ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

PREDICTION OF RNA EDITING SITES BY MACHINE LEARNING AND ROLE OF RNA EDITING IN FUNCTIONAL DISEASE MECHANISMS

HÜSEYİN AVNİ TAÇ

PH.D. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Prof. Uğur Özbek

ISTANBUL-2024

ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

PREDICTION OF RNA EDITING SITES BY MACHINE LEARNING AND ROLE OF RNA EDITING IN FUNCTIONAL DISEASE MECHANISMS

HÜSEYİN AVNİ TAÇ

PH.D. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Prof. Uğur Özbek

ISTANBUL-2024

DECLARATION

I declare that this thesis work is my own work, I had no unethical behavior at any stages from the planning to the writing of the thesis, I obtained all the information in this thesis in accordance with academic and ethical rules, I cited all the information and comments that were not obtained with this thesis work, and I provided resources in the list of references. I also declare that there was no violation of any patents and copyrights during the study and writing of this thesis.

Date: 25.07.2024
Hüseyin Avni Taç

PREFACE AND ACKNOWLEDGEMENT

I want to express my gratitude to Prof. Osman Uğur Sezerman, who has served as an outstanding supervisor and mentor, demonstrating remarkable patience, motivation, enthusiasm, and providing exceptional knowledge and guidance. He has consistently offered support and constructive feedback, propelling my career forward from day one. I would also like to thank Prof. Dr. Uğur Özbek, Prof. Dr. Elif Erson for their invaluable evaluation and support.

I am thankful to numerous friends and former members of the Sezermanlab, Berk Gürdamar, Tuğçe Bozkurt, Zehra Hazal Sezer, Tayyip Karaman, and Eray Şahin for their unwavering support throughout this journey.

Finally, I would like to thank my spouse, Ela Taç, to my daughter Dilara Elif Taç and to my son Yiğit Taç for their patience and support throughout my doctoral studies.

TABLE OF CONTENTS

LIST OF ABBREVATIONS AND SYMBOLS

ADAR	Adenosine deaminase acting on RNA
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide
BRCA	Breast invasive carcinoma
dbSNP	Short Genetic Variations database
dsRBDs	Double-stranded RNA-binding domains
FN	False-negative
FP	False positive
GDC	Genomic Data Commons
GSEA	Gene set enrichment analysis
HNSC	Head and neck squamous cell carcinoma
IFN	Interferon
KICH	Kidney chromophobe
KIRP	Kidney renal papillary cell carcinoma
LUAD	Lung adenocarcinoma
MFE	Minimum free energy
ML	Machine learning
MLP	Multi-Layer Perceptron
NCI	The National Cancer Institute
NES	Nuclear export signal
NLS	Nuclear localization signal
RBP	RNA binding protein
RDDSVM	RNA-DNA Differences with Support Vector Machines
RES	RNA editing sites
RF	Random forests
SNV	Single Nucleotide Variant
SVM	Support vector machines
TCGA	The Cancer Genome Atlas
THCA	Thyroid carcinoma
TN	True negative
TP	True positive

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Prediction of RNA Editing Sites by Machine Learning and Role of RNA Editing in Functional Disease Mechanisms

Adenosine to inosine editing plays a crucial role in several biological functions such as gene regulation, alternative splicing, mRNA stability, and is linked to various human diseases. Thus, accurately identifying RNA editing sites holds significant value for both research and medicine. However, computational analysis of RNAseq data often leads to false-positives due to mapping errors. In the initial section of the dissertation, a machine learning approach was introduced, utilizing support vector machines to train on sequence and structural data from previously confirmed editing sites, aiming to predict new RNA editing sites. The classifier demonstrated robust performance on validated data, achieving a sensitivity of 92,8%, a specificity of 77,1%, and an overall accuracy of 90,2%. To make this method widely accessible, an open-source software called RDDSVM was created. The correlation of RNA editing, and gene expression were also investigated in TCGA tumor samples in search for regulatory sites and new cancer specific prognostic RNA editing biomarkers. After determination of regulatory RNA editing sites in BRCA, THCA and LUAD, the impact of the editing frequency of those regulatory sites on the prognosis of the patients were evaluated. Only the RNA editing sites in LUAD tumor samples were found to be important for the prognosis of the patients. EROL1, PSMB2 were novel genes which had multiple regulatory RNA editing sites with editing levels are closely related to prognosis of LUAD patients. The quantification of those single editing events in those genes can be used as prognostic biomarkers in tumor tissue of LUAD patients.

Keywords: RNA editing**,** RNAseq, Machine learning, Support vector machines, Prognostic RNA editing biomarkers

ÖZET

RNA Düzenleme Noktalarının Makine Öğrenimi ile Tahmin Edilmesi ve RNA Düzenlemenin Fonksiyonel Hastalık Mekanizmalarındaki Rolü

Adenozin- inozine RNA düzenlemesi, gen regülasyonu, alternatif uçbirleştirme, mRNA stabilitesi gibi birçok biyolojik işlevde kritik bir rol oynar ve çeşitli hastalıklarla bağlantılıdır. Bu nedenle, RNA düzenleme noktalarını doğru bir şekilde tanımlamak, araştırma ve tıp için büyük önem taşımaktadır. Ancak, RNAseq verilerinin analizi, haritalama hataları nedeniyle yanlış pozitif sonuçlara yol açmaktadır. Tezin ilk bölümünde, daha önce onaylanmış düzenleme noktalarından elde edilen sekans ve yapısal verileriyle kullanan destek vektör makineleri tabanlı bir makine öğrenimi yaklaşımı oluşturulmuştur. Bu yaklaşım yeni RNA düzenleme noktalarını doğru biçimde tahmin etmeyi amaçlamaktadır. Sınıflandırıcı, doğrulanmış veriler üzerinde güçlü bir performans sergileyerek %92,8 duyarlılık, %77,1 özgüllük ve %90,2 genel doğruluk elde etmiştir. Bu yöntemi geniş kitlelere ulaştırabilmek için RDDSVM adında açık kaynaklı bir yazılım geliştirilmiştir. RNA düzenlemesi ve gen ifadesi arasındaki korelasyon kullanılarak, TCGA tümör örneklerinde düzenleyici noktalar ve yeni kanser spesifik prognostik RNA düzenleme biyobelirteçleri araştırılmıştır. BRCA, THCA ve LUAD'da belirlenen düzenleyici RNA düzenleme noktalarının hastaların prognozu üzerindeki etkisi değerlendirilmiştir. Sadece LUAD tümör örneklerindeki RNA düzenleme noktaları hastaların prognozu için önemli bulunmuştur. EROL1, PSMB2 gibi yeni genler, düzenleme seviyeleri LUAD hastalarının prognozuyla yakından ilgili olan birden fazla düzenleyici RNA düzenleme noktasına sahiptir. Bu genlerdeki düzenleme olaylarının niceliklendirilmesi, LUAD hastalarının tümör dokusunda prognostik biyobelirteçler olarak kullanılabilir.

Anahtar sözcükler: RNA düzenlemesi, RNAseq, Makine öğrenmesi, Destek vektör makineleri, Prognostik RNA düzenleme biyobelirteçleri

1 INTRODUCTION AND AIM

1.1 RNA Editing

RNA editing refers to a process occurring after transcription that modifies the genome of organisms by changing the sequence of nucleotides in RNA through insertion, deletion, deamination, or substitution. This phenomenon is present in various types of organisms, including eukaryotes (1), viruses (2), archaea (3), and prokaryotes (4). Mammals exhibit two main forms of RNA editing: one involves the conversion of adenosine to inosine (A-to-I editing) catalyzed by the ADAR (adenosine deaminase acting on RNA) family of enzymes, and the other involves the conversion of cytidines to uridines (C-to-U editing) mediated by the APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide) protein family. Among these, A-to-I editing predominates in humans. In the process of reverse transcription, inosine is interpreted as guanosine by reverse transcriptase, resulting in the incorporation of cytidine into the complementary DNA (cDNA) strand. Consequently, A-to-I editing events are detected as adenosine to guanosine (A-G) alterations during sequencing.

RNA editing plays a crucial role in numerous biological processes such as gene expression and translation (5,6), alternative splicing (7), and mRNA degradation (8). Its significance extends to its association with carcinogenesis (9,10) and various human diseases (11). Therefore, understanding the mechanism of RNA editing and identifying RNA editing sites within the transcriptome hold significant value for both research and medical purposes.

1.2 Prediction of RNA Editing Sites from Sequence Using Support Vector Machines

RNA-seq, a high-throughput sequencing technique of the transcriptome, is highly valuable for identifying RNA editing events in cells under specific conditions. However, computational analysis of RNA-seq data carries significant risks of false positives due to mapping errors arising from factors such as short sequencing reads, polymorphic regions, genomic repeats and duplications, and splice junction sites (11–

14). To address these challenges, numerous studies have employed machine learning approaches, integrating various factors through complex methodologies, to enhance the accuracy of RNA editing detection (15–18).

