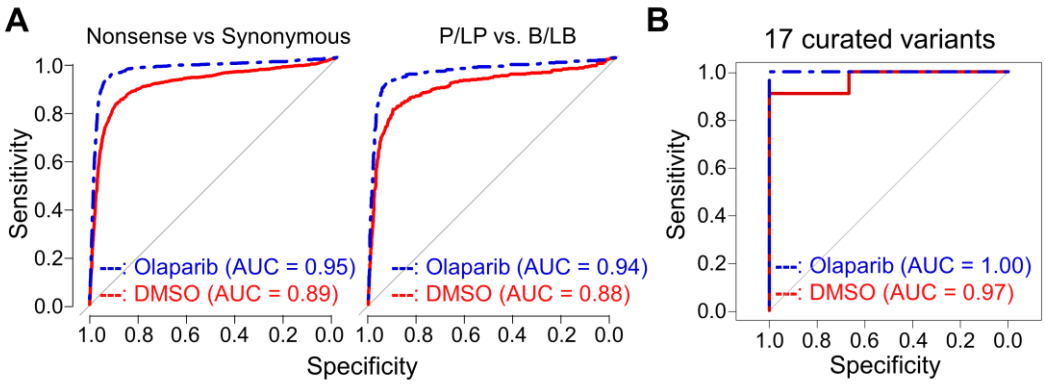


Figure 10. Receiver-operating-characteristic (ROC) curves for sLFCs of SNVs. (A) The left panel shows ROC curves for discriminating nonsense (n = 1,141) vs. synonymous variants (n = 4,837). Exons 63 and 62 were excluded from the analysis due to the possibility of mutations in these exons escaping nonsense-mediated decay. The right panel shows ROC curves for discriminating pathogenic/likely pathogenic (P/LP) (n = 440) vs. benign/likely benign (B/LB) variants (n = 1,163) annotated by ClinVar database. Area under the curve (AUC) values are shown. **(B)** ROC curves for discriminating 17 variants that do not include nonsense variants and that were classified by an expert panel (6 B/LB + 11 P/LP) in DMSO- and olaparib-treated conditions. Olaparib (blue) or the solvent control (DMSO, red).

Next, we examined the sLFC distributions across different variant types in both the olaparib and DMSO conditions. Nonsense and splice site variants, which are commonly associated with loss of *ATM* function, were significantly depleted compared to synonymous variants, with the effect being more pronounced in the olaparib group (**Figure 11**). These findings confirm that olaparib treatment enhances the signal-to-noise ratio in functional evaluations. As a result, we proceeded with functional assessments using olaparib-treated cells and designated the sLFCs from this group as "function scores" (Methods).

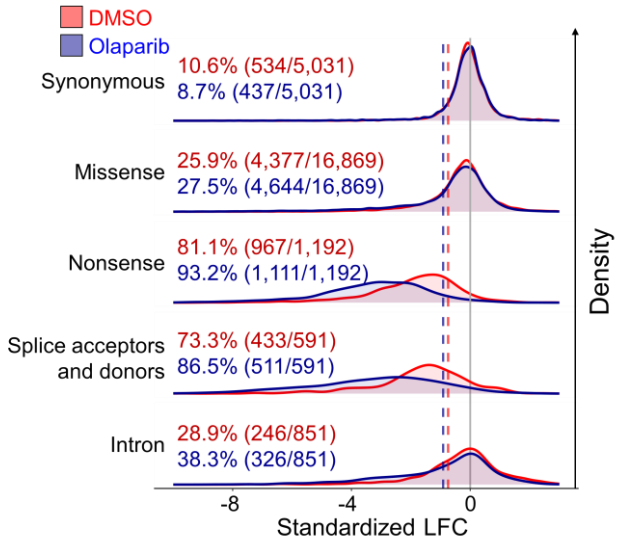


Figure 11. Kernel density estimation plots of SNV sLFCs. For each variant category, the number and percentage of SNVs with adjusted LFC values lower than cutoffs, representing Youden's indices (-0.745 and -0.912 in the solvent control (DMSO, red) and olaparib (blue) groups, respectively), are shown. The cutoffs are shown in blue (olaparib) or red (DMSO) dashed lines. The dark gray line represents sLFC = 0.

Each SNV was introduced by 2-3 epegRNAs, with each epegRNA incorporating a distinct synonymous mutation. We compared function scores among internal replicates and observed strong correlations (r = 0.61–0.69, mean = 0.65, **Figure 12**). Among the 18,651 amino acid substitutions analyzed, 2,012, 1,069, and 97 were encoded by two, three, or four distinct SNVs, respectively. A strong correlation (r = 0.64) was observed between function scores of SNV pairs that resulted in the same amino acid substitution (**Figure 13**). To ensure accuracy, we used the mean function scores across replicates from different experimentalists for further analyses. Additionally, only 0.16% of sequencing reads containing both the intended edits and synonymous edits included indels (**Figure 14**), and these rare cases were excluded from subsequent analyses (Methods).

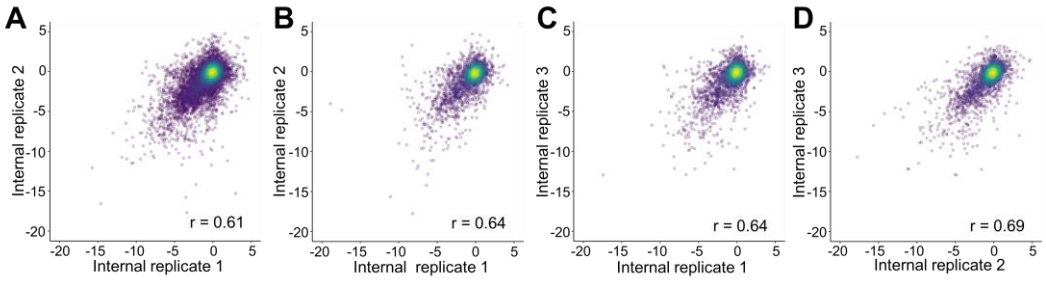


Figure 12. Correlations between sLFCs of replicates in the olaparib-treated group. (A)

Correlation between the sLFCs of two different synonymous indicator mutations, designated as internal replicates 1 and 2, for the same intended SNVs. **(B-D)** Correlation between the sLFCs of three different synonymous indicator mutations, designated as internal replicates 1, 2, and 3, for the same intended SNVs. Pearson's correlation coefficients (r) are shown.

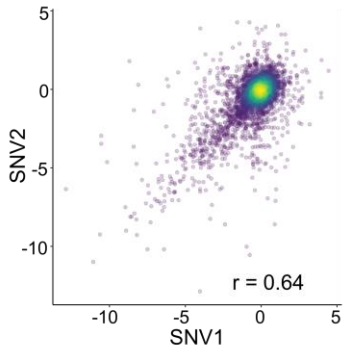


Figure 13. Correlation between the sLFCs of different SNVs encoding the same amino acid variants. The Pearson's correlation coefficient (r) is shown.

Figure 14. Proportions of reads containing indels (shown in red) with or without targeted SNVs.

To determine whether our method could be applied to other cancer predisposition genes, we conducted a similar screening in exons 4 and 19 of BRCA1. Nonsense and canonical splice site variants were significantly depleted relative to synonymous variants (**Figure 15**). ROC analyses comparing nonsense and synonymous SNVs showed that the AUC in the olaparib-treated group (0.91) was higher than in the DMSO-treated group (0.84) (DeLong's test, P = 0.017) (left panel, **Figure 16**). When analyzing 77 ClinVar-annotated variants (53 P/LP and 24 B/LB), the AUCs for the olaparib and DMSO groups were 0.94 and 0.82, respectively (middle panel, **Figure 16**). Furthermore, when focusing solely on 32 missense variants (30 P/LP and 2 B/LB), the AUCs were 0.98 for the olaparib group and 0.85 for the DMSO group (right panel, **Figure 16**). These findings suggest that our approach is applicable to evaluating BRCA1 variants as well.

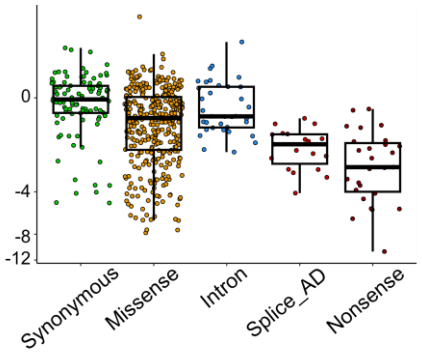


Figure 15. Distribution of function scores for different categories of BRCA1 variants. 'Intron' refers to mutations positioned -5, -4, and -3 bp from intron-exon junctions and +3, +4, and +5 bp from exon-intron junctions, whereas 'splice acceptors and donors' (splice AD) refers to mutations positioned -2, -1, +1, and +2 bp from intron-exon and exon-intron junctions. Boxes represent the 25th, 50th, and 75th percentiles, and whiskers show the 10th and 90th percentiles.

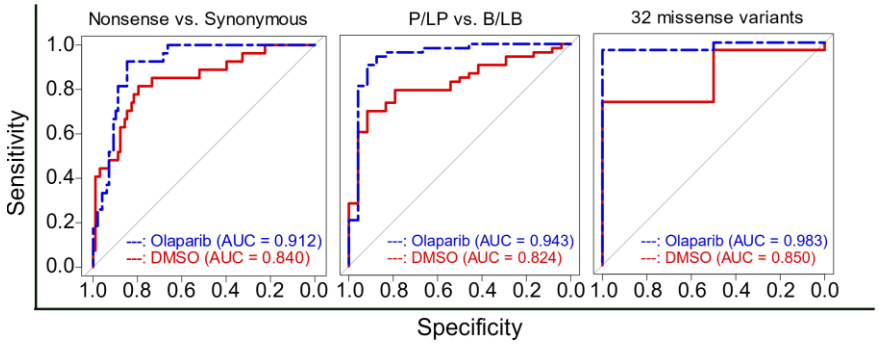


Figure 16. ROC curves for sLFCs of BRCA1 SNVs. The left panel shows ROC curves for discriminating nonsense (n = 27) vs. synonymous variants (n = 98). The middle panel shows ROC curves for discriminating pathogenic/likely pathogenic (P/LP) (n = 53) vs. benign/likely benign (B/LB) variants (n = 24) as annotated by the ClinVar database. The right panel shows ROC curves for discriminating 32 missense variants (30 P/LP + 2 B/LB) in DMSO- and olaparib-treated conditions. Area under the curve (AUC) values are shown. Olaparib (blue) or the solvent control (DMSO, red).

3.4. Function scores of 24,534 *ATM* variants

We experimentally determined function scores for 24,534 SNVs and categorized them into three groups: 'non-functional' (function score < -1.360), 'intermediate' (-1.360 ≤ function score < -0.912), and 'functional' (function score ≥ -0.912). These cutoffs were established using the 5th percentile of synonymous variant function scores and Youden's index (-0.912) (Methods). Variants falling into the non-functional or intermediate categories were collectively referred to as "depleting

variants.”

To further validate our large-scale functional assessments, we selected six non-functional missense variants that had previously been classified as variants of uncertain significance (VUSs). When these variants were introduced into cells and co-cultured with wild-type cells, a depletion of variant-containing cells was observed, with four out of six showing a more pronounced depletion in the presence of olaparib (**Figure 17**). Additionally, we examined two functional variants (R337C and R337H), which had conflicting pathogenicity reports, alongside one non-functional (L969P) and two functional (S1981C and C2991G) variants [24], all of which were initially classified as VUSs. The results of these experiments were consistent with our high-throughput functional assessments (**Figure 18**).

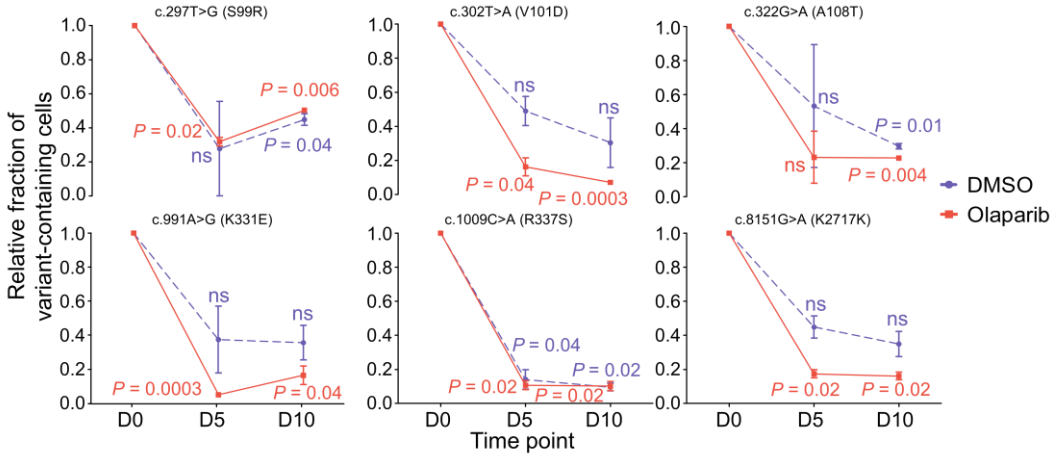


Figure 17. Comparison of non-functional SNVs’ function scores with individual evaluation results. Relative fraction of variant-containing cells after culturing them with wild-type cells in the absence (blue) or presence (red) of olaparib. DMSO is the solvent control. Statistical significance in comparision with D0 (paired t-test) is shown. ns, statistically not significant (P > 0.05).

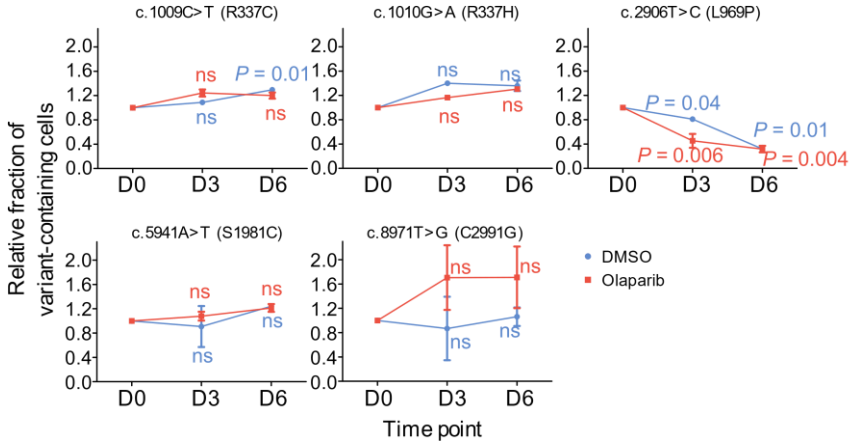
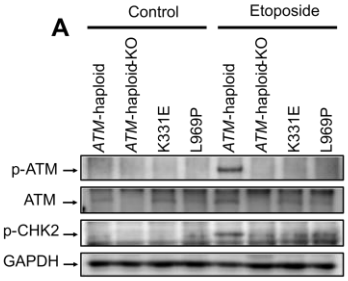


Figure 18. Comparison of functional and non-functional SNVs' function scores with individual evaluation results. Relative fraction of variant-containing cells after culturing them with wild-type cells in the absence (blue) or presence (red) of olaparib. DMSO is the solvent control. Statistical significance in comparision with D0 (paired t-test) is shown. ns, statistically not significant (P > 0.05).

Western blot analysis demonstrated that two non-functional variants (c.991A>G (K331E) and c.2906T>C (L969P)), previously considered VUSs, exhibited significant reductions or near-complete loss of ATM signaling. This was evident from decreased phosphorylation of ATM and CHK2 following etoposide treatment, a DNA-damaging agent (**Figure 19A**). Furthermore, deep sequencing of potential off-target sites in cells carrying these two non-functional variants showed no detectable off-target effects (**Figure 19B**). While prime editing has a low probability of off-target modifications [21, 56, 61], we cannot entirely exclude the possibility of unobserved or unassessed off-target effects, particularly for other variants.