In the initial section of the thesis, a notably straightforward technique was devised for predicting new A-to-I editing sites by leveraging sequence and structural details extracted from the surrounding sequences of experimentally confirmed A-to-I editing sites. For data extraction from these known editing sites, two distinct approaches were deployed: a structural model and a sequence model. The structural model integrated RNA structure's dot-bracket notation with triplet sequence data, a method previously applied in the computational forecasting of microRNAs (19,20). Conversely, the sequence model relied solely on triplet sequence details (such as the specific arrangements of triplet sequences: AUU, AAU, etc.) without incorporating RNA structural information. These details were processed through support vector machines (SVM), a technique prevalently utilized in bioinformatics and computational biology for training. Using this model, the SVM classifier showed high performance on experimentally verified data providing a sensitivity of 92,8%, specificity of 77,1%, and accuracy of 90,2%.

To the best of our understanding, this is the first case of utilization of triplet sequence elements adjacent to A-to-I editing sites for predicting new A-to-I editing sites in RNA. The approach stands out for its simplicity compared to methods outlined in comparable studies, offering a less computationally intensive solution while maintaining high efficiency. An open-source software named RDDSVM was designed to predict potential A-to-I RNA editing sites.

1.3 Discovery of Regulatory RNA Editing Sites and Prognostic RNA Editing Markers in Cancer

RNA editing plays several roles in numerous biological processes such as gene expression and translation. It is also possible to see some examples of those consequences in carcinogenesis. Recoding of AZIN1 protein (encoding antizyme inhibitor 1) results in higher AZIN1 protein stability and promotes cell proliferation

predisposing to hepatocellular carcinoma (21). In another study, Zhang et al. showed that RNA editing of the miRNA target regions in 3′ UTR of MDM2 gene, can abolish mir-200b mediated repression of the MDM2 mRNA resulting in elevated MDM2 levels in some cancers (22). In prostate cancer, RNA editing of prostate cancer antigen 3 (PCA3), an intronic lncRNA, increases its stability and expression which forms a double-stranded RNA complex with precursor mRNA of prune homolog 2 (PRUNE2). The formation of the PCA3-PRUNE2 complex promotes tumorigenesis in prostate cancer (23).

In the second part of the thesis, the objective was to search for RNA editing sites that regulate gene expression in cancer to identify new cancer specific prognostic RNA editing markers. For this purpose, the correlation of RNA editing frequency and gene expression levels was investigated in tumor samples. The editing frequency data used in this study was obtained from study of Han and colleagues (9) in which RNA editing data was generated from RNAseq results of TCGA (The Cancer Genome Atlas) tumor and non-tumor tissue samples. Out of RNA editing sites with increased editing (referred as differential sites) in tumor samples of BRCA, THCA and LUAD, the sites which were correlated with expression of the closest gene (referred as regulatory sites) were determined. The impact of the editing frequency of those regulatory sites on the survivability of the cancer patients was also evaluated.

1303, 460 and 370 RNA editing sites were found to be statistically correlated with the expression level of matching gene in BRCA, THCA and LUAD respectively. After investigation of the 10 genes with the highest correlation in each cancer type, most of those regulatory sites in LUAD tumors were crucial for the survivability of the LUAD patients (61,7%) compared to very few sites in BRCA (7,8%) and THCA (3,1%) patients. EROL1, PSMB2 were novel genes which have multiple regulatory RNA editing sites closely related to prognosis of LUAD patients. The quantification of those single RNA editing events in those genes can be used as prognostic biomarkers in tumor tissue of LUAD patients and help guide their therapy.

2 BACKGROUND

2.1 Definition and Overview of RNA editing

RNA editing is a post-transcriptional molecular process through which the nucleotide sequence of an RNA molecule is altered, resulting in a different sequence from that encoded by the DNA. This alteration can lead to a variety of functional consequences, including the modification of the encoded protein (5,6), the alteration of RNA stability (8), and the regulation of gene expression (22). RNA editing occurs in a range of organisms, from unicellular protozoans to humans, and involves several different mechanisms (1–3). The two most well-studied types of RNA editing in humans are Adenosine-to-Inosine (A-to-I) editing and Cytidine-to-Uridine (C-to-U) editing.

2.2 Adenosine-to-Inosine (A-to-I) Editing

A-to-I editing is the most common form of RNA editing in higher eukaryotes, particularly in mammals. It is catalyzed by a family of enzymes called adenosine deaminases acting on RNA (ADARs). During A-to-I editing, adenosine (A) residues in double-stranded RNA (dsRNA) regions are deaminated to inosine (I), which is interpreted as guanosine (G) by the cellular machinery (Figure 1). The initial exploration into A-to-I RNA editing was rooted in the analysis of adenosine deamination and the alteration of coding sequences within the GRIA2 transcript of mammals, which codes for GluA2, a component of the AMPA-type glutamate receptor (24). Through the comparison of genomic DNA and cDNA from mouse brains, Melcher and his team discovered that the GRIA2 gene undergoes a modification from A to G in its mRNA form (25). This alteration leads to a switch in the codons, transforming glutamine to arginine at the affected site. Subsequent research established that these changes were not the result of mutations in the DNA, but rather the consequence of RNA editing processes.

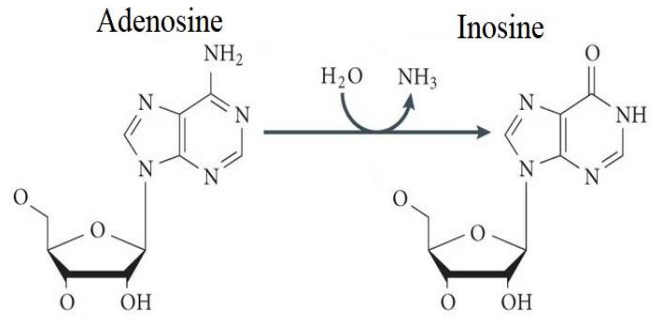

Figure 1. Deamination of Adenosine to inosine

A-to-I editing can lead to changes in the amino acid sequence of encoded proteins, affect splicing, and influence RNA stability and localization. A-to-I editing plays a crucial role in the brain and has been implicated in various neurological conditions (11,26).

2.3 Cytidine-to-Uridine (C-to-U) Editing

C-to-U editing involves the deamination of cytidine (C) to uridine (U) and is less common than A-to-I editing. It is catalyzed by a different set of enzymes, notably the APOBEC family (APOBEC1, APOBEC2, APOBEC3A-H, APOBEC4, and activation-induced cytosine deaminase -AICDA-). C-to-U editing can result in changes to the protein-coding sequence of mRNAs and is essential for the generation of diversity in certain proteins, such as the apolipoprotein B (APOB) in mammals (27). These enzymes are expressed in immune cells such as macrophages, monocytes, and natural killer (NK) cells in response to conditions of low oxygen (hypoxia) and stimulation by interferons (IFNs), playing a significant role in the immune response (28). Evidence also exists of C-to-U RNA processing occurring in the mitochondria of plants, indicating that this function holds biological significance beyond mammals (29). APOBEC1 was initially discovered due to its role in editing the mRNA of apolipoprotein B (ApoB), leading to the production of two different protein versions: one truncated and one full-length. The full-length version of ApoB is involved in cholesterol transportation, whereas the truncated version plays a role in transporting triglycerides in the bloodstream. Like the function of ADAR enzymes, APOBEC family proteins predominantly act on noncoding and intronic areas of transcripts that

include Alu elements. Furthermore, APOBEC proteins are known to combat retroviruses, endogenous retroelements, and other viral entities, with their antiviral actions being either linked to or separate from their capacity to edit RNA (30,31). Since C-to-U editing is less common than A-to-I editing, it will not be discussed here in detail.

2.4 ADAR Protein Family

In humans, the ADAR family comprises three distinct proteins: ADAR1 (produced by the ADAR gene), ADAR2 (produced by the ADARB1 gene), and ADAR3 (produced by the ADARB2 gene). ADAR1 exists in two forms, ADAR1p150 and ADAR1p110. ADAR1 and ADAR2 possess enzymatic activity, enabling them to perform RNA editing. In contrast, ADAR3, the third member of the ADAR family, lacks catalytic activity. Despite this, ADAR3, a brain-specific dsRNA binding protein, has been demonstrated to inhibit RNA editing carried out by ADAR1 and ADAR2. This suggests that ADAR3 may play a regulatory role in the RNA editing process (32,33).

Proteins in the ADAR family share a high degree of conservation and exhibit a comparable arrangement of domains. The catalytic deaminase domain is in the C-terminal region of these proteins (Figure 2). Of the family, only ADAR1 and ADAR2 possess active deaminase domains, whereas ADAR3 does not have a functional deaminase domain (34). Located upstream of the catalytic domain, there are two to three double-stranded RNA-binding domains (dsRBDs) in these proteins, highlighting that ADARs' attachment to RNA is guided by the double-stranded RNA (dsRNA) structure rather than a specific sequence. The p110 isoform of ADAR1 is universally present across major human tissues, whereas the longer ADAR p150 isoform is produced in response to activation by an interferon (IFN)-inducible promoter. The ADAR p150 isoform distinguishes itself with an extended N-terminal region, which functions as the nuclear export signal (NES) and a Z-DNA binding domain. While both ADAR1 p110 and ADAR1 p150 isoforms have the capacity to shuttle between the nucleus and the cytoplasm, the p110 isoform is mainly found in the nucleus and the p150 isoform is typically located in the cytoplasm.