B

Clone	Intended	Region	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Wild-type	Intended-edit	Indel	Other
K331E	On	chr11:108247030	T	A	A	G	T	C	A	T	A	T	A	G	G	A	A	G	T	A	G	A	G	G	0.0%	98.7%	0.0%	1.3%
	Off_1	chr8:107553890	T	A	A	G	T	g	A	T	A	T	A	G	G	A	A	G	a	A	G	G	G	G	99.6%	0.0%	0.0%	0.4%
	Off_2	chr20:38320364	T	g	A	G	T	C	A	g	A	T	A	G	G	A	A	G	T	A	G	G	G	G	100.0%	0.0%	0.0%	0.0%
	Off_3	chr4:188264845	T	A	A	G	T	g	A	T	A	T	A	G	a	A	A	G	T	A	G	A	G	G	99.5%	0.0%	0.0%	0.5%
	Off_4	chr6:90874607	T	A	A	G	T	t	A	T	t	T	A	G	G	A	A	G	T	A	G	A	G	G	99.5%	0.0%	0.0%	0.5%
L969P	On	chr11:108271094	G	A	A	G	A	G	T	A	C	C	C	C	T	T	G	C	C	A	A	T	G	G	0.0%	99.6%	0.0%	0.4%
	Off_1	chr11:46297704	G	g	A	G	A	G	T	A	C	C	C	C	-	T	G	C	C	A	A	T	G	G	99.6%	0.0%	0.0%	0.4%

Mismatch = 1 or 2, or Mismatch=1 and Bulge=1

Figure 19. Western blot and off-target evaluation of two non-functional variant clones. (A) Western blotting to examine the total ATM, phosphorylated ATM (p-ATM), and phosphorylated CHK2 (p-CHK2) protein levels in *ATM*-haploid clones containing newly identified non-functional variants (K331E or L969P). Arrows indicate the molecular weights of the indicated proteins. GAPDH was used as a loading control. **(B)** Off-target effects. DNA sequences at the on- and potential off-target sites were evaluated using deep sequencing in two clones containing non-functional variants, which were generated with the epegRNAs used for the high-throughput functional evaluations of the SNVs. The numbers at the top represent positions in the protospacer (1-19) and protospacer adjacent motif (NGG PAM, 20-22, gray). Base pair mismatches between the on- and off-target sites are highlighted in orange. Deep sequencing of target DNA sequences was performed to examine the frequencies of wild-type sequences, sequences containing intended edits, sequences containing indels, and other sequences.

We also compared our experimentally derived function scores with predictions from various computational tools, including CADD [40], REVEL [39], SIFT [32], PROVEAN [38], GERP [41], AlphaMissense [44], EVE [42], and BoostDM [29]. Among these, AlphaMissense and CADD showed the strongest correlations with our function scores, though the correlations remained modest (r = -0.47 and -0.45, respectively) (**Figure 20**). This underscores the necessity of experimental validation, as has been similarly observed for BRCA1 and BRCA2 variants [58, 62, 63].

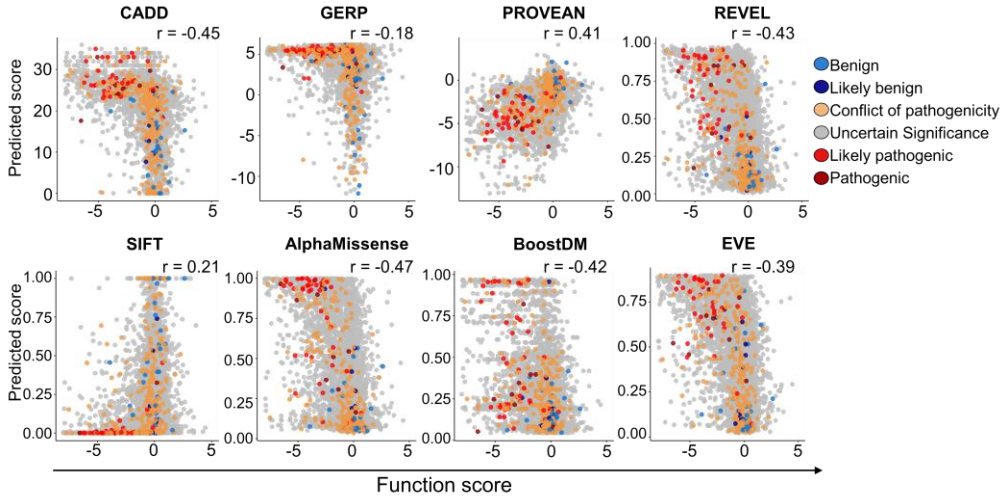


Figure 20. Correlations between experimentally measured function scores and functional effects predicted by previously developed computational models for missense SNVs. The functional classifications of variants in the ClinVar data are shown using different colored dots. Pearson correlation coefficients (r) are shown.

PhyloP, an in silico tool that predicts conservation scores at the nucleotide level, exhibited a strong correlation between the average PhyloP score per exon and the proportion of non-functional

variants within that exon (r = 0.76, P = 8.4 × 10^{-13}). This suggests that regions under stronger evolutionary constraint are less tolerant to amino acid changes (**Figure 21A**). Additionally, the BLOSUM substitution matrix, which predicts the likelihood of one amino acid replacing another [64], indicated that functional SNVs generally had higher BLOSUM scores than non-functional ones (**Figure 21B**). A weak but significant positive correlation was observed between BLOSUM scores and function scores (r = 0.22, P = 1.2 × 10^{-243}) (**Figure 21C**).

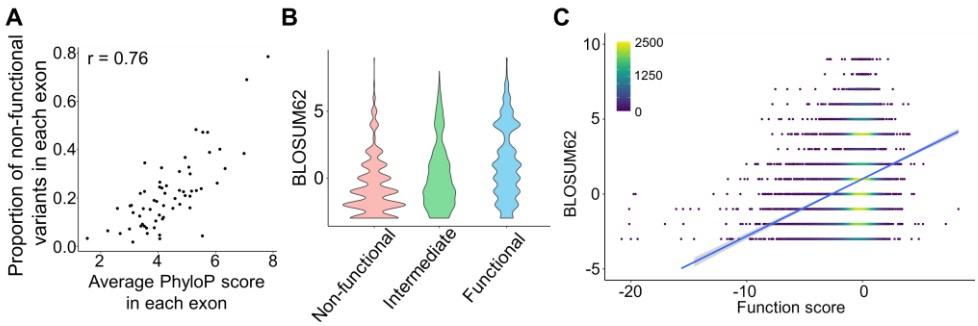


Figure 21. Correlation between the conservation scores and function scores. (A) Proportions of non-functional variants among missense variants in each exon are plotted versus the average PhyloP score for each exon. **(B)** Violin plots showing the distribution of BLOSUM62 scores for each of our functional classifications. **(C)** Correlation between the BLOSUM62 scores and the function scores. A trend line based on a linear regression is shown. The color of each dot was determined by the number of neighboring dots (that is, dots within a distance that is 1.5 times the default radius of the dot).

Function scores were generally low for nonsense and splice site variants, whereas missense variants showed a wider range of functional effects (**Figure 22A**). When classified by variant type, 88% of nonsense variants, 79% of splice site variants, 30% of intronic variants, and 20% of missense variants were identified as non-functional (**Figure 22B**). These findings align with previously published results for BAP1, VHL, and DDX3X [57, 60, 64, 65]

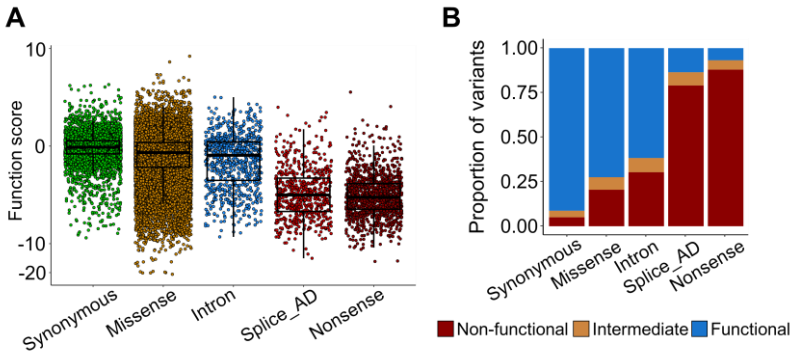


Figure 22. Distribution of function scores and proportions of SNV classification for the categories of variants. (A) Distribution of function scores for different categories of variants. 'Intron' refers to mutations positioned -5, -4, and -3 bp from intron-exon junctions and +3, +4, and +5 bp from exon-intron junctions, whereas 'splice acceptors and donors' (splice AD) refers to mutations positioned -2, -1, +1, and +2 bp from intron-exon and exon-intron junctions. Boxes represent the 25th, 50th, and 75th percentiles, and whiskers show the 10th and 90th percentiles. **(B)** Proportions of non-functional, intermediate, and functional SNVs for the indicated categories of variants. splice AD, splice acceptors and donors.

We further analyzed the impact of different amino acid substitutions on function scores. Among 150 possible amino acid substitution types, those involving tryptophan (W>G, W>C, W>R, W>S, and W>L) frequently resulted in non-functional variants (**Figure 23A**). Similarly, substitutions such as V>D, L>P, R>P, Y>D, and L>R often led to non-functional effects, consistent with prior studies indicating that L>P, L>R, and R>P substitutions are frequently associated with phenotypic changes [66]. Grouping amino acids into categories based on polarity and charge—nonpolar, polar uncharged, positively charged, and negatively charged—we found that substitutions from nonpolar to charged amino acids most commonly led to reduced function scores (**Figure 23B**).

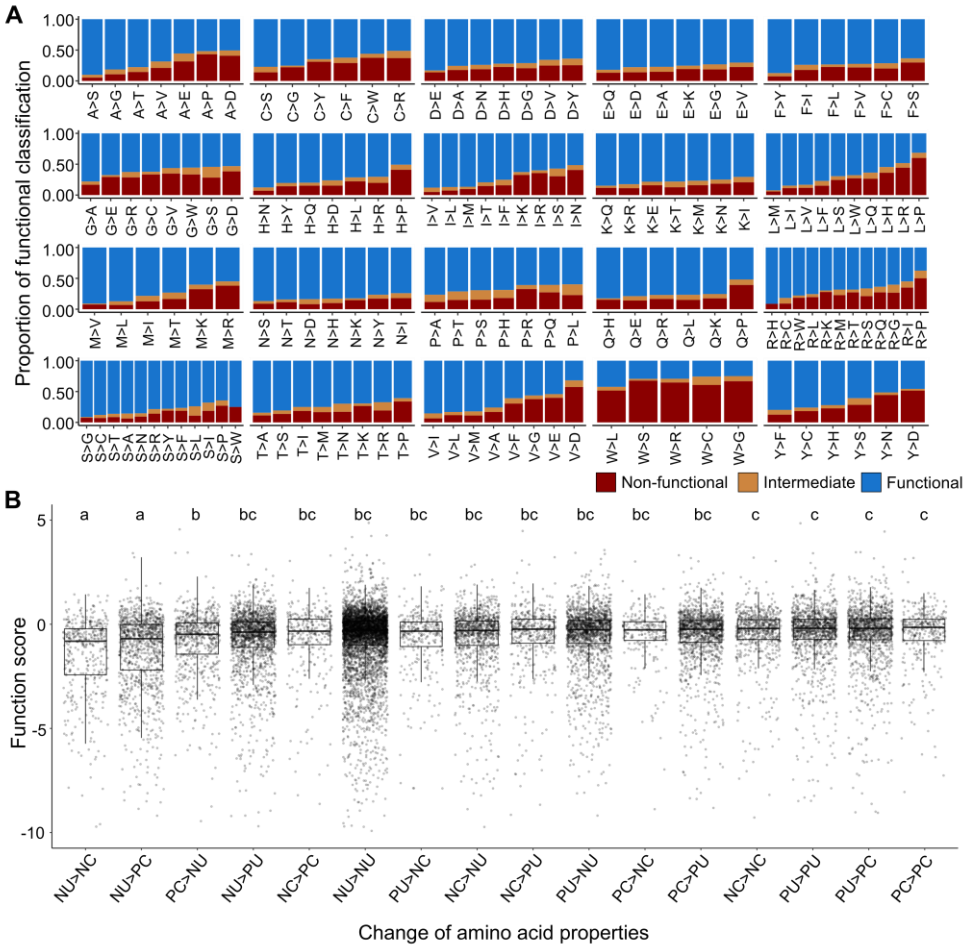


Figure 23. Proportions of functional categories for each amino acid substitution and distribution of functions scores for amino acid substitution types. (A) Proportions of non-functional and intermediate SNVs for each type of amino acid substitution shown on the x-axis. **(B)** Effect of the type of amino acid substitution on function score distributions. Subsets of types of amino acid changes without statistically significant differences between them ($P > 0.05$, analysis of variance (ANOVA) followed by Tukey's post hoc test) in the function scores are indicated with a, b, and c. NU, non-polar uncharged; NC, negative-charged; PC, positive-charged; PU, polar uncharged.

Examining function scores in relation to the observed frequency of SNVs at day 0, we found that nonsense variants with higher starting frequencies tended to show greater reductions in function scores (**Figure 24A**). ROC analysis further revealed that the reliability of functional assessments decreased when SNV frequencies were below 0.001% (**Figure 24B and 24D**). Based on SNV frequency at day 0, we classified functional evaluation confidence levels into three categories: high confidence (68% of functionally assessed missense SNVs, SNV frequency > 0.001%), medium-high

confidence (29% of missense SNVs, SNV frequency between 0.0001% and 0.001%), and medium confidence (2.8% of missense SNVs, SNV frequency < 0.0001%) (**Figure 24C**).

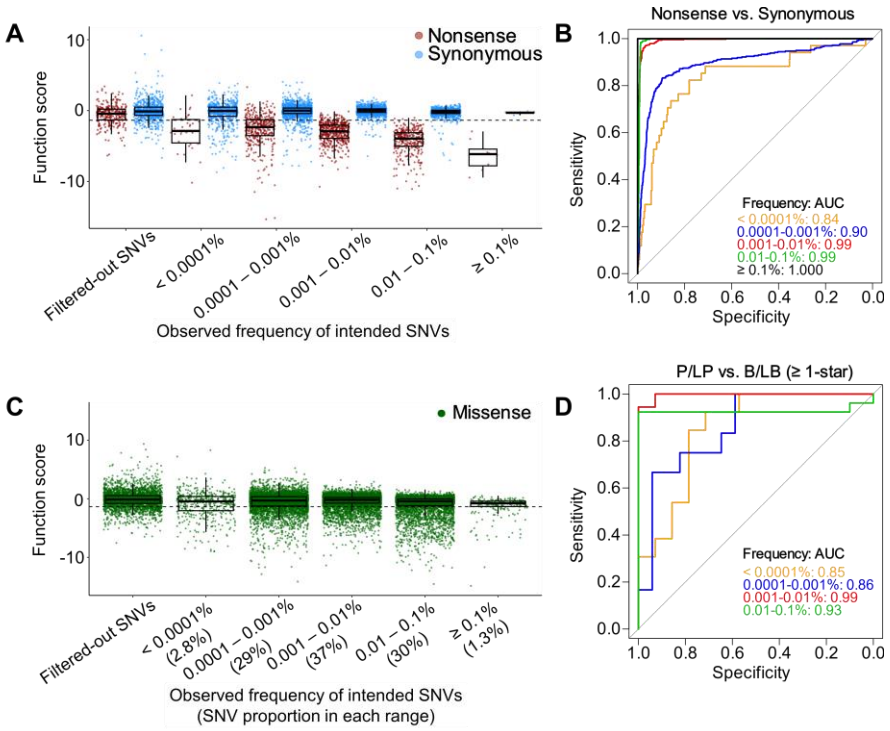


Figure 24. Distributions and functional classification accuracies of function scores for different ranges of variant frequencies at day 0 (D0). (A) Distribution of function scores for nonsense and synonymous SNVs for different ranges of SNV frequencies at D0. **(B)** ROC curves for different ranges of SNV frequencies at day 0 for discriminating nonsense vs. synonymous variants. **(C)** Distribution of function scores for missense variants for different ranges of SNV frequencies at D0. The proportion of variant numbers in each range of SNV frequencies among all missense SNVs that were functionally evaluated and classified is shown in parenthesis on the x-axis. Boxes represent the 25th, 50th, and 75th percentiles, and whiskers show the 10th and 90th percentiles. **(D)** ROC curves for different ranges of SNV frequencies for discriminating pathogenic/likely pathogenic (P/LP) vs. benign/likely benign (B/LB) ClinVar variants with ≥ one-star status (the number of variants n = 116). AUC values are shown.

3.5. Effect of the variant position on *ATM* function scores

We proposed that our functional assessments could help identify key regions essential for ATM protein function. To explore this, we analyzed function scores of SNVs across the entire coding sequence (**Figure 25A**). Most synonymous variants within exons had neutral function scores, although a few, particularly those located near exon-intron boundaries, showed lower scores. Nonsense variants were consistently depleted across exons, including in the penultimate and final

exons. Since nonsense mutations occurring in the last exon or within the last 50 nucleotides of the penultimate exon can sometimes bypass nonsense-mediated decay [67], the depletion of these variants in such regions suggests that these exons play a significant role in *ATM* function. Additionally, variants located at splicing donor and acceptor sites were largely depleted (**Figure 25B**).

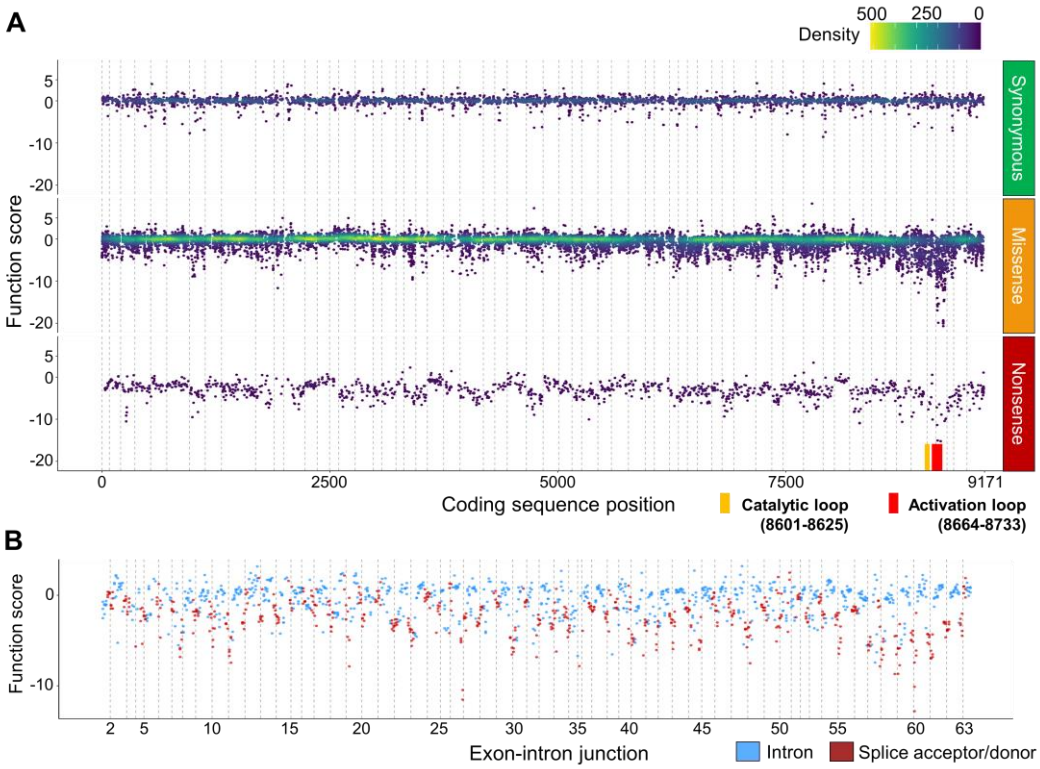


Figure 25. Effect of variant position on *ATM* function scores. (A) Function score map for SNVs across exons 2 to 63, categorized by variant type. Exons are separated by vertical dashed lines. The numbers of variants are indicated using dot colors. The color of each dot was determined by the number of neighboring dots (that is, dots within a distance that is 1.5 times the default radius of the dot). **(B)** Function score map for intronic and splice acceptor and donor variants. Vertical dashed lines indicate the exons (exons 2 to 63) located between consecutive introns.

For missense SNVs, the most significant depletion was observed in exons 57 to 60 (coding positions 8,269 to 8,786), where the mean function scores were -2.1 for exon 57, -2.1 for exon 58, -3.6 for exon 59, and -5.5 for exon 60. More than half of the missense SNVs within this region were classified as non-functional, compared to only 18% of missense SNVs in other exons (**Figure 26**), indicating that this region is critical for *ATM* activity. This segment corresponds to the highly conserved kinase domain (exons 55 to 63) [68], which is crucial for *ATM*'s function in the DNA

damage response, particularly its kinase activity [69].

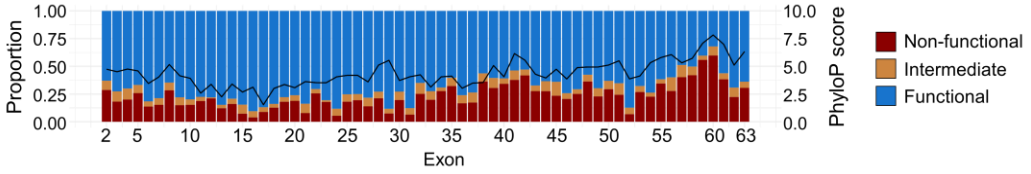


Figure 26. Proportions of depleting missense SNVs per exon. The black line represents the average PhyloP score for each exon, with the values on the right y-axis.

We further mapped the susceptibility of ATM amino acid residues to missense mutations (**Figure 27**). The activation loop (residues 2888 to 2911) and catalytic loop (residues 2867 to 2875) within the kinase domain, which are essential for substrate recognition and phosphorylation, were especially sensitive to missense alterations [70]. For instance, missense SNVs at residues D2870 and H2872, which interact with p53 at S15, had function scores ranging from -5.7 to -8.6 and -5.6 to -7.0, respectively, while all synonymous variants at these positions remained neutral. Similarly, residues involved in critical interactions with p53, such as T2902 (hydrogen bonding with Q16), L2900 (hydrophobic interaction), and F3049 (hydrophobic interaction), exhibited notably negative function scores, with mean values of -3.9, -9.4, and -3.1, respectively. These findings suggest that these regions are essential for *ATM* activity. Conversely, only 3.3% (13/396) of missense variants in exon 17, which contains many VUSs, were classified as non-functional, implying that this exon may be less crucial for *ATM* function. In conclusion, our functional analyses effectively pinpointed regions that are critical for ATM protein activity.

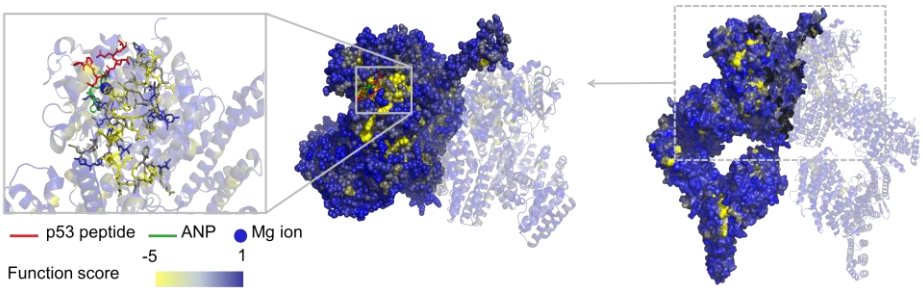


Figure 27. Mapping of intolerance to missense SNVs on the three-dimensional ATM structure. The average function score of missense SNVs at each amino acid position is shown on a color spectrum from yellow to blue (ranging from a minimum of -5 to a maximum of 1). In this dimeric representation of ATM, one of the two monomers is shown as a transparent secondary structure, for simplicity. In the magnified view of the boxed region, the red sticks, blue dot, and green sticks represent a p53 peptide, a magnesium ion, and ANP (phosphoaminophosphonic acid-adenylate ester, a synthetic analog of ATP), respectively. Amino acid residues encoded in exons 59 and 60 are depicted as sticks.

3.6. Clinical relevance of *ATM* function scores

We investigated the clinical relevance of our functional scores by examining their ability to differentiate variants classified as pathogenic/likely pathogenic (P/LP) and benign/likely benign (B/LB) in ClinVar. Our scores effectively distinguished these variant categories (**Figure 28A**). Additionally, we assessed our scores for splice site variants in relation to the ACMG/AMP interpretation criteria established by the ClinGen Hereditary Breast, Ovarian, and Pancreatic Cancer Expert Committee [14]. We observed a pattern in functional scores following PVS1 classification, ranging from PVS1 (indicating the strongest evidence of pathogenicity) to PVS1-strong, PVS1-supporting, and PVS1 N/A (**Figure 29**), demonstrating alignment between our scores and existing clinical classifications.

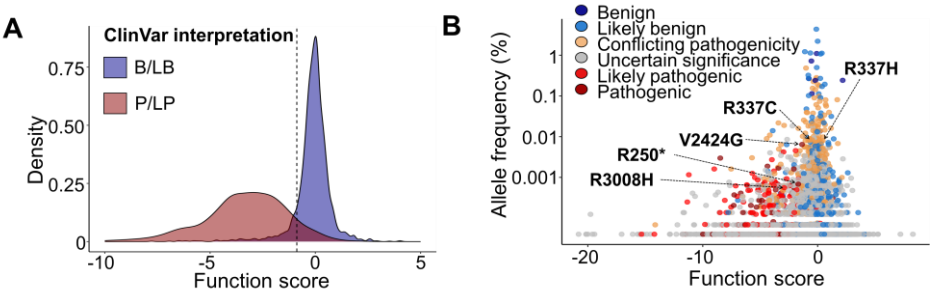


Figure 28. Clinical correlation between the ClinVar and GnomAD database and function scores. (A) Kernel density estimation plots of function scores for SNVs reported in ClinVar as P/LP (pathogenic or likely pathogenic) (n = 848), or B/LB (benign or likely benign) (n = 2,289). The cutoff for depleting variants, -0.912, is indicated with the dashed vertical line. **(B)** Function scores plotted against allele frequencies of SNVs in the general population (gnomAD v.4.1). ClinVar classifications are shown using different colored dots. Four variants most frequently observed in tumor samples and a variant with a strong association with breast cancer (c.7271T>G) are shown with arrows

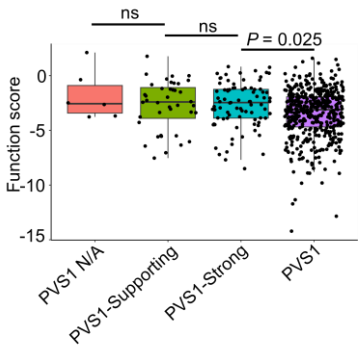


Figure 29. Box plots showing function scores of splice acceptor and donor SNVs. The functional

categories suggested by the ACMG guideline are shown on the x-axis. PVS1, Pathogenic very strong evidence; N/A, not applicable. The expected function of *ATM* decreases in the order of PVS1 N/A > PVS1-Supporting > PVS1-Strong > PVS1.

Individuals with biallelic pathogenic *ATM* variants develop ataxia-telangiectasia, a hereditary condition, while heterozygous carriers face an elevated risk of developing cancers such as breast, ovarian, and pancreatic cancer [2, 71-74]. Based on this, we hypothesized that non-functional variants would be less frequent in the general population. A comparison of variant frequencies from the gnomAD v4.1 dataset (n = 807,162) [26] with different function scores revealed that SNVs with low function scores were rare, whereas those with neutral scores were more prevalent, as expected (**Figure 28B**). Notably, SNVs with a population allele frequency exceeding 0.05% (classified as benign according to ACMG's BS1 criterion) had an average function score of -0.012, significantly higher than the -0.80 average score observed for variants with frequencies below 0.05% ($P = 1.1 \times 10^{-8}$) [14].

ATM mutations have been implicated in increased cancer risk, particularly for breast cancer [6, 9, 10, 75]. To evaluate whether our functional analysis could predict cancer susceptibility, we assessed cumulative cancer incidence using UKB data. Among 424,909 participants without a prior cancer diagnosis, 2,427 individuals carried 382 non-functional SNVs, 15,557 had 122 intermediate SNVs, and 107,625 possessed 1,612 functional SNVs, all of which were functionally assessed in this study. Participants were categorized based on their *ATM* SNVs, with those carrying multiple variants assigned to the most functionally disruptive category. Individuals with non-functional variants exhibited a significantly increased cancer incidence ($P = 8.0 \times 10^{-8}$) compared to those with intact *ATM* (**Figure 30**, left). The intermediate SNV group also demonstrated a slightly elevated cancer incidence relative to the intact group ($P = 0.005$), whereas the functional variant group showed no significant difference from the intact group. These trends remained consistent when the analysis was restricted to missense SNVs (**Figure 30**, right). Importantly, 91% (245/268) of the non-functional missense SNVs identified in this cohort were previously unreported or classified as variants of uncertain significance (VUS), underscoring the novel contributions of this study in refining *ATM* functional assessments.

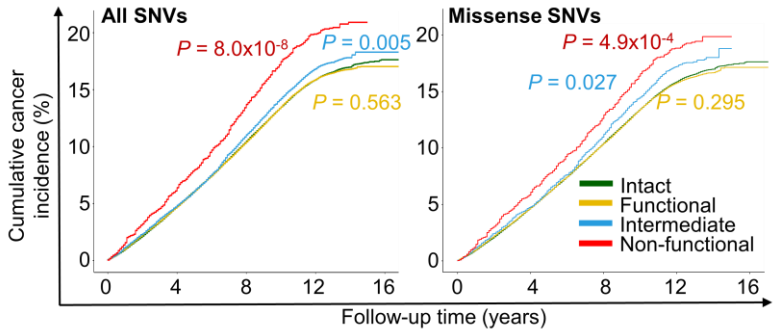


Figure 30. Cumulative cancer incidence in UKB participants (n = 424,909) with different functional categories of *ATM* variants. The left panel includes participants with all SNV mutation types, and the right panel includes only participants with missense SNVs and intact *ATM*. *P*-values

are shown for each group in comparison with the intact *ATM* group.

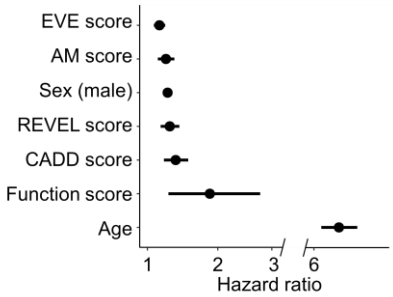


Figure 31. Hazard ratios of cancer incidence for various computational scores and the function score. Black bars represent 95% confidence intervals. AM score, AlphaMissense score.

We further investigated cancer hazard ratios (HRs) based on function scores using Cox proportional hazards regression, adjusting for age and sex. Among all variables, age had the highest HR (6.7), followed by function score (HR, 1.9), while other predictors such as sex (HR, 1.3) and computational models (EVE, 1.2; AlphaMissense, 1.2; REVEL, 1.3; CADD, 1.4) had comparatively lower HRs (**Figure 31**). To evaluate lifelong cancer risk across different functional groups, we analyzed the age at first cancer diagnosis. Cancer onset occurred significantly earlier in the non-functional group compared to the intact group, whereas the functional group showed no significant difference from the intact group (**Figure 32**). When focusing on breast cancer among female participants, the non-functional group exhibited a markedly higher cumulative cancer incidence for this cancer type (**Figure 33A and 33B**), consistent with previous studies linking *ATM* mutations to elevated breast cancer risk [6, 9, 75]. These findings indicate that our *ATM* function scores may serve as a useful predictor of cancer risk in individuals carrying *ATM* variants.

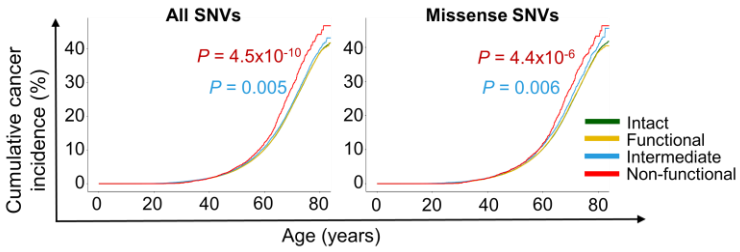


Figure 32. Lifelong cancer incidence in UK Biobank participants with different functional categories of *ATM* variants determined using the function score. The left panel includes participants with all SNV mutation types, and the right panel includes only participants with missense SNVs and intact *ATM*. *P*-values are shown for each group in comparison with the intact *ATM* group. Classifications of *ATM* SNVs based on function scores are indicated with different colors (intact, green; functional, yellow; intermediate, blue; non-functional, red)

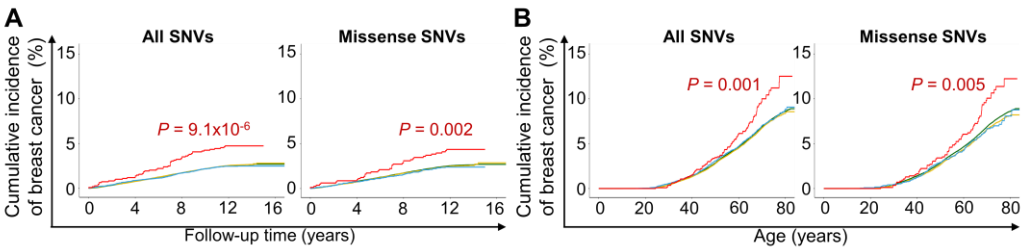


Figure 33. Cumulative incidence of breast cancer in UK Biobank female participants with different functional categories. (A) Cumulative breast cancer incidence in female UK Biobank participants with different functional categories of *ATM* variants determined using the function score. The left panel includes participants with all SNV mutation types, and the right panel includes only participants with missense SNVs and intact *ATM*. **(B)** Lifelong breast cancer incidence in female UK Biobank participants with different functional categories of *ATM* variants determined using the function score. The left panel includes participants with all SNV mutation types, and the right panel includes only participants with missense SNVs and intact *ATM*. The *P*-value is shown for the non-functional group in comparison with the intact *ATM* group.