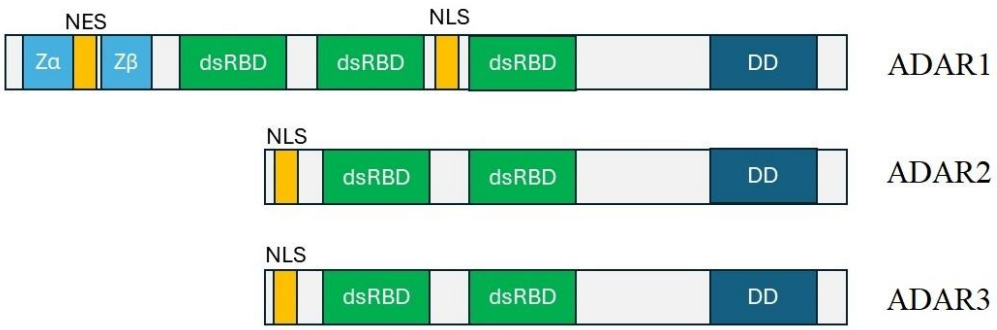

Figure 2. Structure of ADAR family proteins.

They feature a catalytic deaminase domain (DD) at their C-terminus, alongside two or three double-stranded RNA-binding domains (dsRBDs) located in the N-terminal region. Additionally, ADAR1 possesses an N-terminal Zα domain housing a nuclear export signal (NES) and a Zβ domain. Furthermore, a nuclear localization signal is present in these proteins.

ADAR2 is an essential protein capable of moving between the nucleolus and nucleoplasm, driven by its expression levels and interaction with substrates. It also self-regulates both its expression and activity by editing its own pre-mRNA, which introduces 3’ splice sites leading to a decrease in the production of functional ADAR2 (35).

ADAR3, the family's third member, is exclusively found in specific brain regions (36). It has a 72% sequence similarity with ADAR2 and features two double-stranded RNA-binding motifs along with a deaminase catalytic domain. Despite the absence of RNA editing activity in ADAR3, it likely acts as a double-stranded RNA-binding protein. This role enables ADAR3 to influence the interaction of the other ADAR proteins with double-stranded RNA, thus impacting the process of RNA editing (37).

2.5 Molecular Functions of A-to-I Editing

ADAR proteins target adenosine for deamination within double-stranded RNA (dsRNA) or the A-form helix of DNA: RNA hybrids. This action introduces non-standard base pairs into the dsRNA structure, adding complexity and influencing the

adenosine editing efficiency. The two bases next to the edited adenosine in the RNA sequence play an important role in the efficiency of RNA editing (38). Specifically, editing is more probable when the target adenosine is flanked by a pyrimidine before it and a guanine (G) after it, such as in the sequence 5'-UAG-3' (39). Furthermore, mismatches at the editing sites can enhance processing efficiency. While classical A-U pairings can undergo editing, mismatches involving A-C are found to be the most conducive to efficient adenosine editing, in contrast to A-A or A-G mismatches, which are less effective. Thus, both the specific neighboring nucleotides to the edited adenosine and the occurrence of an A-C mismatch at the site of editing are critical factors that enhance the deamination efficiency of dsRNA by ADAR enzymes.

RNA editing is pivotal in creating diversity among protein isoforms and their functions by modifying mRNA coding sequences. Yet, in the protein-coding regions of mammalian transcripts, the number of edits is relatively limited, with only 40 conserved editing sites identified (40). Conversely, the bulk of A-to-I RNA editing activities are found in noncoding RNAs, introns, and 3' untranslated regions (UTRs) in humans (41,42). A fascinating element of ADAR1's biological role is its interaction with Alu elements, which originate from the duplication of short interspersed nuclear elements (SINEs) in humans (43). Alu elements, comprising two repeats facing in opposite directions, create a double-stranded hairpin configuration that is frequently subjected to RNA editing (44). While Alu elements are mainly located in introns and untranslated regions (UTRs) of genes, they can also appear within coding sequences. ADAR enzymes have the ability to identify and bind to these RNA hairpin structures, facilitating the conversion of adenosine to inosine via hydrolytic deamination (37).

Cotranslational A-to-I editing events have the potential to cause intron retention or the creation of new exons, as the elimination of splice sites can lead to the production of various protein isoforms. Specifically, A-to-I editing within intronic Alu elements may change the standard 5' splice donor site from AU to IU and/or the standard 3' splice acceptor site from AA to IA, which could influence the splicing process of transcripts (45,46).

A-to-I editing contributes to the regulation of RNA structural stability as well. The substitution of adenosine (A) with inosine (I) in an A-U base pair decreases stability, whereas the insertion of inosine in an A-C mismatch enhances stability (47). Wone and colleagues showed that A-to-I editing in the 3' untranslated region (UTR) of the stearoyl-CoA desaturase (SCD1) gene led to an increase in KHDRBS1 protein binding, which in turn, improved the stability of SCD1 mRNA (48). ADAR2 also binds to and stabilizes RNA in a manner that is independent of editing.

RNA editing is also present in microRNAs (miRNAs), where primary miRNAs in the nucleus are processed into miRNAs through the formation of hairpin double-stranded RNA structures, which are targets for ADAR enzymes (49). A-to-I editing can influence several aspects of miRNA biogenesis, from competing for binding with miRNA precursors to interacting with RNA interference (RNAi) processing components, thereby affecting the efficiency of miRNA maturation (50). Moreover, A-to-I editing can modify RNAi activity by changing the miRNA-binding sites on mRNA and altering miRNA specificity. Long non-coding RNAs (lncRNAs), capable of forming double-stranded RNA structures, serve as potential substrates for ADARs. A-to-I editing can affect the stability and functionality of lncRNAs by modifying their structure and interactions with miRNAs, indicating a link between RNA editing and gene silencing mechanisms that regulate gene expression (42).

2.6 Biological Functions of RNA Editing

In HEK293 cells, the deletion of ADAR genes triggers interferon production, believed to be a response to endogenous double-stranded RNA (dsRNA) that the cell's machinery mistakes for viral genetic material (51). The complete absence of ADAR1 or ADAR2 results in spontaneous lethality (52), with ADAR1-deficient mice perishing around embryonic day 12.5, and ADAR2-deficient mice dying at or shortly after birth, around postnatal day 20 (53). ADAR2 plays a vital role in editing the neuronal glutamate receptor Gria2 mRNA, with only the edited version of Gria2 pre-mRNA undergoing efficient splicing and export from the nucleus. The survival of mice with inserted sequences that mimic edited Gria2 indicates the critical nature of Gria2 editing for survival, underscoring ADAR2's essential role (24). Mice lacking ADAR3

(ADAR3-KO) generally appeared normal but exhibited cognitive impairments, specifically in learning and memory functions, even though the overall levels of A-to-I editing in the genome did not significantly alter (54).

2.7 RNA Editing in Immunity

ADAR1 is recognized as a crucial element in the innate immune mechanism, modifying host RNAs through A-to-I editing (55). This process helps prevent inappropriate interferon (IFN) signaling and the detection of endogenous double-stranded RNA (dsRNA) by pattern recognition receptors (PRRs), potentially averting pathogenic responses (56). The mitochondrial antiviral-signaling protein (MAVS) is essential for signaling pathways that uphold immune equilibrium and antiviral activities. PRRs, including RIG-I, MDA5, and toll-like receptors (TLRs), detect dsRNA or single-stranded RNA (ssRNA) and prompt the oligomerization of MAVS (57). This action initiates downstream factors, leading to an interferon response. These receptors identify distinct RNA structural motifs, with RIG-I targeting dsRNA ends and MDA5 focusing on internal RNA duplexes. A-to-I editing alters A-U base pairs in dsRNA to I-U mismatches, disrupting the RNA duplex structure. This indicates that ADAR1, through its influence on MDA5 polymerization to dsRNA and MAVS activation, plays a vital role in moderating IFN production and curbing inappropriate innate immune reactions (53,55).

It is believed that isoforms of p150 induced by interferon contribute to a signaling pathway, possibly by modifying certain p150 targets obtained via their presence in the cytoplasm and/or their ability to bind to Z-DNA/RNA domains (58,59). Evidence supporting this includes observations that the lethal effect on embryos caused by the deletion of ADAR1 can be mitigated by additionally knocking out MAVS or MDA5/Ifih1. This suggests that ADAR1 is crucial for initiating the double-stranded RNA (dsRNA) recognition pathway, highlighting its importance in innate immune responses.

Mutations in ADAR1 are linked to Aicardi-Goutières syndrome (AGS), a severe autoinflammatory condition characterized by irregular production of interferon (IFN)

(60). Beyond suppressing the MDA5/MAVS pathway, ADAR1 also influences the activity of protein kinase R (PKR). As another sensor for double-stranded RNA (dsRNA) crucial in antiviral defenses, PKR activates a response that halts protein synthesis and leads to cell death by phosphorylating eukaryotic initiation factor 2α (eIF2α) (61,62). Immune tolerance is established when cells label immunogenic double-stranded RNA (dsRNA) as "self" by converting adenosine to inosine in RNA sequences, which helps in avoiding an overly aggressive immune reaction.