We further assessed the clinical implications of function scores using data from two cohort studies that examined germline *ATM* variants in relation to breast cancer susceptibility (6,796 cases and 3,388 controls) [76, 77], supplemented with gnomAD data as additional controls (**Data not shown**). Focusing on 159 missense variants classified as VUSs (out of 276 total variants), we examined breast cancer odds ratios (ORs) based on function scores. Individuals with non-functional and intermediate variants had ORs of 4.0 ($P = 2.2 \times 10^{-15}$, n = 44) and 2.0 (P = 0.010, n = 12), respectively (**Figure 34**). By comparison, ORs based on classifications from AlphaMissense, CADD, and REVEL were either statistically insignificant or only marginal. Among them, variants with AlphaMissense scores exceeding 0.56 (n = 38) showed the highest OR (1.6, P = 0.002), but this result was less robust. The stronger association between non-functional variants and breast cancer risk, as determined by our functional approach, further supports the clinical utility of our scoring method in assessing cancer risk among *ATM* variant carriers.

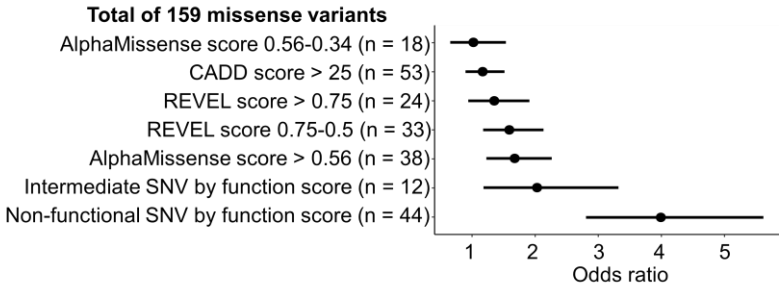


Figure 34. Associations between functional subsets of missense variants and their occurrence as germline variants in breast cancer patients. Pathogenic variant subsets were determined using

the known cutoff values of computational scores calculated using AlphaMissense, REVEL, and CADD, or using our function scores. Odds ratios were calculated by comparing the occurrence of each pathogenic variant subset in tumor samples to that of the benign variant subset. Black bars represent 95% confidence intervals.

To investigate the correlation between function scores and cancer genomics, we analyzed tumor sequencing data from the AACR GENIE Cohort v16.0 (Genomics Evidence Neoplasia Information Exchange; hereafter referred to as GENIE) [25], which comprises data from 184,988 cancer patients. Among the 5,343 *ATM* SNVs identified in the GENIE dataset, 4,338 were functionally assessed in this study. Of the 938 variants classified as 'Oncogenic' or 'Likely oncogenic' according to OncoKB, 724 (77%) were non-functional (**Figure 35A**). Furthermore, 29% (984/3,392) of variants labeled as 'Uncertain' by OncoKB were also classified as non-functional in our analysis, suggesting that a substantial fraction of variants with uncertain oncogenic potential may, in fact, be pathogenic. Four variants frequently observed in cancer samples (c.1009C>T [R337C], c.1010G>A [R337H], c.9023G>A [R3008H], and c.748C>T [R250*]) have been primarily classified as oncogenic due to their high prevalence in cancer cases (**Figure 35B**) [25]. However, our experimental data indicate that R337C and R337H exhibit neutral function (scores: -0.65 and 0.33, respectively) and are rare in the general population (0.017% and 0.007%, respectively) (**Figure 28B**), suggesting that their current oncogenic classification should be reevaluated. Conversely, R3008H and R250* were found to be non-functional, aligning with their ClinVar pathogenic classifications. Additionally, the c.7271T>G (V2424G) variant, associated with a 69% breast cancer risk [78], was also determined to be non-functional in our analysis.

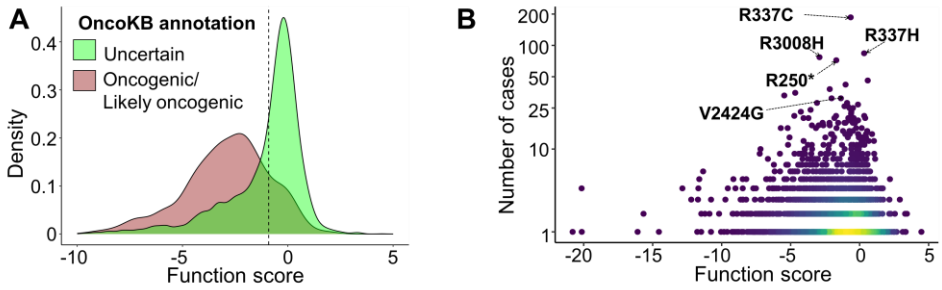


Figure 35. Distribution of function scores in GENIE database. (A) Kernel density estimate plots of function scores for SNVs (n = 4,338) found in tumor sequencing data, classified by the OncoKB database. The cutoff for depleting variants, -0.912, is indicated with the dashed vertical line. **(B)** Function scores of SNVs plotted against the number of observations in tumor samples. Four variants most frequently observed observed in tumor samples and a variant with a strong association with breast cancer (c.7271T>G) are shown with arrows.

Given that pathogenic variants in *ATM* have been implicated in an increased susceptibility to multiple cancer types, including breast and pancreatic cancer [7, 79], we investigated the association between *ATM* missense SNVs and cancer. We assessed their frequency in tumor and non-cancer populations. Specifically, we classified missense SNVs based on their functional impact using our

system and compared their distribution between cancer cases (n = 7,611) derived from the GENIE tumor sequencing dataset and non-cancer controls (n = 74,023) from gnomAD v3.1.2. ORs were computed for each SNV based on its frequency in cases and controls. The analysis revealed a strong association between non-functional SNVs (n = 1,069) and pan-cancer occurrence (OR = 6.2, P = 1.0×10^{-171}), while intermediate variants (n = 268) exhibited a weaker yet statistically significant association (OR = 1.2, P = 0.012) (**Figure 36A**). It is important to note that elevated pan-cancer occurrence denotes an overall increased frequency across all cancer types rather than a uniformly high prevalence within individual cancer types. Furthermore, the ORs of non-functional variants classified by AlphaMissense, CADD, and REVEL were notably lower than those identified as non-functional through our experimental assessments. Among these, variants with AlphaMissense scores exceeding 0.56 (the threshold corresponding to 90% precision, n = 1,007) demonstrated the highest OR (3.8). Given that the ORs of predicted non-functional variants are contingent on the chosen score thresholds, we systematically varied cutoff values to examine their influence. The results indicated that function scores derived from our system yielded higher ORs than those based on alternative methods such as AlphaMissense (**Figure 36B**), suggesting that function-based classification may provide clinically relevant insights into the cancer-associated risk of *ATM* variants.

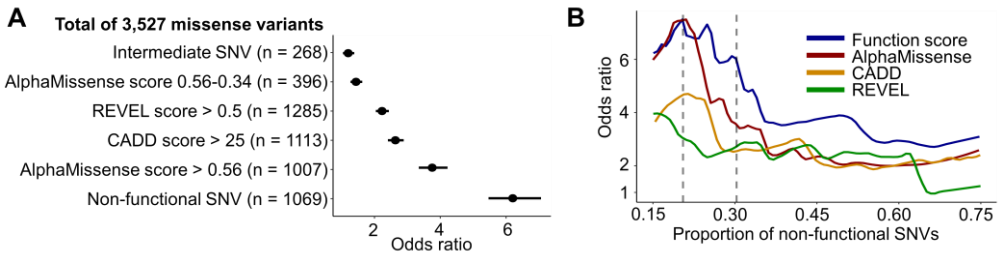


Figure 36. Associations between functional subsets of missense variants and their occurrence in tumor samples. (A) Pathogenic variant subsets were determined using the known cutoff values of computational scores determined by AlphaMissense, REVEL, and CADD, or using our function scores. Odds ratios were calculated by comparing the occurrence of each pathogenic variant subset in tumor samples to that of the benign variant subset. Black bars represent 95% confidence intervals. **(B)** Odds ratios plotted across varying proportions of non-functional SNVs. The variation in the proportions of non-functional SNVs was induced by changing cutoff values for each scoring system. The dashed lines represent the proportions of non-functional SNVs at 20% and 30%, which correspond to the proportions of non-functional missense SNVs in our dataset and the GENIE tumor sequencing data, respectively.

ATM mutations have also been associated with poorer prognosis in chronic lymphocytic leukemia (CLL) [18, 80, 81] but with improved outcomes in bladder cancer [20, 82]. In CLL patients [83] with non-functional or intermediate *ATM* variants (n = 60), survival outcomes were significantly worse than in those with intact *ATM* (n = 829) (**Figure 37**). To assess the prognostic impact of *ATM* missense variants, we conducted an analysis excluding patients with premature truncation variants (nonsense, frameshift, or splice site acceptor/donor). Among patients harboring only missense variants, those with depleting missense variants (n = 36) exhibited significantly reduced failure-free survival compared to other missense variant carriers (32 vs. 61 months, P = 3.7

$\times 10^{-4}$) (**Figure 38**).

To further investigate this relationship, we analyzed genomic data from 623 patients with stage III-IV bladder cancer [84-89] obtained from cBioPortal [30, 31]. Patients with depleting *ATM* variants (n = 34) demonstrated significantly longer overall survival than those with intact *ATM* (n = 557) (77 vs. 22 months, P = 0.044). Additionally, although progression-free survival was longer in the depleting *ATM* group, the difference did not reach statistical significance (P = 0.118). Patients carrying functionally intact *ATM* variants (n = 32) had overall and progression-free survival comparable to those in the intact *ATM* group (**Figure 38**).

These findings, together with the analysis of patients with chronic lymphocytic leukemia (CLL), support the utility of our functional classification in predicting the clinical prognosis of cancer patients harboring *ATM* variants.

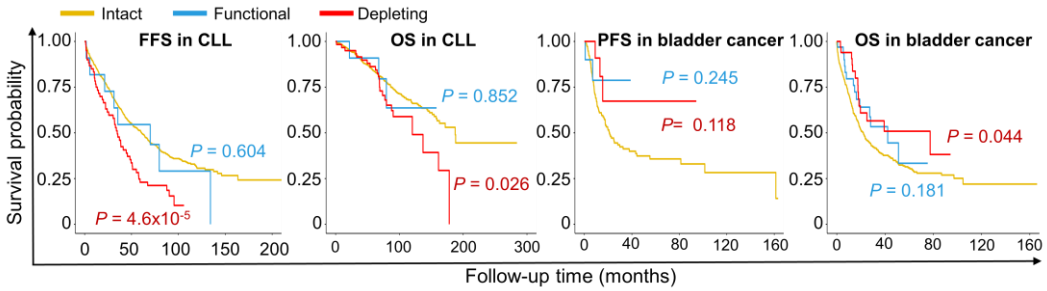


Figure 37. Prognosis of cancer patients with different functional categories of *ATM* variants. Patients with chronic lymphocytic leukemia (CLL, the number of patients, n = 900) and those with stage III or IV bladder cancer (n = 623) were categorized into three groups based on the functional classes of their somatic variants: depleting (non-functional + intermediate) variants, functional variants, and wild-type *ATM*. Survival analysis was conducted using the Kaplan-Meier estimator. P-values of survival comparisons between the intact *ATM* group and functional (blue) or depleting groups (red) are shown. FFS, failure-free survival; OS, overall survival; PFS, progression-free survival.

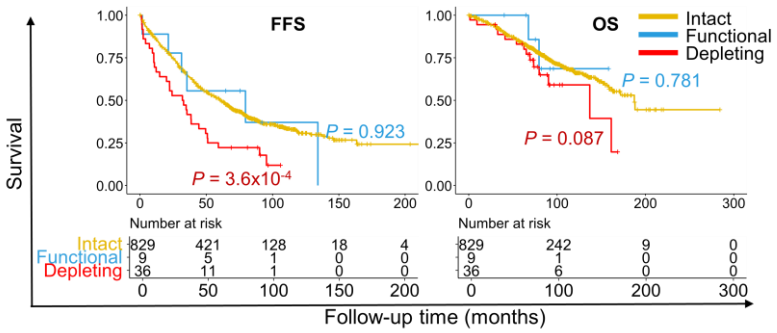


Figure 38. Prognosis of cancer patients with different functional categories of *ATM* missense variants. Patients with chronic lymphocytic leukemia (CLL, the number of patients, n = 874) harboring missense variants were categorized into three groups based on the functional classes of

their somatic variants: depleting (non-functional + intermediate) variants, functional variants, and wild-type *ATM*. Survival analysis was conducted using the Kaplan-Meier estimator. P-values of survival comparisons between the intact *ATM* group and functional (blue) or depleting groups (red) are shown. FFS, failure-free survival; OS, overall survival; PFS, progression-free survival.

3.7. Deep learning-based prediction of the functional effects of *ATM* variants

Out of the 27,513 potential SNVs within the *ATM* coding sequence, 4,421 could not be analyzed due to insufficient prime editing, particularly in AT-rich regions lacking the NGG PAM motif (**Figure 39A and 39B**). To address this, we proposed that the function scores for these 4,421 SNVs could be computationally estimated using experimentally determined scores from the remaining 23,092 SNVs. For this purpose, we applied a transformer-based deep learning model that incorporated variant positions, classifications, the AlphaFold 3-derived ATM protein structure [46], and scores from existing models such as AlphaMissense (**Figure 40**, Methods). The model was assessed using 116 variants functionally categorized as P/LP or B/LB in ClinVar. After filtering out variants that influenced the same amino acid sites as those in the test dataset, we retained 16,275 missense, 1,183 nonsense, and 4,395 synonymous variants for model training (Methods). The model, named DeepATM, was validated through five-fold cross-validation, yielding a median Pearson correlation coefficient of 0.65. Excluding structural data from AlphaFold 3 led to a slight reduction in the median Pearson correlation coefficient to 0.61 (P = 0.032, **Figure 41A**), suggesting that including ATM structural information enhances functional effect predictions. Using random forest instead of deep learning for missense variants alone, the median Pearson correlation coefficients for DeepATM with and without structural data, and random forest, were 0.61, 0.57, and 0.55, respectively (**Figure 41B**).

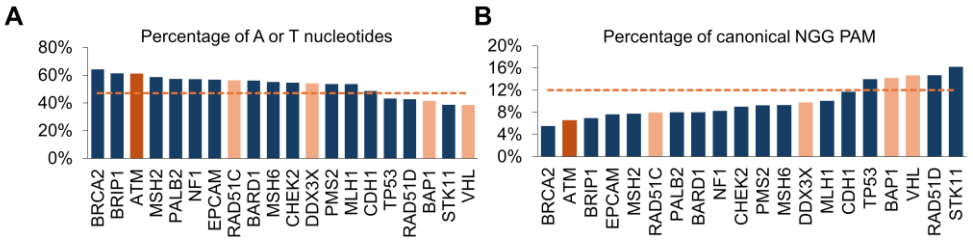


Figure 39. Sequence compositions of cancer-related genes. (A) Percentages of A and T nucleotides within the coding sequences of the hereditary cancer-associated genes shown on the x-axis. The mean percentage of A and T nucleotides for 19,284 human genes is 47%, the value that is indicated with the dashed horizontal line. **(B)** Percentages of 3-bp sequences that are canonical PAM sequences (NGG or CCN) within the coding sequences of the hereditary cancer-associated genes shown on the x-axis. The mean percentage of such sequences for 19,284 human genes is 12%, the value that is indicated using the dashed horizontal line. Four genes, whose variants were functionally evaluated in a high-throughput manner, are indicated using light red bars.

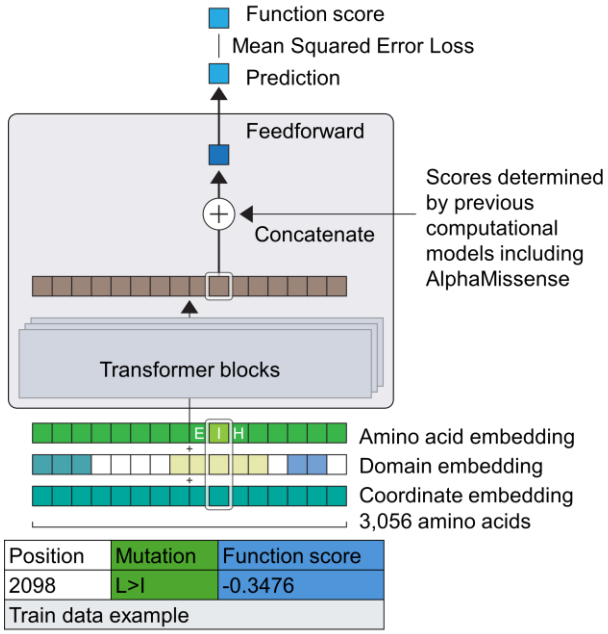


Figure 40. Schematic representation of DeepATM.

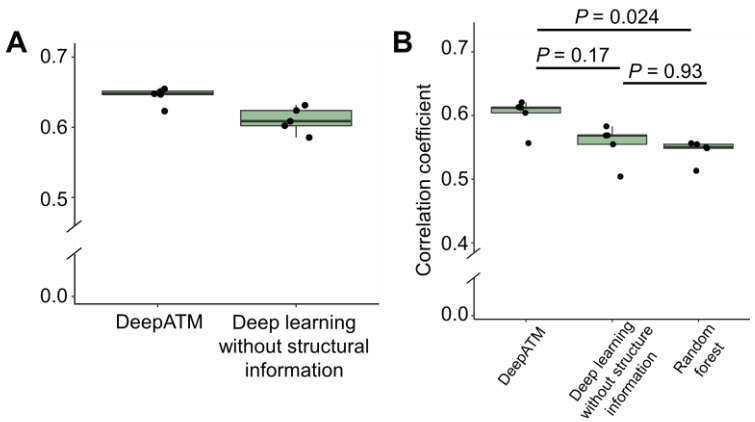


Figure 41. Results of five-fold cross-validation for machine learning models. (A) Pearson's correlation coefficients in five-fold cross-validation for DeepATM and the same transformer-based model trained without protein structural information. **(B)** Pearson's correlation coefficients for missense variants in five-fold cross-validations for DeepATM, the same transformer-based model trained without protein structural information, and the random forest model. Statistical significance by Wilcoxon's test with Bonferroni correction is shown.

DeepATM scores were transformed into experimentalized DeepATM scores (eDA scores) via a rank-based adjustment and regression, aligning their distribution with function scores (Methods, **Figure 42A**). The eDA scores for both the 23,092 and 4,421 SNVs exhibited comparable distributions (**Figure 42A**), and a strong correlation was observed between eDA scores and function scores (r = 0.70, **Figure 42B**).

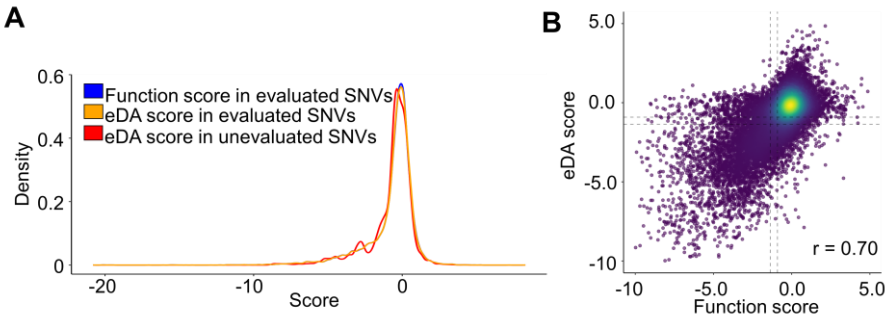


Figure 42. Relationships between the eDA scores and function scores. (A) Distribution of the function and eDA scores of the 23,092 experimentally evaluated SNVs, and of the eDA scores for 4,421 unevaluated SNVs. The distributions of function and eDA scores for the 23,092 SNVs were almost identical. **(B)** Correlations between the eDA and function scores for 23,092 SNVs. The cutoff values (-1.360 and -0.912) used for functional classification of SNVs are shown with dashed lines.

These scores effectively differentiated P/LP variants from B/LB variants (**Figure 43**). ROC analysis using 116 ClinVar-classified missense variants as the test dataset demonstrated that DeepATM had the highest AUC (0.95) among evaluated models (**Figure 44A**), with AlphaMissense ranking next at an AUC of 0.91. Under a stricter classification criterion (ClinVar two-star or higher), the test dataset was reduced to 68 variants, resulting in DeepATM achieving an AUC of 0.99, which significantly surpassed AlphaMissense (0.94, DeLong's test, P = 0.034) (**Figure 44B**). For the 4,421 previously unassessed variants, using 240 variants with a ClinVar classification of at least two stars and 455 variants with at least one-star classification, the AUCs were 1.00 and 0.99, respectively (**Figure 45**), demonstrating DeepATM's high predictive accuracy.

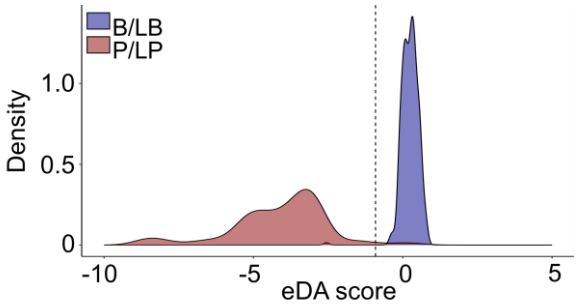


Figure 43. Kernel density estimation plots of eDA scores for unevaluated SNVs reported in ClinVar as P/LP (pathogenic or likely pathogenic) (n = 220), or B/LB (benign or likely benign) (n = 343). The cutoff for depleting variants, -0.912, is indicated with the dashed vertical line.

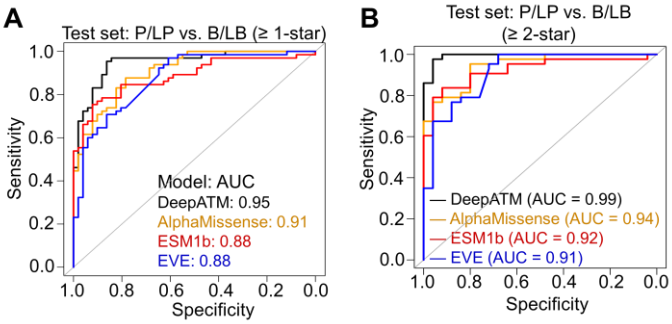


Figure 44. ROC curves for computationally calculated function scores. (A) ROC curves of 116 SNVs in the test set that have been functionally classified as either pathogenic/likely pathogenic or benign/likely benign in ClinVar with ≥ one-star status. **(B)** ROC curves of 68 SNVs in the test set that have been functionally classified as either pathogenic/likely pathogenic or benign/likely benign in ClinVar with ≥ 2-star status.

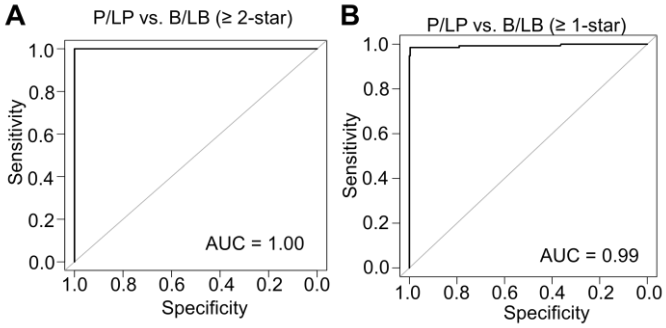


Figure 45. ROC curves for eDA-based functional classification. (A) ROC curves for eDA-based functional classification of 240 SNVs in the unevaluated set with ClinVar classifications with ≥ 2-

star status. **(B)** ROC curves for eDA-based functional classification of 455 SNVs in the unevaluated set with ClinVar classifications with ≥ one-star status. Area under the curve (AUC) values are shown.

Similar to function scores, variants with low eDA scores tended to have lower allele frequencies, while those with high eDA scores were more frequent (**Figure 46A**). In UK Biobank data, 425 of the 4,426 SNVs without function scores but with eDA scores were identified. Individuals carrying non-functional eDA scores had significantly increased cancer incidence ($P = 7.5\times10^{-4}$) compared to those with intact *ATM* (**Figure 46B**, left). A similar trend was observed for missense variants (**Figure 46B**, right), and individuals with non-functional eDA scores exhibited a significantly higher lifelong cancer risk (**Figure 47A**). Cumulative breast cancer risk during follow-up and overall lifetime risk were also significantly elevated in this group (**Figure 47B and 47C**). These findings indicate that eDA scores can be used to predict cancer susceptibility in individuals with *ATM* variants.

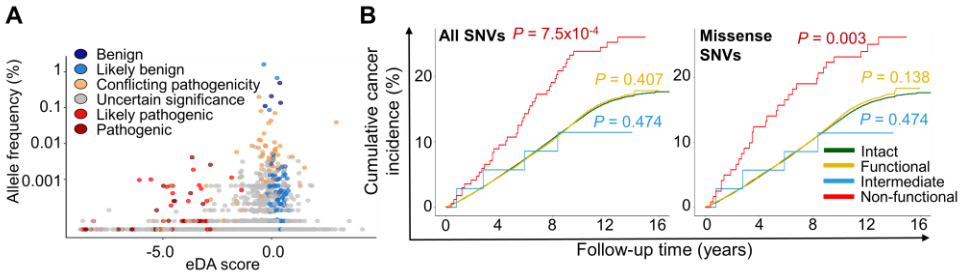


Figure 46. Analyses of clinical databases for unevaluated SNVs using eDA scores. (A) eDA scores plotted against allele frequencies of SNVs in the general population (gnomAD v.4.1 and UK Biobank). ClinVar classifications are shown using different colored dots. **(B)** Cumulative cancer incidence in UKB participants (n = 323,897) with different functional categories of *ATM* variants determined using the eDA score. Experimentally unevaluated variants only were analyzed. The left panel includes participants with all types of SNVs, and the right panel includes only participants with missense SNVs and intact *ATM*. P-values are shown for each group in comparison with the intact *ATM* group.

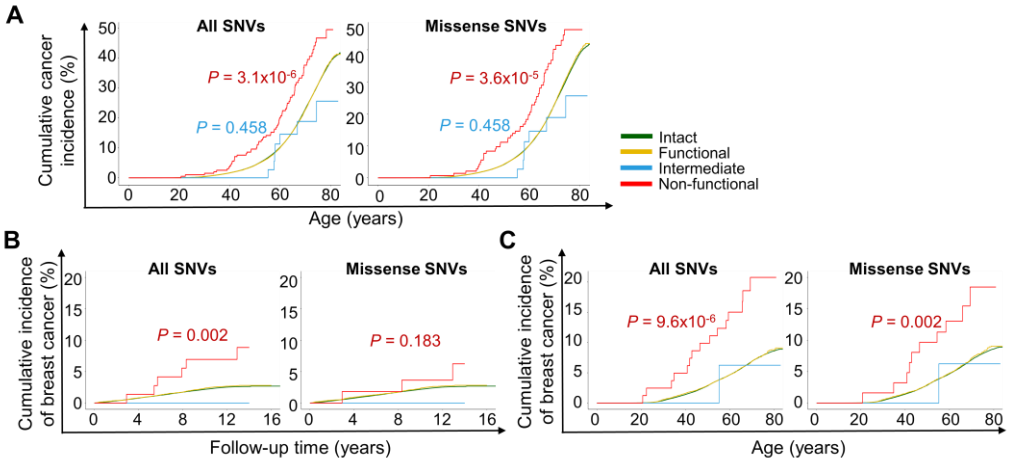


Figure 47. Cumulative cancer incidence in UK Biobank with different functional categories of *ATM* variants determined using the eDA scores. (A) Lifelong cancer incidence, the left panel includes participants with all types of unevaluated SNVs, and the right panel includes only participants with unevaluated missense SNVs and intact *ATM*.
(B) Cumulative breast cancer incidence in female UK Biobank participants, the left panel includes participants with all types of unevaluated SNVs, and the right panel includes only participants with unevaluated missense SNVs and intact *ATM*.
(C) Lifelong breast cancer incidence in female UK Biobank participants, the left panel includes participants with all types of unevaluated SNVs, and the right panel includes only participants with unevaluated missense SNVs and intact *ATM*. The P-value is shown for the non-functional group in comparison with the intact *ATM* group. (intact, green; functional, yellow; intermediate, blue; non-functional, red)

In the GENIE dataset, 698 missense variants that had not been experimentally assessed were classified as non-functional based on their eDA scores. These variants showed an increased odds ratio (OR = 52, P = 1.1×10^{-82}) for cancer occurrence compared to variants classified as non-functional by other models, such as AlphaMissense (**Figure 48A**). Furthermore, odds ratios for non-functional variants across different eDA score thresholds were consistently higher than those observed with previous models (**Figure 48B**), underscoring the clinical utility of eDA scores in evaluating the cancer relevance of *ATM* variants.

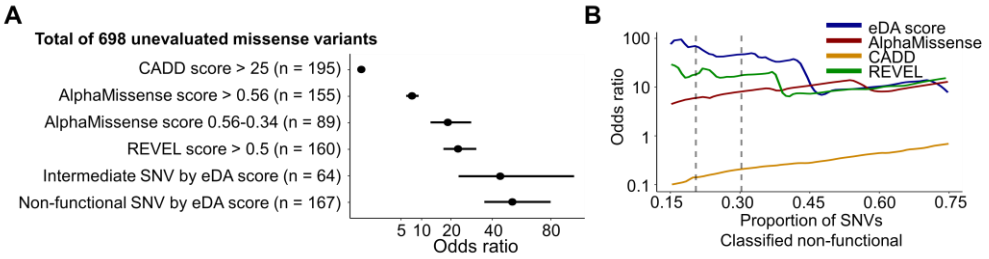


Figure 48. Associations between functional subsets of unevaluated missense variants and their occurrence in tumor samples. (A) Pathogenic variant subsets were determined using the known cutoff values of computational scores calculated by AlphaMissense, REVEL, and CADD, or using our function scores. Odds ratios were calculated by comparing the occurrence of each pathogenic variant subset in tumor samples to that of the benign variant subset. Black bars represent 95% confidence intervals. **(B)** Odds ratios plotted across varying proportions of non-functional SNVs. The variation in the proportions of non-functional SNVs was induced by changing cutoff values for each scoring system. The dashed lines represent the proportions of non-functional SNVs at 20% and 30%, which correspond to the proportions of non-functional missense SNVs in our dataset and the GENIE tumor sequencing data, respectively.

3.8. Complete functional classification of all 27,513 possible *ATM* SNVs

In total, we generated 23,092 function scores and 4,421 eDA scores, covering all 27,513 possible *ATM* SNVs across 62 protein-coding exons. When multiple SNVs resulted in the same single amino acid variant (SAAV) in *ATM*, we derived a representative function or eDA score by averaging the individual scores from those SNVs. These consolidated scores are reported for SAAVs within residues 2,712 to 3,056, including the kinase domain, as illustrated in **Figure 49**. Among non-functional SNVs, 24% were nonsense mutations, while 16% were missense mutations occurring in the kinase domain. In the GENIE dataset, these proportions were 11% and 28%, respectively, while in the bladder cancer patient database, they were 16% and 36%. In the CLL patient database, nonsense mutations comprised 12% and missense mutations 56% of non-functional SNVs.

We further reassessed the clinical significance of the function and eDA scores for all 27,513 SNVs. The combined scores effectively distinguished B/LB variants from P/LP variants (**Figure 50A**). The frequencies of deleterious *ATM* variants in both the general population and cancer samples, along with the odds ratios (ORs) for deleterious variants in cancer samples relative to controls, closely matched those calculated using the 23,092 function scores (**Figure 50B-F**). The ability to predict cancer patient prognoses—such as worse outcomes for patients with deleterious *ATM* variants in CLL and improved outcomes in bladder cancer—showed slight enhancements (**Figure 50G**), likely due to the expanded patient sample size.

Additionally, we assessed cancer risk using UKB data and the combined scores. Individuals carrying non-functional *ATM* variants exhibited an increased risk for both total and breast cancers compared to those with functional *ATM* (**Figures 51A-E**). When analyzing hazard ratios (HRs) across different cancer types, breast cancer (HR = 1.5) and prostate cancer (HR = 1.4) demonstrated the highest HRs among individuals with non-functional *ATM* variants (**Figure 51F**). These findings align with prior studies indicating elevated risks for breast and prostate cancers in individuals with *ATM* variants [8, 10]. Collectively, our results suggest that these combined scores can serve as useful

predictors of cancer prognosis and risk in individuals carrying *ATM* SNVs.