2.8 RNA Editing in Cancer

While the connection between RNA editing and cancer has been recognized for some time, the exact role it plays in the development of cancer remains unclear. Various cancers have shown abnormal levels of ADARs, including instances of heightened RNA editing due to the overexpression of RNA editing enzymes, as well as cases of diminished RNA editing stemming from their reduced expression (63,64). In certain cancers, the relationship between RNA editing levels and the presence of RNA editing enzymes is not straightforward and the impact and relevance of RNA editing differ across cancer types (65). Current insights indicate that RNA editing significantly influences cancer development through a variety of mechanisms.

Variations in RNA editing levels can result in changes to protein sequences and the buildup of improperly expressed proteins, which may encourage cancer progression. Consequently, the role of RNA editing could be comparable in clinical significance to the accumulation of DNA mutations found in tumors with a high mutational load. The DNA damage response is essential in preventing cells from becoming cancerous, with the DNA repair mechanism being vital for preserving cellular balance. Notably, RNA editing has been detected in the RNA transcripts of genes that are part of the DNA repair systems, indicating that RNA editing might play a role in the early phases of cancer development.

Many research findings have highlighted the effects of ADAR-mediated RNA editing on cellular growth. There are several examples of those incidents. Elevated expression of the ADAR gene in non-small cell lung cancer enhances A-to-I RNA

editing at the K_{12} site of NEIL1, altering the codon from arginine to lysine at position 242, a phenomenon also noted in myeloma (66). Such mutations at this site weaken the cell's capacity for repairing DNA damage induced by oxidative stress. Elevated ADAR1 activity leads to increased RNA editing in antizyme inhibitor 1 (AZIN1), which results in the creation of a variant AZIN1 protein with a serine to glycine substitution at position 376 (S376G), affecting cell growth (67). In myeloma, excessive activity of ADAR1 causes the R701G mutation by editing the RNA of Glioma-associated oncogene 1 (GLI1), influencing cancer cell growth and their resistance to chemotherapy (68). While GRIA2 mRNA typically undergoes RNA editing by ADAR2, a decrease in this editing within glioblastoma correlates with reduced ADAR2 levels and altered GluA2 receptor functionality.

Besides A-to-I RNA editing, C-to-U RNA editing is also crucial in driving the proliferation of cancer cells. Elevated levels of APOBEC1 result in the editing of novel APOBEC1 target number 1 (NAT-1) mRNA, leading to a reduction in NAT-1 protein levels (31). This change impacts cell cycle regulation. Apart from changes in coding regions, RNA editing can also result in changes in non-coding regions which is important in carcinogenesis. Zhang et al. showed that RNA editing of the miRNA target regions in 3′ UTR of MDM2 gene, can abolish mir-200b mediated repression of the MDM2 mRNA resulting in elevated MDM2 levels in some cancers. (22). In prostate cancer, RNA editing of prostate cancer antigen 3 (PCA3), an intronic lncRNA, increases its stability and expression which forms a double-stranded RNA complex with precursor mRNA of prune homolog 2 (PRUNE2). The formation of the PCA3-PRUNE2 complex promotes tumorigenesis in prostate cancer (23). While numerous reports have highlighted the link between RNA editing and cancer, the exact mechanisms by which RNA editing contributes to cancer remain largely uncharted. RNA editing is also known to broaden the array of tumor antigens on cancer cells, making them identifiable to the immune system. Additionally, RNA editing is implicated in cancer advancement, promoting tumor growth, spread, immune system avoidance, and the formation of secondary tumors. As such, RNA editing presents a promising avenue for cancer treatment strategies.

2.9 Detection of RNA-Editing Events From RNA-Seq Data and Current Challenges

RNA editing can be identified by comparing RNA-seq sequencing reads with the reference genome, focusing on A-to-G nucleotide alterations that occur exclusively in the RNA sequence and not in the genomic DNA. Identifying A-to-G discrepancies compared to the reference genome may seem simple, but it's essential to apply a series of filters to eliminate biases. These biases can originate from the sequencing process, such as sequencing errors, incorrect alignment of reads, and PCR duplicates created during library preparation, which could lead to incorrect identification of RNA-editing sites, as well as from genomic differences like single-nucleotide polymorphisms (11–14).

Furthermore, the Short Genetic Variations database (dbSNP), which serves as a repository for sequence variations and is a key resource for distinguishing RNA-editing sites from known SNPs, includes sequence variations from both genomic DNAs and cDNAs, thus nucleotide changes from cDNA templates may represent genuine RNA editing sites. The complexity in the computational workflow for identifying RNA-editing sites is exacerbated by the likelihood of errors and the significant demand for computational resources. Moreover, a standard method for normalizing identified RNA-editing sites has yet to be established, complicating the comparison of the quantity of RNA-editing sites identified between different samples.

The precision of both the count of RNA editing locations and the degree of penetrance (the ratio of edited messages at a specific site) is closely linked to the sequencing depth and the coverage of the input RNA-Seq. A higher density of reads enhances the sensitivity for identifying editing sites and provides stronger statistical backing for the predictions (69). For instance, conducting extremely deep sequencing (over 5000 reads) on chosen Alu elements in humans demonstrated that nearly every adenosine within the Alu regions of primates is susceptible to editing (70). The consistency of read coverage across entire transcripts also influences the identification of editing events: certain genomic regions that are expressed may receive more comprehensive coverage than others due to varying expression levels, biases

introduced during sequencing and mapping, or incomplete annotations of transcripts. In situations where sequencing coverage is inadequate or where there is significant deviation from uniform coverage of transcripts, having access to both biological and technical replicates can serve as a useful strategy (71,72).

The methods used in library preparation and sequencing can impact the precision with which reads are mapped to the reference genome. Over recent years, there have been significant advancements in sequencing technologies, and the typical length of reads has steadily increased. For example, with Illumina technology, reads usually fall between 50 and 300 nucleotides in length. Although read length has not been shown to influence the analysis of differential gene expression significantly, it plays a crucial role in identifying RNA editing (73). Specifically, longer reads (100 nucleotides or more) enhance the genome's mappability, are essential for accurately mapping exon-exon junctions, and aid in identifying reads that are extensively edited. Additionally, longer reads are beneficial for distinguishing between functional gene structures and pseudogenes, which might otherwise obscure the detection of true editing events (74,75).

Another important consideration in the planning of RNA-Seq experiments for the detection of editing involves choosing between single-end or paired-end sequencing modes. Single-end sequencing is more budget-friendly, yet it decreases the chance of uniquely mapping reads. Therefore, paired-end sequencing is favored because it effectively extends the length of reads, enhances the accuracy of mapping, and as a result, diminishes the likelihood of incorrectly identifying RNA editing events.

Furthermore, conventional RNA-Seq methods do not preserve the directionality of the original transcripts, making them less effective for detecting and measuring RNA editing in genomic areas where overlapping transcripts come from opposite directions. In contrast, RNA-Seq data obtained through strand-specific techniques allow for more accurate alignments and aid in pinpointing the precise genomic locations of edits (76).

In conclusion, RNA-Seq experiments that utilize long reads and strand-specific paired-end sequencing provide an effective foundation for accurately identifying RNA editing occurrences. This approach ensures more reliable genome alignments and reduces the number of reads mapping to multiple locations. The accuracy of RNA editing predictions improves with greater sequencing depth, making it advisable to conduct experiments where the RNA is sequenced extensively. For Illumina sequencing specifically, RNA-Seq datasets comprising 80–100 million paired-end reads are typically suitable for RNA editing profiling.

2.9.1 RNA editing detection tools

RNA editing sites (RES) are detectable by comparing RNA sequencing (RNA-seq) with DNA sequencing (DNA-seq) data (77). This method reveals RES as either adenosine to guanine (A > G) or cytidine to thymidine (C > T) mismatches, corresponding to A-to-I or C-to-U RNA editing processes, respectively (69). However, the necessity for corresponding RNA and DNA sequencing data from the same cell or tissue samples makes this method potentially impractical and costly. To overcome these challenges, computational methods have been devised to identify RES solely from RNA-seq data. To accurately pinpoint RES through these means, it is essential to incorporate various filters in the analysis pipeline to diminish the rate of false positives. These filters aim to weed out sequencing errors and inherent genetic variations, like single nucleotide polymorphisms (SNPs) (78). Furthermore, employing strand-specific RNA-seq techniques is advised because they enable more precise identification of RES in areas where transcripts from opposite strands overlap (76). Here, a few major examples of those RNA editing detection tools will be discussed.

REDItools, the pioneering tool for RNA editing detection, was introduced to the field for identifying both known RNA Editing Sites (RES) catalogued in RNA editing databases, such as REDIportal, and for the de novo discovery of RES without prior RNA editing knowledge (79). It can perform these tasks using solely RNA-seq data or by contrasting RNA-seq with DNA-seq data. In scenarios where corresponding DNA-seq data is unavailable, it compares variants against a benchmark substitution

distribution, reporting only those positions that are statistically significant. The tool operates on pre-aligned reads in BAM format. It incorporates various filters and quality controls to exclude positions based on several criteria, including base quality, mapping quality scores, coverage at the position, and the exclusion of substitutions in areas of homopolymer repeats. The creators of REDItools enhanced the original tool by developing REDItools2, an advanced version that operates more efficiently across multiple nodes in a parallel manner (80). REDItools2 offers two analysis modes: a serial mode and a parallel mode, the latter of which necessitates the setup of a Message Passing Interface (MPI) implementation.