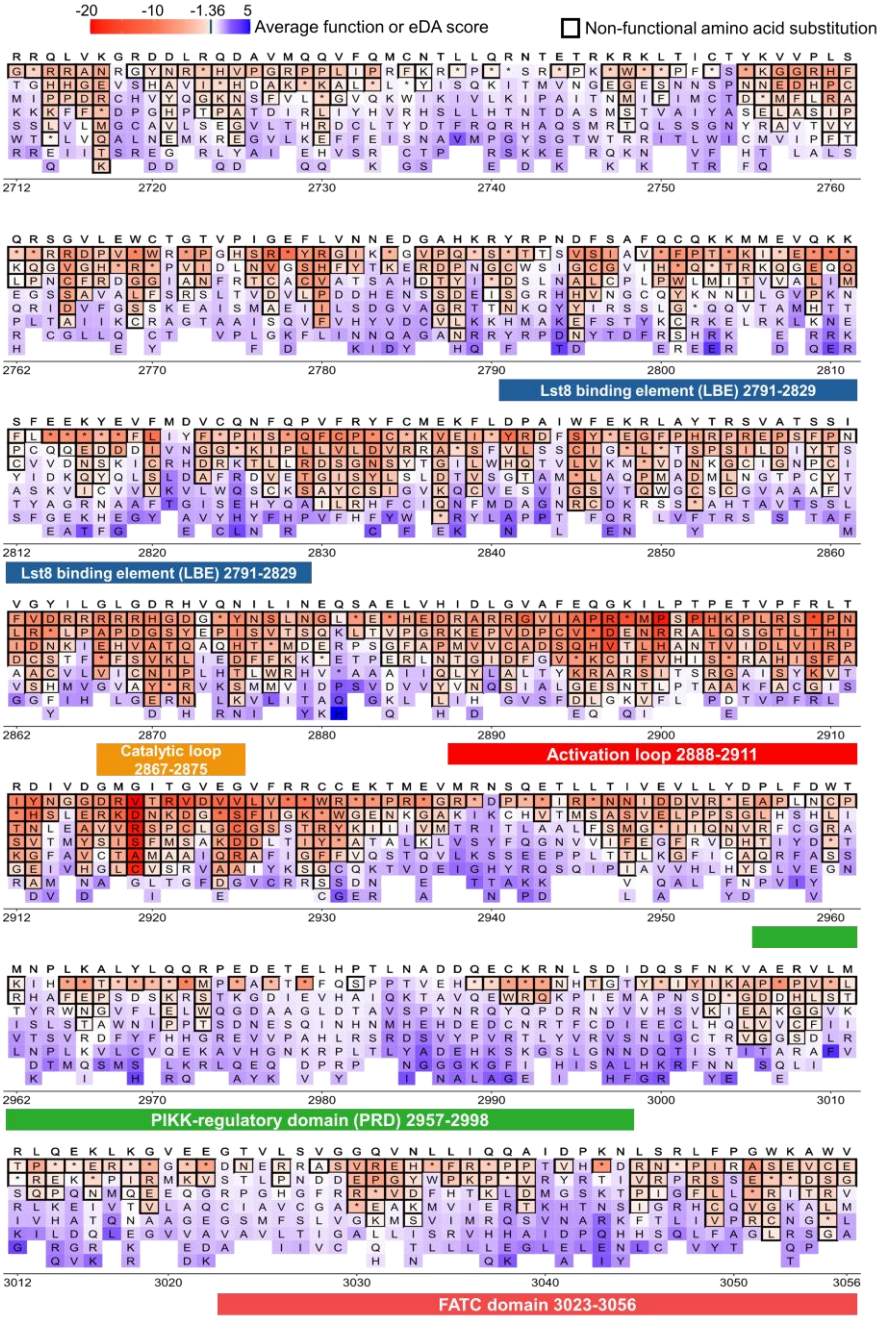


Figure 49. Heatmap showing the functional effects of variants in the kinase domain. Letters within the boxes indicate the amino acid substitutions that have been generated at each position in the reference sequence, which is shown at the top of each segment. Asterisks represent stop codons, and boxes outlined in black indicate non-functional variants. The numbers at the bottom of the heatmap indicate the positions in the amino acid sequence. Functional domains are shown below the amino acid positions. The color spectrum, from red to blue, represents the average function or eDA scores for the single amino acid variants.

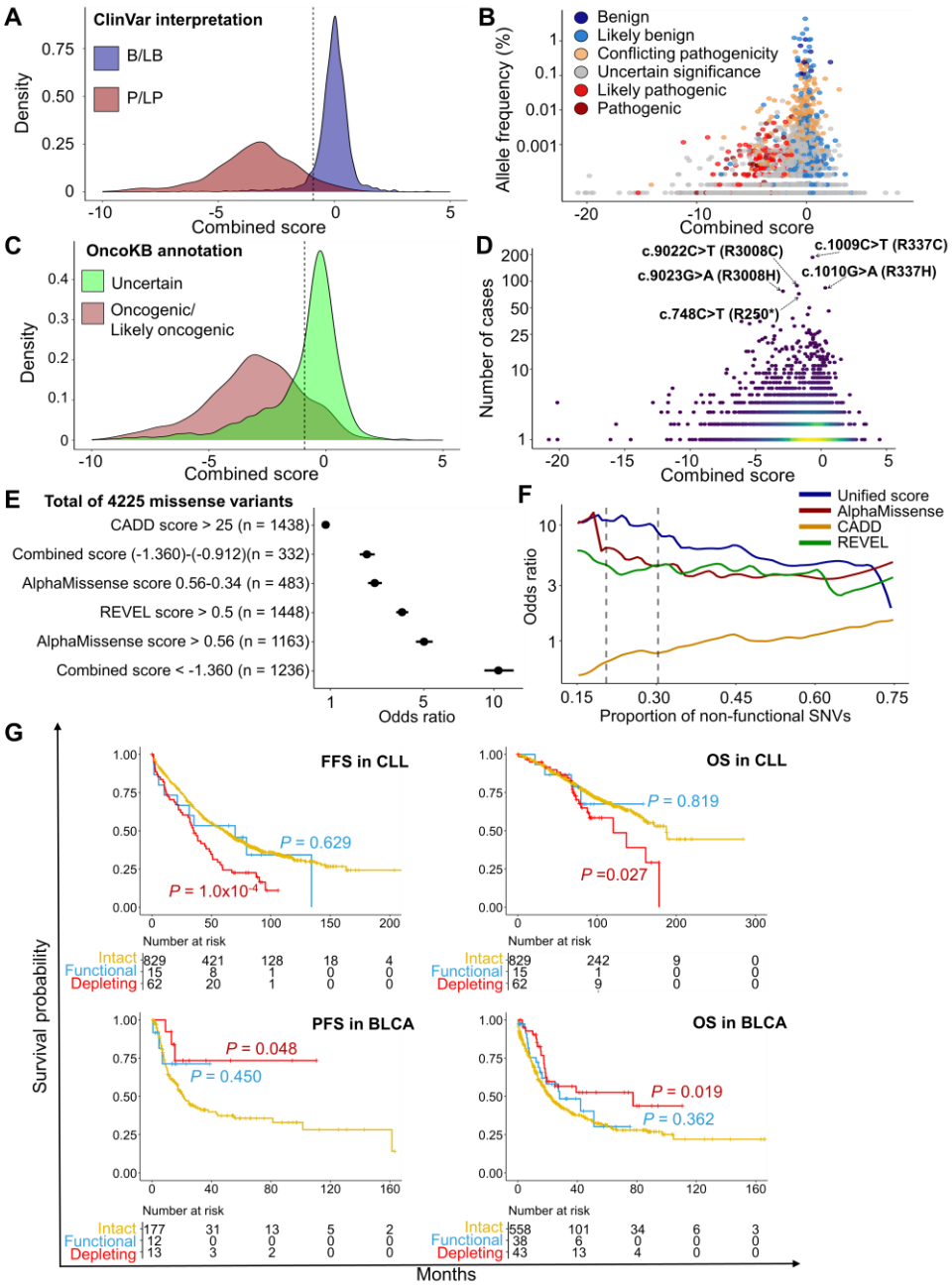


Figure 50. Clinical relevance of combined scores. (A) Kernel density estimation plots of combined scores (function scores for evaluated SNVs and eDA scores for unevaluated SNVs) for all SNVs in the coding sequence reported in ClinVar as P/LP (pathogenic or likely pathogenic) (n = 690), or B/LB (benign or likely benign) (n = 2,560). The cutoff for depleting variants, -0.912, is indicated

with the dashed vertical line. **(B)** Combined scores plotted against allele frequencies of SNVs in the general population (gnomAD v.4.1). ClinVar classifications are shown using different colored dots. **(C)** Kernel density estimate plots of combined scores for all SNVs in the coding sequence (n = 5,250) found in tumor sequencing data, classified by the OncoKB database. The cutoff for depleting variants, -0.912, is indicated with the dashed vertical line. **(D)** Combined scores of SNVs plotted against the number of observations in tumor samples. Four variants most frequently observed observed in tumor samples and a variant with a strong association with breast cancer (c.7271T>G) are shown with arrows. **(E)** Associations between functional subsets of missense variants and their occurrence in tumor samples. Pathogenic variant subsets were determined using the known cutoff values of computational scores calculated by AlphaMissense, REVEL, and CADD, or using our combined scores. Odds ratios were calculated by comparing the occurrence of each pathogenic variant subset in tumor samples to that of the benign variant subset. Black bars represent 95% confidence intervals. **(F)** Odds ratios plotted across varying proportions of non-functional SNVs. The variation in the proportions of non-functional SNVs was induced by changing cutoff values for each scoring system. The dashed lines represent the proportions of non-functional SNVs at 20% and 30%, which correspond to the proportions of non-functional missense SNVs in our dataset and the GENIE tumor sequencing data, respectively. **(G)** Prognosis of cancer patients with different functional categories of *ATM* variants. Patients with chronic lymphocytic leukemia (CLL, the number of patients, n = 906) and those with stage III or IV bladder cancer (n = 639) were categorized into three groups based on the functional classes of their somatic variants: depleting (non-functional + intermediate) variants, functional variants, and wild-type *ATM*. Survival analysis was conducted using the Kaplan-Meier estimator. P-values of survival comparisons between the intact *ATM* group and functional (blue) or depleting groups (red) are shown. FFS, failure-free survival; OS, overall survival; PFS, progression-free survival.

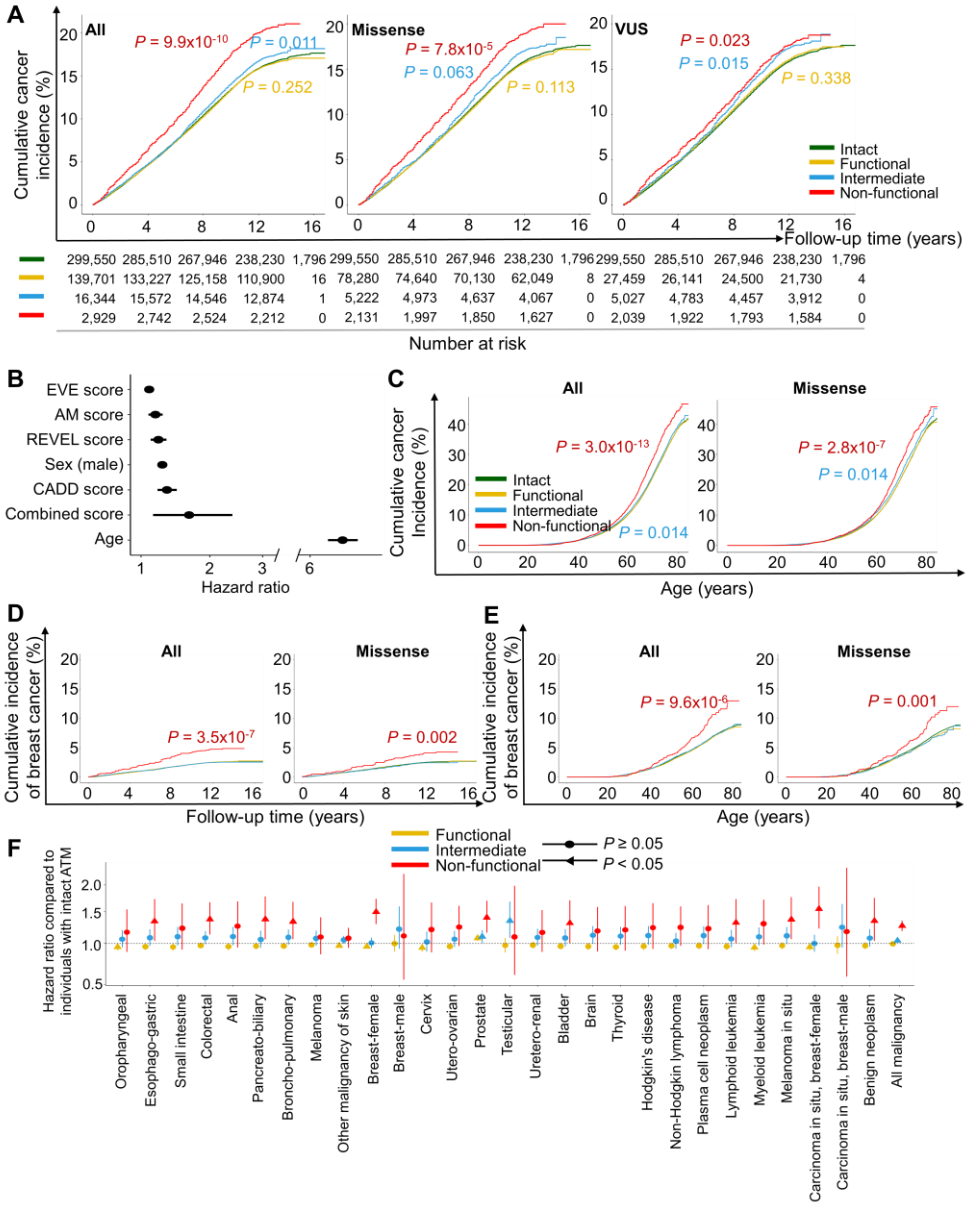


Figure 51. Cancer risks determined using combined scores for all 27,513 possible *ATM* SNVs based on UKB data. (A) Cumulative cancer incidence in UKB participants (n = 458,524) with different functional categories of *ATM* variants determined using the combined score. The left panel includes participants with all types of SNV mutations, the middle panel includes only participants with missense SNVs and intact *ATM*, and the right panel includes only participants with VUSs and intact *ATM*. The numbers of participants are shown below. P-values are shown for each group in comparison with the intact *ATM* group. **(B)** Hazard ratios of cancer incidence for various

computational scores and the combined score. Black bars represent 95% confidence intervals. AM, AlphaMissense. **(C)** Lifelong cancer incidence in UKB participants with different functional categories of *ATM* variants determined using the combined score. The left panel includes participants with all types of SNVs, and the right panel includes only participants with missense SNVs and intact *ATM*. P-values are shown for non-functional (red) and intermediate groups (blue) in comparison with the intact *ATM* group. **(D)** Cumulative breast cancer incidence in female UKB participants with different functional categories of *ATM* variants determined using the combined score. The left panel includes participants with all types of SNVs, and the right panel includes only participants with missense SNVs and intact *ATM*. The P-value is shown for the non-functional group in comparison with the intact *ATM* group. **(E)** Lifelong breast cancer incidence in female UKB participants with different functional categories of *ATM* variants determined using the combined score. The left panel includes participants with all types of SNVs, and the right panel includes only participants with missense SNVs and intact *ATM*. The P-value is shown for the non-functional group in comparison with the intact *ATM* group. **(F)** Hazard ratio of developing various types of cancer in participants with *ATM* variants in each functional category. SNVs were categorized functionally based on the combined score. Hazard ratios were adjusted for the effects of sex and age. Each type of cancer is defined with a set of ICD-10 codes (Methods). Vertical bars represent 95% confidence intervals.

4. DISCUSSION

In our study, we assessed cell fitness in the presence of the PARP inhibitor olaparib. Because PARP is responsible for repairing single-strand breaks (SSBs), its inhibition leads to the accumulation of SSBs, which eventually convert into double-strand breaks (DSBs), increasing the burden on the cell's DSB repair pathway [90, 91]. *ATM* plays a key role in activating DSB repair pathways, such as homologous recombination, and the failure in DSB repair can result in excessive DNA damage, leading to genomic instability, and, in most cases, subsequent cell death [3, 92]. Thus, *ATM* variants depleted in the presence of olaparib are likely impaired in DSB repair. Furthermore, given that *ATM*-directed DSB repair is mainly mediated by the interaction of ATM with NBS1, a member of the MRN (MRE11-RAD50-NBS1) complex, and the kinase activity of ATM, *ATM* variants depleted in the presence of olaparib are likely to exhibit disruptions in the interaction with NBS1 or impaired kinase activity. In addition, ATM phosphorylates Chk2 and p53, leading to cell cycle arrest and the inhibition of uncontrolled cell proliferation [93]. Therefore, *ATM* variants depleted in the presence of olaparib are also likely impaired in cell cycle regulation and the maintenance of genomic stability, possibly increasing cancer risk.

Independent of DNA damage, ATM can be directly activated via oxidation by reactive oxygen species (ROS) [94]. This ROS-directed ATM activation promotes the clearance of toxic protein aggregates [95, 96] and regulates ROS homeostasis [3, 97, 98]. It is unclear whether the *ATM* function measured in our study can be extrapolated to these functions of *ATM* that are independent of the DNA damage response (DDR). Interestingly, R3047X, an *ATM* variant that has intact DDR activity, but impaired ROS homeostasis regulation [94, 99], was classified as non-functional in our study, suggesting that such extrapolation might be possible. However, further research is necessary to draw a more generalized conclusion on this issue.