GIREMI utilizes RNA-seq data exclusively to detect RES by focusing on allelic linkage, the principle where genetic markers in proximity on a chromosome tend to be inherited together during the chromosomal crossover in meiosis (81). The method assumes that SNPs close to each other should share the same haplotype, whereas RES might exhibit variation in relation to adjacent SNPs across different reads. Leveraging this concept, GIREMI calculates the mutual information between known SNP sites from public databases -like the dbSNP database- (82) and unidentified RNA variants within the RNA-seq sample. To improve its accuracy, the method also incorporates a generalized linear model. For input, GIREMI requires BAM files and a curated list of single nucleotide variants (SNVs), including both publicly available SNVs and those identified within the RNA-seq samples, necessitating users to undertake a genotype-calling process.

In response to the growing need for longer reads to span the entire transcriptome, as facilitated by technologies from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), the same research team introduced L-GIREMI. This new technique adopts a similar strategy to GIREMI for RES prediction from long-read RNA-seq data, catering to the advancements in sequencing technology (83).

RES-Scanner facilitates the comprehensive detection and annotation of RES across various species (84). However, its application is constrained by the necessity for corresponding RNA- and DNA-seq data from the same sample, which may not always be feasible. The tool is capable of annotating RES if annotation files containing

pertinent features are available. Employing a suite of statistical models, including Bayesian, binomial, and frequency-based approaches, RES-Scanner excels in calling homozygous genotypes and employs a range of filters such as mapping quality, duplicate removal, and read depth considerations to mitigate the risk of false-positive RES detection. It accepts either unprocessed reads in FASTQ format or pre-aligned reads in BAM format. An enhanced version, RES-Scanner2, introduces the capability to pinpoint hyper-editing sites, areas characterized by elevated editing activity, which arises from the propensity of ADAR1 to edit multiple adjacent sites, thereby creating clusters of edits.

RNAEditor is a comprehensive, automated tool designed to process uncompressed FASTQ files for the prediction of RES (78). It offers both command line and graphical user interface (GUI) options for users. RNAEditor's methodology unfolds in three main stages: first, it maps reads using BWA (85); second, it performs SNP calling and refinement employing the GATK tool (86); and third, it annotates identified RES.

To minimize the occurrence of false positives in RES prediction, RNAEditor filters out known SNPs by leveraging data from the dbSNP database, the 1000 Genomes Project (87), and the HAPMAP project (88). Furthermore, it utilizes a clustering algorithm to identify regions with a high frequency of editing, known as editing islands, a feature attributed to the enzyme ADAR1's tendency to target and modify clusters within double-stranded RNA sequences.

JACUSA stands out for its ability to identify Single Nucleotide Variants (SNVs) by juxtaposing RNA-DNA or RNA-RNA sequencing data, enriched by the integration of replicate experiment information (89). The tool adeptly identifies position-specific RNA Editing Sites (RES) using RNA- and DNA-seq data, leveraging the principle that RNA editing, a post-transcriptional modification, is absent in genomic DNA. For RNA-RNA comparisons, JACUSA contrasts RNA-seq data from distinct conditions (such as control versus knockdown samples) to pinpoint differential RES. It also excels in eliminating well-known artifacts stemming from mapping algorithms or sequencing technologies. JACUSA2, its updated version, boasts improved efficiency with a

reduced run time and enhanced detection capabilities for complex read signatures, including substitutions, insertions, deletions, and read truncations. Furthermore, JACUSA2 introduces a novel analysis mode aimed at detecting read arrest events in paired end read samples, showcasing its advanced utility in RNA editing research.

SPRINT uniquely identifies RNA Editing Sites (RES) and hyper-editing regions without necessitating a preliminary step to exclude known SNPs (90). This approach acknowledges the importance of individual variability at SNP sites, which may not uniformly represent false-positive RES indicators due to the genetic distinctions between individuals. SPRINT's methodology unfolds in three main phases: initial read processing, Single Nucleotide Variant (SNV) detection, and the conclusive identification of RES. This final stage involves the aggregation of SNV duplets, pairs of sequential SNVs exhibiting identical mutation patterns, such as successive A > G alterations when juxtaposed with the reference genome in RNA-seq data. The underlying hypothesis suggests that closely situated consecutive SNVs, specifically within a 400-nucleotide range, are likely authentic RES due to the propensity of ADAR enzymes to modify adjacent sites. Conversely, consecutive SNVs spaced within 1600 nucleotides are presumed to be SNPs. SPRINT is compatible with both uncompressed FASTQ files, where BWA serves as the mapping tool, and BAM files. Notably, the feature to detect hyper-editing sites is exclusively accessible when analyzing raw FASTQ files.

2.9.2 Machine learning and its applications in RNA editing prediction

Machine learning (ML) encompasses the techniques and processes involved in developing predictive models from data or identifying meaningful patterns within data which aims to replicate or simulate human pattern recognition capabilities through objective, computational methods (91). Machine learning proves invaluable for analyzing datasets that are either in high volumes or too intricate for manual examination, facilitating the automation of data analysis workflows for reproducibility and efficiency. Biological research generates datasets characterized by significant size and complexity, making machine learning an essential tool for interpreting such data. The application of machine learning in biology has been evolving for decades,

increasingly becoming a cornerstone in virtually all areas of biological studies due to the exponential growth in data volume and complexity, underscoring the necessity for both effective data analysis methodologies and a thorough comprehension of these approaches. Before I go into details of applications of ML in RNA editing prediction, I will go over some basics of ML and types of ML methods used in this study.

2.9.3 Basics of machine learning

In broad terms, a “dataset” consists of numerous data points or instances, with each representing a unique observation from an experiment. These data points are characterized by a set of “features”, which are typically consistent in number across the dataset. Such features might include attributes like length, time, concentration, and gene expression level. A “machine learning task” defines the specific goal or objective that a machine learning model aims to achieve.

For instance, in a study examining gene expression over time, one might aim to predict the transformation rate of a particular metabolite. In this scenario, 'gene expression level' and 'time' would serve as input features for the model, whereas the 'conversion rate' represents the model's target output—the variable the model is designed to predict. Models can accommodate any number of input and output features, which may either be continuous, representing an uninterrupted range of numerical values, or categorical, indicating distinct, separate values. Often, categorical features are binary, indicating a state of being either present (1) or absent (0) (91).

"Supervised machine learning" involves developing a model based on data that has been labeled with some form of "ground truth". This ground truth, which serves as a reference for training the model, is typically determined through experimental measurements or human annotation. The foundational truth originates from empirical lab observations, although these initial data often undergo preprocessing before being utilized in model training. In contrast, unsupervised learning techniques detect patterns within unlabeled data, eliminating the necessity for inputting ground truth as predefined labels. Occasionally, these methods are merged into a semi-supervised learning framework, where a limited quantity of labeled data is integrated with a larger

pool of unlabeled data. This hybrid approach can enhance model performance, especially when acquiring labeled data is expensive or difficult (92).

When the task at hand is to categorize data points into distinct groups, such as labeling them as 'cancerous' or 'non-cancerous,' this is referred to as a 'classification problem.' Algorithms designed to execute this task are known as classifiers. In contrast, regression models predict a continuous range of outcomes, for instance, estimating the change in free energy resulting from a mutation in a protein's residue. It's often feasible to convert regression outputs into categorical data by applying thresholds or categorization techniques, effectively transforming regression tasks into classification problems. For instance, changes in free energy could be categorized into ranges that either promote or detract from protein stability (93).

The results produced by a machine learning model are not perfect and will vary from the actual ground truth. The mathematical functions used to quantify this variation, or broadly, the degree of 'disagreement' between the predicted outputs and the true outputs, are known as 'loss functions' or 'cost functions'. In the context of supervised learning, these functions gauge the difference between the model's predictions and the actual ground truth values. Common examples of such measures include the mean squared error loss for regression tasks and binary cross-entropy for classification tasks.

Clustering methods aim to identify groups of data points that share similarities within a dataset, relying on measures of similarity between those points to form clusters. These methods fall under unsupervised learning, meaning they don't depend on pre-labeled data for training. An example of their application is in gene expression studies, where clustering can reveal patient subsets with comparable gene expression profiles.

Models fundamentally act as mathematical functions that process a given set of input features to generate one or more outputs or results. To learn from training data, these models are equipped with tunable parameters. The values of these parameters are iteratively adjusted throughout the training phase to optimize the model's

performance. Before deployment for prediction tasks, models must undergo a training phase where their parameters are automatically adjusted to enhance performance. In supervised learning scenarios, this adjustment aims to refine the model's accuracy on a training dataset by reducing the average loss or cost function value. Typically, an additional, separate validation dataset is employed to oversee the training progress without affecting it. In unsupervised learning contexts, even though there are no ground truth outputs to compare against, the optimization still focuses on minimizing a cost function. Following training, the model's effectiveness is evaluated using a test dataset that was not part of the training process.