The importance of ATM autophosphorylation in the mechanism of ATM activation has been debated for decades. Surprisingly, most amino acid substitutions at one of the well-known (the number of substitutions n = 24) and potential autophosphorylation sites (n = 12) (well-known: S367, S1893, S1981, and S2996, potential: T1885 and C2991) [3] showed no depleting effects, with an average function score of -0.20 and 0.48, respectively; two exceptions did exhibit such effects (S1893L and S2996T). Two previous studies suggested that, in response to radiation exposure, *ATM* variants with mutations in one of the autophosphorylation sites (e.g., S367A, S1893A, and S1981A) retained, albeit slightly reduced, protein kinase activity and exhibited autophosphorylation at other autophosphorylation sites (e.g., phosphorylation at S1893 in S1981A-mutant cells) [100, 101]. Other studies even proposed that the primary activation mechanism of ATM is not its autophosphorylation, but its interaction with the MRN (Mre11-Rad50-Nbs1) complex [102-104]. Based on these results, a single missense mutation at one of the autophosphorylation sites may not completely inactivate ATM, due, perhaps, to possible autophosphorylation at other sites. Further research is needed to draw a solid conclusion on this issue.

Five missense variants with discordant ClinVar interpretations as LB or B were classified as non-functional in this study. After a thorough manual review of the evidence and reclassification using ClinGen guidelines [14], we determined that none met the exact criteria for LB or B. As a result, these variants were reclassified as either VUS or LP (Table 5). Furthermore, we observed significant differences in SpliceAI scores synonymous variants between our classifications suggesting that the function score is also reliable for synonymous variant classifications (**Figure 52**).

Two clinical trials failed to establish a solid correlation between deleterious *ATM* mutations and PARP inhibitor responses [17, 105]. If clinical trials were conducted again using the functional classification results provided in our study, it might be possible to draw a more solid conclusion about the correlation between the functional status of *ATM* and the response to PARP inhibitors. Furthermore, our data could also be utilized for planning other clinical trials relevant to *ATM* mutations.

We used a single cell line, which aligns with other representative studies for the functional evaluation of all possible SNVs across entire coding sequences [57, 59, 60]. We cannot rule out the possibility that the functional evaluation results could vary depending on the cell type and genetic background. Although prime editing rarely induces off-target effects [21, 56, 61], we cannot completely rule out the possibility that some functional effects associated with SNVs could be, at least partly, attributable to potential off-target effects, which we did not assess in a high-throughput manner. We used a single readout of cell survival and proliferation in the presence of olaparib for the functional evaluation of *ATM* variants. While we cannot rule out the possibility that the functional evaluation results might differ if a different readout were used, or that variants classified as non-functional might retain some functionality in other processes, our current results align with clinical data. We experimentally evaluated 84% of all possible SNVs. The inability to reach 100% of them is mainly attributable to AT-rich regions that lack the canonical NGG PAM. However, using deep learning, we accurately evaluated the remaining 4,421 variants, which also showed clinical usefulness. We envision that this approach can be expanded to other genes to address the issue of VUSs, enabling precision medicine.

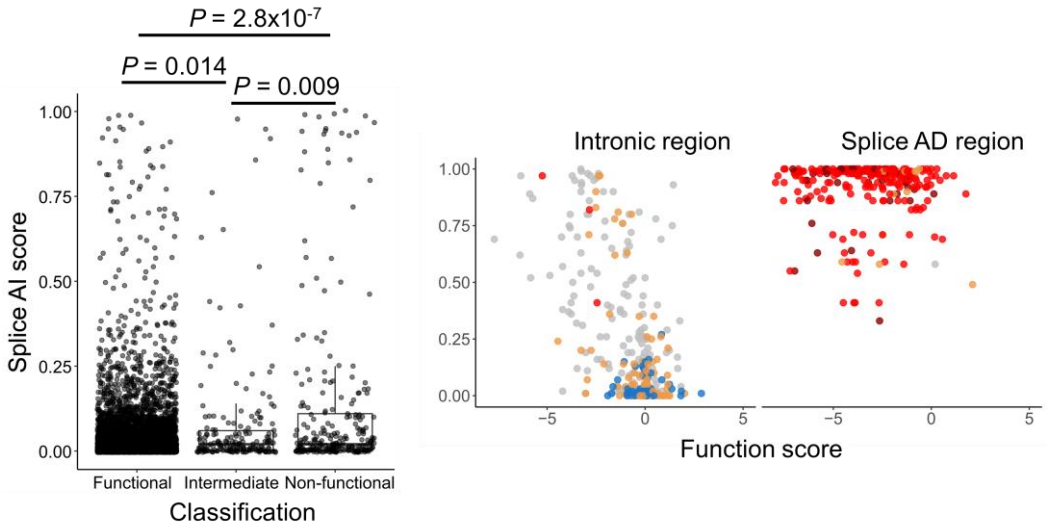


Figure 52. Splice AI score distribution of variants. Left panel shows the splice AI score distribution of synonymous variants for functional classifications, and the right panel shows the splice AI score distributions for intronic variant types.

Table 5. Manual review of discordant variants and reclassification

HGVSc	HGVSp	Type	CADD	REVEL	AM	SpliceAI	GnomAD	ClinVar	Score	Confidence	Star	Evidence	Final
c.987A>G	R329R	Syn	NA	NA	NA	0	0	LB	-1.878	High	2	BP7, PM2, PS3	LP
c.1062C>T	H354H	Syn	NA	NA	NA	0.04	0	LB	-1.592	High	2	BP7, PM2, PS3	LP
c.4551T>C	L1517L	Syn	NA	NA	NA	0.01	0	LB	-1.425	High	2	BP7, PM2, PS3	LP
c.4890C>T	D1630D	Syn	NA	NA	NA	0.32	0	LB	-3.408	High	VUS	BP7, PM2, PS3	LP
c.5013T>A	V1671V	Syn	NA	NA	NA	0.35	0	LB	-1.431	High	1	BP7, PM2, PS3	LP
c.5016A>G	G1672G	Syn	NA	NA	NA	0.25	0	LB	-2.205	High	1	BP7, PM2, PS3	LP
c.5016A>T	G1672G	Syn	NA	NA	NA	0.5	0	LB	-4.773	High	1	BP7, PM2, PS3	LP
c.5019C>T	S1673S	Syn	NA	NA	NA	0.18	0	LB	-3.396	High	2	BP7, PM2, PS3	LP
c.5262G>A	K1754K	Syn	NA	NA	NA	0	0	LB	-1.873	High	2	BP7, PM2, PS3	LP
c.6066T>G	G2022G	Syn	NA	NA	NA	0.28	0	LB	-2.176	High	1	BP7, PM2, PS3	LP
c.6663G>A	E2221E	Syn	NA	NA	NA	0.04	6.84E-07	LB	-1.583	High	2	BP7, PM2_supporting, PS3	VUS
c.7081C>T	L2361L	Syn	NA	NA	NA	0	0	LB	-1.404	High	2	BP7, PM2, PS3	LP
c.8148T>C	V2716V	Syn	NA	NA	NA	0	0	LB	-1.568	High	1	BP7, PM2, PS3	LP
c.8700T>C	L2900L	Syn	NA	NA	NA	0	0	LB	-1.673	High	1	BP7, PM2, PS3	LP
c.8844T>C	I2948I	Syn	NA	NA	NA	0.07	1.86E-06	LB	-1.483	High	2	BP7, PM2_supporting, PS3	VUS
c.9096G>A	V3032V	Syn	NA	NA	NA	0	0	LB	-3.964	High	2	BP7, PM2, PS3	LP
c.319T>C	C107R	Miss	22.5	0.16	0.299	0	6.84E-07	LB	-1.906	Medium-high	1	BP4, PM2_supporting, PS3	VUS
c.569T>A	I190K	Miss	25.8	0.376	0.934	0.05	0	B	-2.58	Medium-high	1	PM2, PS3	LP
c.1462T>A	W488R	Miss	25.8	0.65	0.89	0.04	0	LB	-2.553	Computational	1	PM2	VUS
c.5306C>A	T1769K	Miss	22.9	0.433	0.18	0.29	0	LB	-2.163	Medium-high	1	BP4, PM2, PS3	LP
c.5693G>A	R1898Q	Miss	19.9	0.125	0.075	0.14	0.0001407	B/LB	-3.911	Medium-high	2	BP4, PS3	VUS

5. CONCLUSION

In this study, we used prime editing and deep learning for the functional evaluation of 100% of the 27,513 possible *ATM* SNVs across all 62 protein-coding exons. We envision that the approach used in this study can be expanded for the evaluation of other *ATM* functions that cannot be assessed by measuring cell fitness in the presence of olaparib as well as for the complete functional evaluation of variants in other genes, including those with AT-rich regions in which NGG PAMs are rare. We have experimentally evaluated 62 exons and 23,092 variants in this study, making our analysis larger in scale than other saturation genome editing studies that have analyzed the complete coding sequences of *RAD51C* (9 exons and 9,188 evaluated variants) [59], *BAP1* (17 exons and 18,108 variants) [60], *VHL* (3 exons and 2,268 variants) [57], and *DDX3X* (17 exons and 12,776 variants) [65].

Our functional evaluation results provide clinically useful information, in that they can estimate cancer risk (or identify individuals at high-risk for cancer) and predict the prognosis of cancer patients, which is unprecedented. We envision that our results could be applied to guide the use of PARP inhibitors in cancer patients with *ATM* mutations, although solid conclusions about this issue would require further clinical studies. In addition, our results could be used to diagnose A-T.

REFERENCES

1. Yurgelun, M.B., et al., *Identification of a Variety of Mutations in Cancer Predisposition Genes in Patients With Suspected Lynch Syndrome.* Gastroenterology, 2015. **149**(3): p. 604-13.e20.
2. Balmaña, J., et al., *Conflicting Interpretation of Genetic Variants and Cancer Risk by Commercial Laboratories as Assessed by the Prospective Registry of Multiplex Testing.* J Clin Oncol, 2016. **34**(34): p. 4071-4078.
3. Lee, J.H. and T.T. Paull, *Cellular functions of the protein kinase ATM and their relevance to human disease.* Nat Rev Mol Cell Biol, 2021. **22**(12): p. 796-814.
4. Schon, K., et al., *Genotype, extrapyramidal features, and severity of variant ataxia-telangiectasia.* Ann Neurol, 2019. **85**(2): p. 170-180.
5. van Os, N.J.H., et al., *Classic ataxia-telangiectasia: the phenotype of long-term survivors.* J Neurol, 2020. **267**(3): p. 830-837.
6. Dorling, L., et al., *Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women.* N Engl J Med, 2021. **384**(5): p. 428-439.
7. Hsu, F.C., et al., *Risk of Pancreatic Cancer Among Individuals With Pathogenic Variants in the ATM Gene.* JAMA Oncol, 2021. **7**(11): p. 1664-1668.
8. Karlsson, Q., et al., *Rare Germline Variants in ATM Predispose to Prostate Cancer: A PRACTICAL Consortium Study.* Eur Urol Oncol, 2021. **4**(4): p. 570-579.
9. Loveday, C., et al., *Analysis of rare disruptive germline mutations in 2135 enriched BRCA-negative breast cancers excludes additional high-impact susceptibility genes.* Ann Oncol, 2022. **33**(12): p. 1318-1327.
10. Thompson, D., et al., *Cancer risks and mortality in heterozygous ATM mutation carriers.* J Natl Cancer Inst, 2005. **97**(11): p. 813-22.
11. Daly, M.B., et al., *NCCN Guidelines® Insights: Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic, Version 2.2024.* J Natl Compr Canc Netw, 2023. **21**(10): p. 1000-1010.
12. Feliubadaló, L., et al., *A Collaborative Effort to Define Classification Criteria for ATM Variants in Hereditary Cancer Patients.* Clin Chem, 2021. **67**(3): p. 518-533.
13. Lesueur, F., et al., *First international workshop of the ATM and cancer risk group (4-5 December 2019).* Fam Cancer, 2022. **21**(2): p. 211-227.
14. Richardson, M.E., et al., *Specifications of the ACMG/AMP variant curation guidelines for the analysis of germline ATM sequence variants.* medRxiv, 2024.
15. Bang, Y.J., et al., *Randomized, Double-Blind Phase II Trial With Prospective Classification by ATM Protein Level to Evaluate the Efficacy and Tolerability of Olaparib Plus Paclitaxel in Patients With Recurrent or Metastatic Gastric Cancer.* J Clin Oncol, 2015. **33**(33): p. 3858-65.
16. de Bono, J., et al., *Olaparib for Metastatic Castration-Resistant Prostate Cancer.* N Engl J Med, 2020. **382**(22): p. 2091-2102.
17. Hussain, M., et al., *Survival with Olaparib in Metastatic Castration-Resistant Prostate Cancer.* N Engl J Med, 2020. **383**(24): p. 2345-2357.
18. Austen, B., et al., *Mutations in the ATM gene lead to impaired overall and treatment-free survival that is independent of IGVH mutation status in patients with B-CLL.* Blood, 2005. **106**(9): p. 3175-82.
19. Bueno, R.C., et al., *ATM down-regulation is associated with poor prognosis in sporadic*

breast carcinomas. Ann Oncol, 2014. **25**(1): p. 69-75.
20. Zhou, Y., et al., *ATM deficiency confers specific therapeutic vulnerabilities in bladder cancer.* Sci Adv, 2023. **9**(47): p. eadg2263.
21. Anzalone, A.V., et al., *Search-and-replace genome editing without double-strand breaks or donor DNA.* Nature, 2019. **576**(7785): p. 149-157.
22. Kim, Y., et al., *Saturation profiling of drug-resistant genetic variants using prime editing.* Nat Biotechnol, 2024.
23. Bae, S., J. Park, and J.S. Kim, *Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases.* Bioinformatics, 2014. **30**(10): p. 1473-5.
24. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Res, 2014. **42**(Database issue): p. D980-5.
25. *AACR Project GENIE: Powering Precision Medicine through an International Consortium.* Cancer Discov, 2017. **7**(8): p. 818-831.
26. Chen, S., et al., *A genomic mutational constraint map using variation in 76,156 human genomes.* Nature, 2024. **625**(7993): p. 92-100.
27. Liu, X., X. Jian, and E. Boerwinkle, *dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions.* Hum Mutat, 2011. **32**(8): p. 894-9.
28. Liu, X., et al., *dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs.* Genome Med, 2020. **12**(1): p. 103.
29. Muiños, F., et al., *In silico saturation mutagenesis of cancer genes.* Nature, 2021. **596**(7872): p. 428-432.
30. Cerami, E., et al., *The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.* Cancer Discov, 2012. **2**(5): p. 401-4.
31. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
32. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
33. Shihab, H.A., et al., *Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models.* Hum Mutat, 2013. **34**(1): p. 57-65.
34. Schwarz, J.M., et al., *MutationTaster2: mutation prediction for the deep-sequencing age.* Nat Methods, 2014. **11**(4): p. 361-2.
35. Chun, S. and J.C. Fay, *Identification of deleterious mutations within three human genomes.* Genome Res, 2009. **19**(9): p. 1553-61.
36. Quang, D., Y. Chen, and X. Xie, *DANN: a deep learning approach for annotating the pathogenicity of genetic variants.* Bioinformatics, 2015. **31**(5): p. 761-3.
37. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.
38. Choi, Y., et al., *Predicting the functional effect of amino acid substitutions and indels.* PLoS One, 2012. **7**(10): p. e46688.
39. Ioannidis, N.M., et al., *REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants.* Am J Hum Genet, 2016. **99**(4): p. 877-885.
40. Rentzsch, P., et al., *CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores.* Genome Med, 2021. **13**(1): p. 31.
41. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++.* PLoS Comput Biol, 2010. **6**(12): p. e1001025.