2.9.4 Machine learning methods used for data classification

2.9.4.1 Support vector machines (SVM)

Numerous machine learning algorithms exist for classification tasks, and Support vector machines (SVM) represent one of these options. The strength of a SVM lies in its capability to discern data classification trends with consistent accuracy and reliability. While sometimes applied for regression tasks, SVM is primarily celebrated for its utility in classification. Its broad applicability spans a variety of data science contexts, underscoring its versatility as a tool.

The SVM decision function fundamentally identifies an optimal "hyperplane" (Figure 3) designed to distinguish observations in one category from another, leveraging patterns in the data's features (94). This hyperplane is subsequently utilized to predict the likely classification of new, unseen data. The features that inform the construction of the hyperplane generally stem from processed data, typically derived through some form of transformation or analysis conducted during the feature selection phase of machine learning. Features are organized into coordinates reflecting their interrelations, which constitute the support vectors. Similar to other machine learning approaches, utilizing SVM entails a balance between two objectives: (a) enhancing the classifier's ability to correctly identify new instances (thus improving accuracy) and (b) ensuring the classifier's effectiveness on unseen data (thereby ensuring its generalizability). The success of the first goal depends on the relevance

and significance of the features (feature importance), whereas the achievement of the second goal is influenced by the diversity and quantity of unique examples employed in training the model. Training a Support Vector Machine (SVM) decision function involves finding a consistent hyperplane that maximizes the distance (or "margin") between the support vectors representing different class labels.

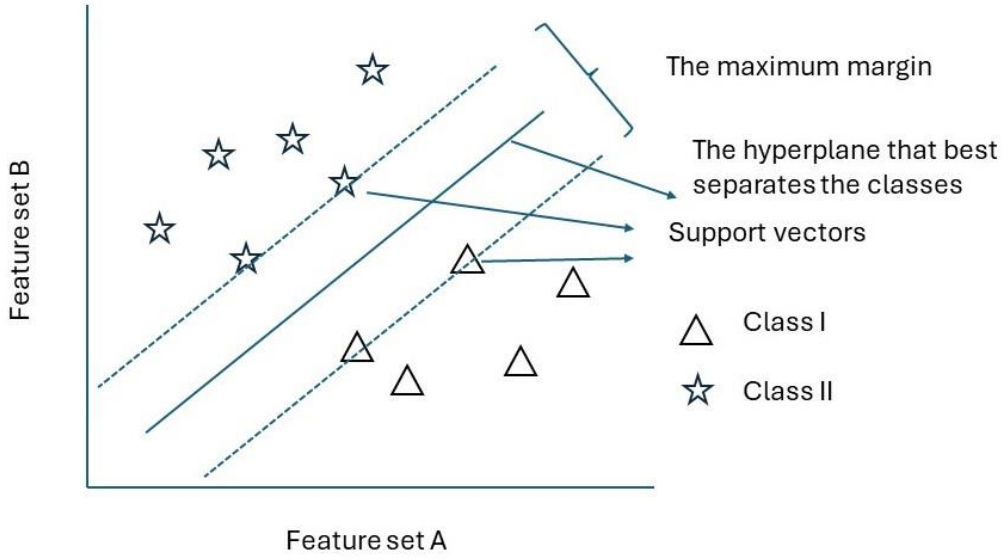

Figure 3. Depiction of the SVM hyperplane

2.9.4.2 Random forests (RF)

Random forests (RF) are among the leading ensemble learning techniques, widely applied in data mining and machine learning (95). This method combines the concept of adaptive nearest neighbors with bagging to facilitate data-adaptive inference without depending on parameter specifications (96). Its sequential, greedy approach to splitting nodes allows trees, and consequently forests, to regularize analysis effectively, especially in situations with many variables but few observations (97). Moreover, the ability of trees to group allows RF to skillfully navigate correlations and interactions between variables. Additionally, RF's capability to assess and prioritize variables through measures of variable importance renders it highly suitable for genomic data analysis and bioinformatics, offering a powerful tool for these research fields.

2.9.4.3 Naïve bayes classifier

A Naïve Bayes Classifier operates on the principle of making probabilistic predictions by applying Bayes' theorem. Essentially, it works under the assumption that the occurrence (or non-occurrence) of a specific feature within a category is independent of the occurrence (or non-occurrence) of any other feature (98). Naïve Bayes represents a method of statistical classification that allows for the prediction of future class probabilities. Originating from the same classification framework as decision trees and neural networks, the Naïve Bayes approach is recognized for its efficiency and accuracy, particularly in scenarios involving large datasets. The benefits of Naïve Bayes include its simplicity in both understanding and implementation, its speed in generating class predictions compared to other classification algorithms, and its ability to be effectively trained with a small amount of data (98).

2.9.4.4 A multi-layer perceptron (MLP)

A Multi-Layer Perceptron (MLP) is an extension of the feedforward neural network, comprising three distinct types of layers: the input layer, the output layer, and one or more hidden layers as depicted in Figure 4 (99). The input layer is responsible for receiving the initial data signals, while the output layer executes tasks such as prediction and classification. Positioned between the input and output layers, the hidden layers serve as the core processing units of the MLP. Data in an MLP moves in a forward direction, from the input to the output layer, mirroring the flow in a feedforward network. The training of neurons within the MLP is accomplished through the backpropagation learning algorithm. MLPs can approximate any continuous function and addressing problems that are not linearly separable, making them highly versatile for tasks like pattern classification, recognition, prediction, and approximation.

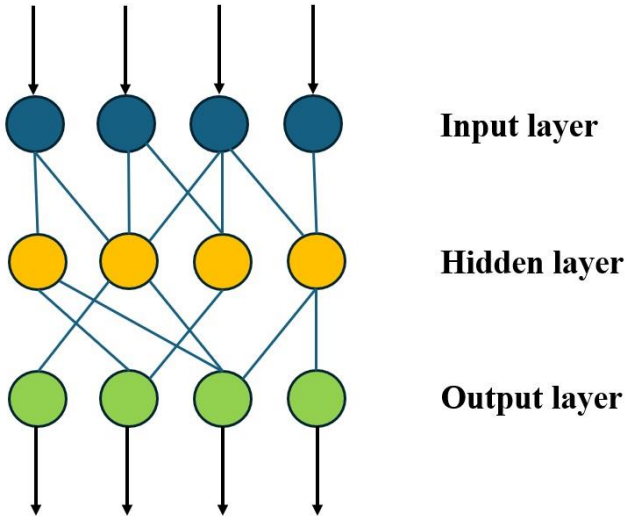

Figure 4. Illustration of a multi-layer perceptron

2.9.4.5 Binary logistic regression

Logistic regression is an advanced statistical technique that surpasses simple linear regression, especially suited for binary (yes/no, high/low) outcomes. Linear regression is less effective with binary dependent variables, whereas logistic regression is specifically designed to assess how different independent variables affect a binary result (100). It considers all independent variables together in a single analysis, which allows for the evaluation of the predictive power of each variable while also considering the impact of all other variables in the model.

This integrated approach provides a thorough insight into how each factor contributes to the binary outcome. For logistic regression to perform optimally, it requires adherence to specific conditions: It's essential to have a proper balance between the number of predictors and the number of participants, the independent variables should not be excessively correlated with one another, and dataset should be free of extreme values that deviate notably from other data points (100).

2.9.5 Machine learning applications for RNA editing

Machine learning (ML) is becoming more prevalent in fields such as computer science and genomics, with numerous ML-driven tools for RES detection emerging in

recent times. One notable instance is RDDpred (16), which stands as the inaugural pipeline crafted using an ML methodology specifically for RES prediction.

RDDpred leverages data from RNA-editing databases such as RADAR (101) and DARNED (102), along with the Mapping Errors Set (MES) technique, to create both positive and negative training datasets for RNA Editing Sites (RES). The MES approach pinpoints areas prone to SNP site errors through the simulation of randomly mutated sequencing reads, which are aligned to the reference genome and evaluated using variant calling tools. Initially, these training sets are used to distinguish between positive and negative RES. Subsequently, this classification aids in the further analysis through a prediction model. This model, based on a random forest algorithm, scrutinizes the remaining sites to identify RES not previously cataloged in the RNA editing databases.

iRNA-AI (17) is a prediction tool that utilizes Support Vector Machines (SVM) to integrate the chemical properties of nucleotides with their frequency distribution across an RNA sequence, forming a pseudo nucleotide composition.

RED-ML (103) integrates a variety of data, including read characteristics, sequencing and alignment artifacts, and specific attributes of RNA Editing Sites (RES) such as sequence context, the presence of candidate sites in Alu regions, and the type of editing, to predict RES across the genome using a logistic regression classifier. When DNA sequencing data are available, Single Nucleotide Polymorphisms (SNPs) can be identified using variant calling tools like GATK and included in the analysis. While RED-ML accepts BAM files as input, it requires a specific reference file when using BAMs produced by the BWA read aligner. This reference combines the human genome assembly release 19 (GRCh37) with exonic sequences from all known splice junctions. However, the tool's utility is limited by factors such as its restriction to human RNA sequencing data, reliance on BAM files based on the older GRCh37 human genome assembly, and its effectiveness only for identifying sites with substantial editing levels.

DeepRed (18) employs deep and ensemble learning techniques to pinpoint potential RNA Editing Sites (RES) directly. The tool operates with a simple input: a list of Single Nucleotide Variants (SNVs) detailing chromosomal locations, reference alleles, and alternate alleles. A prerequisite for generating this list is a variant calling step using tools such as GATK or BCFtools, though further filtering or annotation steps are not necessary. A key advantage of DeepRed is its capacity to analyze RNA sequencing data from multiple species without needing extensive software dependencies. However, its reliance on MATLAB, which requires a paid license, is a notable limitation. This dependency on a proprietary platform was a factor in excluding DeepRed from our comparative benchmarking study.

Besides the stand-alone applications previously discussed, various web-based platforms like DREAM (Detection of RNA Editing Sites associated with miRNAs) (104), AIRlINER (Assessment of RNA editing sites in non-repetitive regions) (105), and REP (Prediction of RNA editing sites and their effects in humans) (106)offer services for identifying RNA Editing Sites (RES). These online resources provide the benefit of requiring no significant computational resources or disk space on the part of the user. However, potential drawbacks include the time it takes to upload data and limitations in processing multiple datasets simultaneously.

3 MATERIALS AND METHODS

3.1 Datasets

The positive dataset was created using the Rediportal database, which catalogues over 4.5 million A-to-I RNA editing events across 55 human body sites from numerous RNA-seq experiments (107). For this dataset, genomic sequences for 5000 A-to-I RNA editing events were randomly selected from Rediportal. These selections included flanking sequences of various lengths (20, 30, 40, 50, and 75 nucleotides) on both the 3' and 5' sides of the editing sites. Conversely, the negative dataset comprised 5000 randomly chosen sequences of equivalent length from the reference genome (hg19), each containing a central adenine but not included in the Rediportal database. Both datasets were randomly and equally split into training and testing sets, with strict measures to prevent overlap between the training and testing data. The sizes of the positive and negative datasets were determined based on preliminary findings from training a support vector machine classifier with varying numbers of sites (500, 1000, 2000, 5000, and 10000 sites per dataset). It was observed that the accuracy of the SVM did not increase beyond using 5000 sites in each dataset.

Following the optimization of parameters and performance, the SVM classifier was evaluated using U87MG cell data from the Bahn study (69), which involved transcriptome sequencing (RNA-seq) of U87MG cells treated with ADAR1 siRNA knockdown or a control siRNA. True RNA editing events were identified as those present only in cells transfected with control siRNA and RNA editing events exclusive to ADAR1 knockdown cells or observed in both cell types were considered as false RNA editing events. The aligned RNA-seq reads were retrieved from the NCBI GEO database under the accession number GSE28040.

3.2 Models for Support Vector Machine Input

3.2.1 Structural model

The hypothesis suggests that the RNA substrate's structure and sequence are pivotal for RNA editing. In support vector machine model, information was incorporated on RNA structure alongside sequence data. Dot-bracket notation was utilized for secondary structure from RNA fold (108), combined with tri-nucleotide sequence information that had been previously applied in the computational prediction of microRNAs (19,20). In dot-bracket notation, unpaired nucleotides are denoted by a dot ".", paired nucleotides at the 5'-end by a left bracket "(", and those at the 3'-end by a right bracket ")". However, in the model, both types of pairings are represented using a left bracket "(" to simplify the data. The notation allows for eight possible combinations with four nucleotide types (A, T, G, C) at the center, there are 32 unique tri-nucleotide combinations (Figure 5). 32 different triplets were counted as they appeared while sliding the window by one nucleotide across the sequence. These counts were normalized and provided to the SVM classifier as a 32-dimensional vector. Only on stem-loops, excluding terminal loops and external single-stranded regions were included. As previously done in the literature (19), the terminal loop and external single-stranded regions were excluded from evaluation and focusing solely on stem-loops. Including these regions has been shown to reduce the classification accuracy of the SVM classifier, likely because these parts are typically overrepresented in RNA structures. Furthermore, the editing sites of ADAR enzymes are predominantly located within extensive duplex structures (109–111) leading to omit these long unpaired loops and tangling sections from the analysis.

3.2.2 Sequence model

In the sequence model, the composition of triplet sequence elements (such as AUU, AAU, etc.) within each nucleotide sequence were focused on, omitting any consideration of structural components. The occurrence of 64 distinct triplet types were tracked by sliding one nucleotide at a time over the sequence. These counts were then normalized and inputted into the SVM classifier as a 64-dimensional vector.

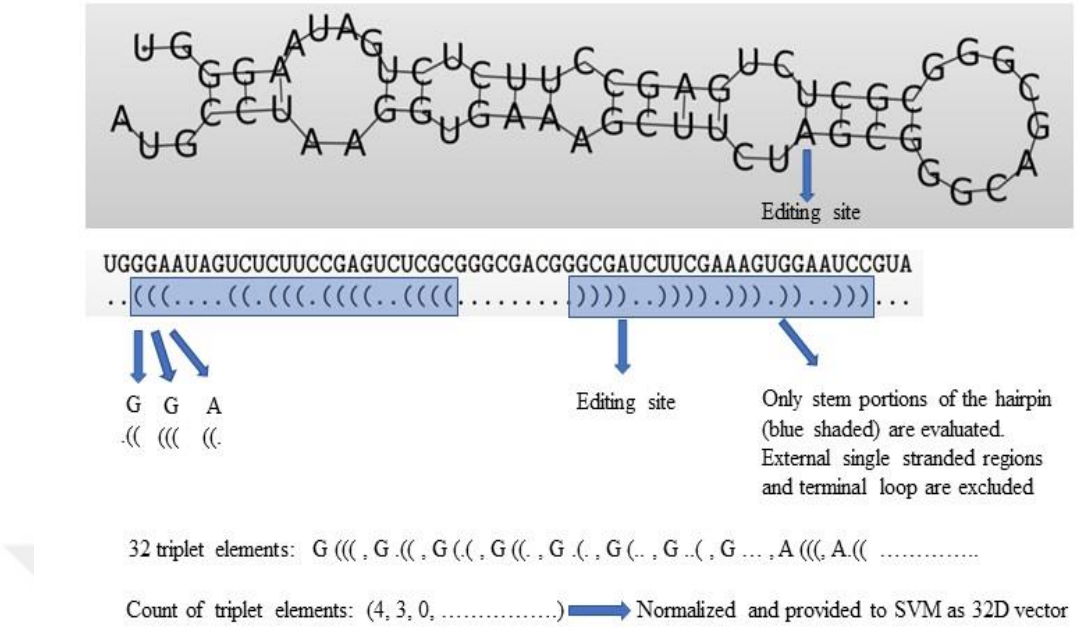

Figure 5. Structural model: The use of dot-bracket notation to capture structure and sequence information around the A-to-I RNA editing sites

3.2.3 Combined model

In the combined model, two vectors from the structure-sequence model, which has 32 dimensions, and the sequence model, which contains 64 dimensions were merged. This aggregated 96-dimensional vector was then inputted into the support vector machine classifier.

3.3 Support Vector Machine

The models were evaluated using various binary classifiers, including Support Vector Machines (SVM), Multilayer Perceptron, Naive Bayes, Random Forest, and Boosted Logistic Regression, implemented using the R library "caret". 10-fold cross-validation was performed on the training dataset (1000 editing points). Optimal parameters for the model were selected using the largest accuracy value. Since SVM delivered the highest accuracy among all tested classifiers for both the sequence and structural models, SVM was used for the rest of the study (Table 1).

Table 1. Performance metrics of different classifiers obtained by sequence and structural model.

	Accuracy		**Sensitivity**		**Specificity**	
Classifier	Sequence Model	Structural Model	Sequence Model	Structural Model	Sequence Model	Structural Model
Support Vector M.	84.7	81.2	86.7	84.4	83	78.7
Multilayer Perceptron	83.9	79.7	85.3	80.7	82.7	78.8
Naive Bayes	82.7	79.4	85.4	83.8	80.5	76.3
Random Forest	84.5	80.9	88.2	84.7	81.7	78.1
B. logistic Regression	78.1	77.5	77.9	81	78.3	74.9

The core concept of a support vector machine (SVM) is to map the input data into a high-dimensional feature space and then identify an optimal separating hyperplane, which serves as the decision boundary between two classes (112). Effective training of SVM models necessitates the appropriate choice of kernel function and its parameters. Linear, radial, and polynomial kernel functions were trained with various parameters (cost for linear; cost and gamma for radial; cost and degree for polynomial) using 10-fold cross-validation on the training dataset to fine-tune these parameters. The following performance metrics were defined and employed to guide the selection of the kernel function and its parameters:

Sensitivity (SE) = TP / (TP + FN),

Specificity (SP) = TN / (TN + FP),

Accuracy (ACC) = (TP + TN) / (TP + FN + TN + FP).

(TP: True positive, TN: True negative, FP: False positive, FN: False-negative)

R package (113) which is an interface for libSVM package (114) was used for the implementation of the SVM.