42. Frazer, J., et al., *Disease variant prediction with deep generative models of evolutionary data.* Nature, 2021. **599**(7883): p. 91-95.
43. Brandes, N., et al., *Genome-wide prediction of disease variant effects with a deep protein language model.* Nat Genet, 2023. **55**(9): p. 1512-1522.
44. Cheng, J., et al., *Accurate proteome-wide missense variant effect prediction with AlphaMissense.* Science, 2023. **381**(6664): p. eadg7492.
45. Jaganathan, K., et al., *Predicting Splicing from Primary Sequence with Deep Learning.* Cell, 2019. **176**(3): p. 535-548.e24.
46. Abramson, J., et al., *Accurate structure prediction of biomolecular interactions with AlphaFold 3.* Nature, 2024. **630**(8016): p. 493-500.
47. Skowronski, K., et al., *Genome-wide analysis in human colorectal cancer cells reveals ischemia-mediated expression of motility genes via DNA hypomethylation.* PLoS One, 2014. **9**(7): p. e103243.
48. Tsherniak, A., et al., *Defining a Cancer Dependency Map.* Cell, 2017. **170**(3): p. 564-576.e16.
49. Wang, C., et al., *ATM-Deficient Colorectal Cancer Cells Are Sensitive to the PARP Inhibitor Olaparib.* Transl Oncol, 2017. **10**(2): p. 190-196.
50. Parsons, R., et al., *Hypermutability and mismatch repair deficiency in RER+ tumor cells.* Cell, 1993. **75**(6): p. 1227-36.
51. Chen, P.J., et al., *Enhanced prime editing systems by manipulating cellular determinants of editing outcomes.* Cell, 2021. **184**(22): p. 5635-5652.e29.
52. Li, H., et al., *Functional annotation of variants of the BRCA2 gene via locally haploid human pluripotent stem cells.* Nat Biomed Eng, 2024. **8**(2): p. 165-176.
53. Gilardini Montani, M.S., et al., *ATM-depletion in breast cancer cells confers sensitivity to PARP inhibition.* J Exp Clin Cancer Res, 2013. **32**(1): p. 95.
54. Schmitt, A., et al., *ATM Deficiency Is Associated with Sensitivity to PARP1- and ATR Inhibitors in Lung Adenocarcinoma.* Cancer Res, 2017. **77**(11): p. 3040-3056.
55. Nelson, J.W., et al., *Engineered pegRNAs improve prime editing efficiency.* Nat Biotechnol, 2022. **40**(3): p. 402-410.
56. Yu, G., et al., *Prediction of efficiencies for diverse prime editing systems in multiple cell types.* Cell, 2023. **186**(10): p. 2256-2272.e23.
57. Buckley, M., et al., *Saturation genome editing maps the functional spectrum of pathogenic VHL alleles.* Nat Genet, 2024. **56**(7): p. 1446-1455.
58. Findlay, G.M., et al., *Accurate classification of BRCA1 variants with saturation genome editing.* Nature, 2018. **562**(7726): p. 217-222.
59. Olvera-León, R., et al., *High-resolution functional mapping of RAD51C by saturation genome editing.* Cell, 2024. **187**(20): p. 5719-5734.e19.
60. Waters, A.J., et al., *Saturation genome editing of BAP1 functionally classifies somatic and germline variants.* Nat Genet, 2024. **56**(7): p. 1434-1445.
61. Kim, D.Y., et al., *Unbiased investigation of specificities of prime editing systems in human cells.* Nucleic Acids Res, 2020. **48**(18): p. 10576-10589.
62. Huang, H., et al., *Functional evaluation and clinical classification of BRCA2 variants.* Nature, 2025.
63. Sahu, S., et al., *Saturation genome editing-based clinical classification of BRCA2 variants.* Nature, 2025.
64. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.

65. Radford, E.J., et al., *Saturation genome editing of DDX3X clarifies pathogenicity of germline and somatic variation.* Nat Commun, 2023. **14**(1): p. 7702.
66. Niu, X., et al., *Prime editor-based high-throughput screening reveals functional synonymous mutations in the human genome.* bioRxiv, 2024: p. 2024.06.16.599253.
67. Abou Tayoun, A.N., et al., *Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion.* Hum Mutat, 2018. **39**(11): p. 1517-1524.
68. Warren, C. and N.P. Pavletich, *Structure of the human ATM kinase and mechanism of Nbs1 binding.* Elife, 2022. **11**.
69. Turenne, G.A., et al., *Activation of p53 transcriptional activity requires ATM's kinase domain and multiple N-terminal serine residues of p53.* Oncogene, 2001. **20**(37): p. 5100-10.
70. Howes, A.C., O. Perisic, and R.L. Williams, *Structural insights into the activation of ataxia-telangiectasia mutated by oxidative stress.* Sci Adv, 2023. **9**(39): p. eadi8291.
71. Berliner, J.L. and A.M. Fay, *Risk assessment and genetic counseling for hereditary breast and ovarian cancer: recommendations of the National Society of Genetic Counselors.* J Genet Couns, 2007. **16**(3): p. 241-60.
72. Foulkes, W.D., *Inherited susceptibility to common cancers.* N Engl J Med, 2008. **359**(20): p. 2143-53.
73. Lynch, H.T., et al., *Clinical/genetic features in hereditary breast cancer.* Breast Cancer Res Treat, 1990. **15**(2): p. 63-71.
74. Pharoah, P.D., et al., *Family history and the risk of breast cancer: a systematic review and meta-analysis.* Int J Cancer, 1997. **71**(5): p. 800-9.
75. Lowry, K.P., et al., *Breast Cancer Screening Strategies for Women With ATM, CHEK2, and PALB2 Pathogenic Variants: A Comparative Modeling Analysis.* JAMA Oncol, 2022. **8**(4): p. 587-596.
76. Girard, E., et al., *Familial breast cancer and DNA repair genes: Insights into known and novel susceptibility genes from the GENESIS study, and implications for multigene panel testing.* Int J Cancer, 2019. **144**(8): p. 1962-1974.
77. Hauke, J., et al., *Gene panel testing of 5589 BRCA1/2-negative index patients with breast cancer in a routine diagnostic setting: results of the German Consortium for Hereditary Breast and Ovarian Cancer.* Cancer Med, 2018. **7**(4): p. 1349-1358.
78. Goldgar, D.E., et al., *Rare variants in the ATM gene and risk of breast cancer.* Breast Cancer Res, 2011. **13**(4): p. R73.
79. van Os, N.J., et al., *Health risks for ataxia-telangiectasia mutated heterozygotes: a systematic review, meta-analysis and evidence-based guideline.* Clin Genet, 2016. **90**(2): p. 105-17.
80. Baghaei Vaji, F., et al., *Prognostic significance of ATM mutations in chronic lymphocytic leukemia: A meta-analysis.* Leuk Res, 2021. **111**: p. 106729.
81. Mashima, K., et al., *Characterizing ATM Aberrations in Chronic Lymphocytic Leukemia (CLL): Prognostic Implications and Sensitivity to PARP Inhibition.* Blood, 2023. **142**(Supplement 1): p. 6507-6507.
82. Yi, R., et al., *ATM Mutations Benefit Bladder Cancer Patients Treated With Immune Checkpoint Inhibitors by Acting on the Tumor Immune Microenvironment.* Front Genet, 2020. **11**: p. 933.
83. Knisbacher, B.A., et al., *Molecular map of chronic lymphocytic leukemia and its impact on outcome.* Nat Genet, 2022. **54**(11): p. 1664-1674.
84. Al-Ahmadie, H.A., et al., *Frequent somatic CDH1 loss-of-function mutations in*

plasmacytoid variant bladder cancer. Nat Genet, 2016. **48**(4): p. 356-8.
85. Clinton, T.N., et al., *Genomic heterogeneity as a barrier to precision oncology in urothelial cancer.* Cell Rep, 2022. **41**(12): p. 111859.
86. Guercio, B.J., et al., *Clinical and Genomic Landscape of FGFR3-Altered Urothelial Carcinoma and Treatment Outcomes with Erdafitinib: A Real-World Experience.* Clin Cancer Res, 2023. **29**(22): p. 4586-4595.
87. Iyer, G., et al., *Prevalence and co-occurrence of actionable genomic alterations in high-grade bladder cancer.* J Clin Oncol, 2013. **31**(25): p. 3133-40.
88. Kim, P.H., et al., *Genomic predictors of survival in patients with high-grade urothelial carcinoma of the bladder.* Eur Urol, 2015. **67**(2): p. 198-201.
89. Pietzak, E.J., et al., *Genomic Differences Between "Primary" and "Secondary" Muscle-invasive Bladder Cancer as a Basis for Disparate Outcomes to Cisplatin-based Neoadjuvant Chemotherapy.* Eur Urol, 2019. **75**(2): p. 231-239.
90. Godon, C., et al., *PARP inhibition versus PARP-1 silencing: different outcomes in terms of single-strand break repair and radiation susceptibility.* Nucleic Acids Res, 2008. **36**(13): p. 4454-64.
91. Zheng, F., et al., *Mechanism and current progress of Poly ADP-ribose polymerase (PARP) inhibitors in the treatment of ovarian cancer.* Biomed Pharmacother, 2020. **123**: p. 109661.
92. Patel, A.G., J.N. Sarkaria, and S.H. Kaufmann, *Nonhomologous end joining drives poly(ADP-ribose) polymerase (PARP) inhibitor lethality in homologous recombination-deficient cells.* Proc Natl Acad Sci U S A, 2011. **108**(8): p. 3406-11.
93. Blackford, A.N. and S.P. Jackson, *ATM, ATR, and DNA-PK: The Trinity at the Heart of the DNA Damage Response.* Mol Cell, 2017. **66**(6): p. 801-817.
94. Guo, Z., et al., *ATM activation by oxidative stress.* Science, 2010. **330**(6003): p. 517-21.
95. Lee, J.H., et al., *ATM directs DNA damage responses and proteostasis via genetically separable pathways.* Sci Signal, 2018. **11**(512).
96. Lee, J.H., et al., *Poly-ADP-ribosylation drives loss of protein homeostasis in ATM and Mre11 deficiency.* Mol Cell, 2021. **81**(7): p. 1515-1533.e5.
97. Guo, Q.Q., et al., *ATM-CHK2-Beclin 1 axis promotes autophagy to maintain ROS homeostasis under oxidative stress.* Embo j, 2020. **39**(10): p. e103111.
98. Xie, X., et al., *ATM at the crossroads of reactive oxygen species and autophagy.* Int J Biol Sci, 2021. **17**(12): p. 3080-3090.
99. Milanovic, M., et al., *FATC Domain Deletion Compromises ATM Protein Stability, Blocks Lymphocyte Development, and Promotes Lymphomagenesis.* J Immunol, 2021. **206**(6): p. 1228-1239.
100. Bakkenist, C.J. and M.B. Kastan, *DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation.* Nature, 2003. **421**(6922): p. 499-506.
101. Kozlov, S.V., et al., *Involvement of novel autophosphorylation sites in ATM activation.* Embo j, 2006. **25**(15): p. 3504-14.
102. Daniel, J.A., et al., *Multiple autophosphorylation sites are dispensable for murine ATM activation in vivo.* J Cell Biol, 2008. **183**(5): p. 777-83.
103. Dupré, A., L. Boyer-Chatenet, and J. Gautier, *Two-step activation of ATM by DNA and the Mre11-Rad50-Nbs1 complex.* Nat Struct Mol Biol, 2006. **13**(5): p. 451-7.
104. Lee, J.H. and T.T. Paull, *ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex.* Science, 2005. **308**(5721): p. 551-4.
105. Mateo, J., et al., *Olaparib in patients with metastatic castration-resistant prostate cancer with DNA repair gene aberrations (TOPARP-B): a multicentre, open-label, randomised,*

phase 2 trial. Lancet Oncol, 2020. **21**(1): p. 162-174.

Abstract in Korean

프라임 에디팅 유전자 편집 기술을 활용한 *ATM* 유전자의 대용량 단일염기변이 기능 탐색

ATM 은 63개의 엑손을 가진 대형 유전자로, DNA 손상 반응에서 중요한 역할을 하며, 기능 소실이 암 발생 위험을 증가시키고 암 환자의 예후에 영향을 미친다. 그러나 대부분의 *ATM* 변이가 불확실한 임상적 의미(VUS, Variant of Uncertain Significance)를 가지므로, 그 기능적 영향을 해석하는 것은 여전히 어려운 과제이다. 본 연구에서는 프라임 에디팅(prime editing)과 딥러닝을 활용하여 *ATM*에서 발생할 수 있는 모든 27,513개의 단일 염기 변이(SNV)의 기능을 평가하였다. 반수체화(haploidization)와 PARP 저해제인 올라파립(olaparib)을 이용한 실험을 통해 23,092개의 SNV 를 분석하여 기능적으로 중요한 잔기들을 규명하였다. 또한, 암 유전체 데이터 및 UK Biobank 데이터를 활용하여 본 연구 결과가 암 발생 위험과 예후 예측에 유용함을 확인하였다. 나아가, 딥러닝 모델인 DeepATM 을 개발하여, 나머지 4,421개의 SNV 의 기능적 효과를 높은 정확도로 예측하였다. 본 연구는 *ATM* 변이의 종합적인 기능적 평가를 제공함으로써 정밀의학의 진보를 이루고, 다른 유전자에서의 VUS 문제를 해결하기 위한 틀을 제시한다.

핵심되는 말: 임상적 의미가 불확실한 변이, 프라임 에디팅, 기능적 스크리닝, 포화 유전체 편집

PUBLICATION LIST

Lee KS, Min JG, Cheong Y, Oh HC, Jung SY, Park JI, Song M, Seo JH, Cho SR, Kim HH. Functional assessment of all ATM SNVs using prime editing and deep learning. Cell. 2025. Accepted

Lee KS, Jang J, Jang H, Kang H, Rim JH, Lim JB. Better Prediction of Clinical Outcome with Estimated Glomerular Filtration Rate by CKD-EPI 2021. J Appl Lab Med. 2024 Oct 4:jfae103. doi: 10.1093/jalm/jfae103. Online ahead of print.

Lee KS, Lee YH, Lee SG. Alanine to glycine ratio is a novel predictive biomarker for type 2 diabetes mellitus. Diabetes Obes Metab. 2024 Mar;26(3):980-988. doi: 10.1111/dom.15395. Epub 2023 Dec 11.

Lee KS, Lee CK, Kwon SS, Kwon WS, Park S, Lee ST, Choi JR, Rha SY, Shin S. Clinical relevance of clonal hematopoiesis and its interference in cell-free DNA profiling of patients with gastric cancer. Clin Chem Lab Med. 2023 Jul 13;62(1):178-186. doi: 10.1515/cclm-2023-0261. Print 2024 Jan 26.

Lee KS, Shin DG, Hwang JH, Kim R, Han CH, Yoo J. Construction of a bone marrow report registry using a clinical data warehouse. Int J Lab Hematol. 2022 Aug;44(4):e140-e144. doi: 10.1111/ijlh.13781. Epub 2021 Dec 10.

Lee KS, Lim HJ, Kim K, Park YG, Yoo JW, Yong D. Rapid Bacterial Detection in Urine Using Laser Scattering and Deep Learning Analysis. Microbiol Spectr. 2022 Apr 27;10(2):e0176921. doi: 10.1128/spectrum.01769-21. Epub 2022 Mar 2.

Lee KS, Cho Y, Kim H, Hwang H, Cho JW, Lee YH, Lee SG. Association of Metabolomic Change and Treatment Response in Patients with Non-Alcoholic Fatty Liver Disease. Biomedicines. 2022 May 24;10(6):1216. doi: 10.3390/biomedicines10061216.

Lee KS, Seo J, Lee CK, Shin S, Choi Z, Min S, Yang JH, Kwon WS, Yun W, Park MR, Choi JR, Chung HC, Lee ST, Rha SY. Analytical and Clinical Validation of Cell-Free Circulating Tumor DNA Assay for the Estimation of Tumor Mutational Burden. Clin Chem. 2022 Dec 6;68(12):1519-1528. doi: 10.1093/clinchem/hvac146.

Nam SW, Lee KS, Yang JW, Ko Y, Eisenhut M, Lee KH, Shin JI, Kronbichler A. Understanding the genetics of systemic lupus erythematosus using Bayesian statistics and gene network analysis. Clin Exp Pediatr. 2021 May;64(5):208-222. doi: 10.3345/cep.2020.00633. Epub 2020 Jul 15.

Lee KS, Rim JH, Lee YH, Lee SG, Lim JB, Kim JH. Association of circulating metabolites with incident type 2 diabetes in an obese population from a national cohort. Diabetes Res Clin Pract. 2021 Oct;180:109077. doi: 10.1016/j.diabres.2021.109077. Epub 2021 Sep 29.

Lee KS, Kim D, Lee H, Lee K, Yong D. Isolation of Non-Hydrogen Sulfide-Producing Salmonella enterica Serovar Infantis from a Clinical Sample: the First Case in Korea. Ann Lab Med. 2020 Jul;40(4):334-336. doi: 10.3343/alm.2020.40.4.334.

Lee KS, Kronbichler A, Pereira Vasconcelos DF, Pereira da Silva FR, Ko Y, Oh YS, Eisenhut M, Merkel PA, Jayne D, Amos CI, Siminovitch KA, Rahmattulla C, Lee KH, Shin JI. Genetic Variants in Antineutrophil Cytoplasmic Antibody-Associated Vasculitis: A Bayesian Approach and Systematic Review. J Clin Med. 2019 Feb 21;8(2):266. doi: 10.3390/jcm8020266.

Lee KS, Kronbichler A, Eisenhut M, Lee KH, Shin JI. Cardiovascular involvement in systemic rheumatic diseases: An integrated view for the treating physicians. Autoimmun Rev. 2018 Mar;17(3):201-214. doi: 10.1016/j.autrev.2017.12.001. Epub 2018 Jan 31.

*This dissertation is based on the works of Lee *et al*. (2025)