3.4 Feature Selection

The F-score is commonly employed to evaluate the discriminative capacity between two sets of real numbers. According to Chen and colleagues (115), the F-score for the i^{th} feature in a vector is defined as follows:

$$F(i) \equiv \frac{\left(\bar{\boldsymbol{x}}_i^{(+)} - \bar{\boldsymbol{x}}_i\right)^2 + \left(\bar{\boldsymbol{x}}_i^{(-)} - \bar{\boldsymbol{x}}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{\boldsymbol{x}}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{\boldsymbol{x}}_i^{(-)}\right)^2}$$

n_+ and n_-: number of positive and negative instances

$\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{x}}_i^{(+)}, \bar{\boldsymbol{x}}_i^{(-)}$: the average of the i^{th} feature of the whole, positive, and negative data sets

$x_{k,i}^{(+)}$: the i^{th} feature of the k^{th} positive instance

$x_{k,i}^{(-)}$: the i^{th} feature of the k^{th} negative instance.

The numerator and denominator of the F-score reflect the discrimination between the positive and negative sets, and the discrimination within each set, respectively. A higher F-score indicates that the feature is more discriminative.

3.5 Identifying RNA Editing Sites in Experimentally Validated Datasets

The SVM classifier was validated using experimentally confirmed U87MG data from the Bahn study (69), where RNA-seq was conducted on ADAR1 knockdown and wildtype U87MG cells. RNA editing sites in these cells were identified using REDItools version 1.0.4 (79), following the exclusion of potential SNPs. These identified RNA editing sites were categorized as true or false editing sites and were used to validate the SVM classifier.

Aligned RNA-seq reads were obtained from the NCBI GEO database under the accession number GSE28040 in the Sequence Alignment/Map (SAM) format. These files were described in the original study (69). The files were then converted from SAM to BAM format, sorted, and indexed using SAMtools version 1.3.1 (116). Duplicate reads were removed using Picard tools version 2.8.3 (117) to eliminate biases from PCR amplification in the RNA-Seq libraries.

The Python script REDItoolDnaRna.py was employed to detect nucleotide changes (A to G) between the reference genome and the RNA-Seq BAM files. Identified mismatch sites were considered as potential RNA editing sites. Stringent criteria were applied in the detection process: a minimum coverage of 10 reads (-c 10), a minimum quality score of 25 (-q 25), a minimum mapping quality of 20 (-m 20), and editing sites had to be supported by at least 3 variant bases (-v 3) with a variation frequency of at least 10% in RNA-Seq (-n 0.1). To reduce false positives, 6 bases at the 3′ ends of each RNA-Seq read were truncated (-a 6-0), duplicates were excluded (-d), substitutions in homopolymeric regions of five or more nucleotides (-l), and substitutions within 4 nucleotides of known splice junctions (-r and -w) were removed. Additionally, possible SNPs were excluded using the web tool SNPnexus version 4 (118).

3.6 RDDSVM

RDDSVM (RNA-DNA Differences with Support Vector Machines) is an open-source R program designed to predict candidate A-to-I RNA editing sites. This tool employs machine learning techniques, specifically support vector machines, and leverages sequence information surrounding the editing sites for predictions. RDDSVM is user-friendly and accessible even to those with limited bioinformatics experience. It is compatible with major operating systems and can be downloaded from its GitHub page (119).

3.7 Comparison of RDDSVM With Other Tools on Validated A-To-I RNA Editing Sites

The predictive performance of RDD-SVM was assessed against other classifiers using human A-to-I RNA editing sites validated by (120). Performance comparisons with Sun et al.'s method and IosinePredict were sourced from (15). The RDD-SVM was given 101 nt long sequences surrounding the validated A-to-I RNA editing sites as input. The same sequences were also input into AIRLINER to estimate the editing probability of the adenosine, setting the probability threshold at 0.5 as specified by (105). AIRLINER can be accessed at (121). For IosinePredict, results reflecting

combined accurate predictions for ADAR1 and ADAR2 protein preferences were used, with cutoffs of 9.6% for unedited sites by ADAR1 and 21% for ADAR2. InosinePredict is available at (122).

3.8 Identification of Differential Editing Sites

RNA editing sites used in this study is obtained from Han's study (9) where RNA editing data was generated from TCGA samples of 17 different cancer types. The method for detection and quality control of RNA editing sites were described in the original study. 12 cancer types which included tumor and normal samples (4257 tumor and 557 normal) were selected to determine sites with differential editing in tumor samples vs normal samples. RNA editing sites which were present in at least 5 pairs of normal and non-tumor samples were selected as informative RNA editing sites. For comparison purposes, Wilcoxon test was used to detect RNA editing sites with differential editing between tumor and non-tumor samples. Significant differential editing sites were defined as sites with FDR < 0.05 and a mean editing level difference ≥ 5% between tumor and non-tumor samples.

3.9 Identification of Regulatory Editing Sites

Out of the consequences of RNA editing in the transcriptome, alteration of gene expression is of paramount importance in cancer. Therefore, the aim was to identify RNA editing events which were correlated to gene expression (referred as regulatory editing sites in this thesis) and help to discover new cancer specific prognostic RNA editing markers. For identification of regulatory RNA editing sites, 4 major cancer types with increased editing in tumor samples were selected: head and neck squamous cell carcinoma (HNSC), breast invasive carcinoma (BRCA), thyroid carcinoma (THCA), lung adenocarcinoma (LUAD). RNA expression data of tumor samples were downloaded for those 4 cancer types and normalized using TCGAbiolinks R package (123). RNA editing sites were annotated using R package "Annotatr" (124) with human genome assembly GRCh37 (hg19) and were matched with corresponding genes (if any).

Sites which have editing in at least %25 of respective tumor samples were selected for analysis. For each selected RNA editing site, RNA editing frequencies were compared to normalized RNA expression values using spearman correlation (FDR-Benjamini Hochberg corrected p-value < 0.05). Spearman's correlation was selected over Pearson's correlation since Spearman's correlation assesses monotonic relationships instead of linear relationships (Pearson's correlation) which is more suitable to investigate a relationship between editing and expression.

3.10 Survival Analysis

The impact of the editing frequency of a regulatory RNA editing site on the survival of the cancer patient was performed as follows: The clinical data of the patients in this study was downloaded from The National Cancer Institute (NCI) Genomic Data Commons (GDC) using TCGAbiolinks R module (123). For a specific RNA editing site, the patients which had editing at that site were grouped into two groups according to their editing frequency. Patients with editing frequency higher than the overall mean was selected as the high editing group whereas the patients with editing frequency lower than the mean was selected as the low editing group. The survival analysis was performed using Kaplan-Meier method and editing sites with p-value <0.05 was selected as the sites that are correlated with the survival of the cancer patients.

3.11 Testing of the Effect of Regulatory Editing Sites For mRNA – miRNA Interaction

Wild type and edited sequences of 51 nucleotide length (25 nucleotides upstream and downstream of editing site) of regulatory and non-regulatory editing sites were submitted to miRNASNP-v3 online prediction module (125). The number of gained or lost miRNA target sites and minimum free energy (MFE - kCal/mol) of the miRNA-mRNA duplex were retrieved and compared for statistical difference between regulatory and non-regulatory group using Wilcoxon test.

4 RESULTS

4.1 Optimization of Models for Support Vector Machine Input

To enhance the prediction accuracy of support vector machines (SVM), three variables were fine-tuned: the length of flanking sequences surrounding the A-to-I editing events (41nt, 61nt, 81nt, 101nt, and 151nt), the method for transforming sequence data for SVM input (structural, sequence, and combined models; refer to Materials and Methods), and the choice of SVM kernel function (linear, radial, and polynomial). The optimization outcomes—including sensitivity, specificity, and accuracy—are displayed in Figure 6 and Table 2. The highest accuracy (86.5%) was achieved using a polynomial kernel function with the sequence model based on 101nt sequences around the editing events, as shown in Figure 6C. The sequence model consistently outperformed the structural model across all lengths. Although the combined model increased sensitivity, it did not enhance overall accuracy and showed lower specificity compared to the sequence model, as indicated in Figures 6A and 6B. Furthermore, the polynomial kernel generally outperformed the linear kernel, while the radial kernel's weaker performance is not presented.

4.2 Optimization of the Support Vector Machine Parameters

Accurate parameter selection is critical for effectively training SVM models. Linear, radial, and polynomial kernel functions using various parameters—cost for linear, cost and gamma for radial, and cost and degree for polynomial—through 10-fold cross-validation on the training dataset to fine-tune these parameters. The parameters that yielded the highest accuracy are documented in Table 3.

4.3 Feature Selection

The prevalence of triplet structures in a 101nt sequence from the training data were analyzed comparing the editing dataset (positive) with the control (negative) dataset. Figure 7 displays the frequency of sequence triplet elements, revealing a higher occurrence of triplets beginning with G and C, and a lower occurrence of those

starting with A and U in the editing dataset. A similar pattern is noted in the number of structural triplet elements in the structure-sequence model, shown in Figure 8. There is an enrichment of structural elements such as G((, G.((, G((., C(((, C.((, and C((., while elements like U(((and A(((are less common in the editing dataset. To identify the most discriminative features (either sequence or structural triplets) between the editing and control groups, F-score was employed as the feature selection method, which evaluates the discriminative power between two sets of real numbers (see Materials and Methods). The F-scores for these features derived from the 101 nt sequences are listed in Table 4.

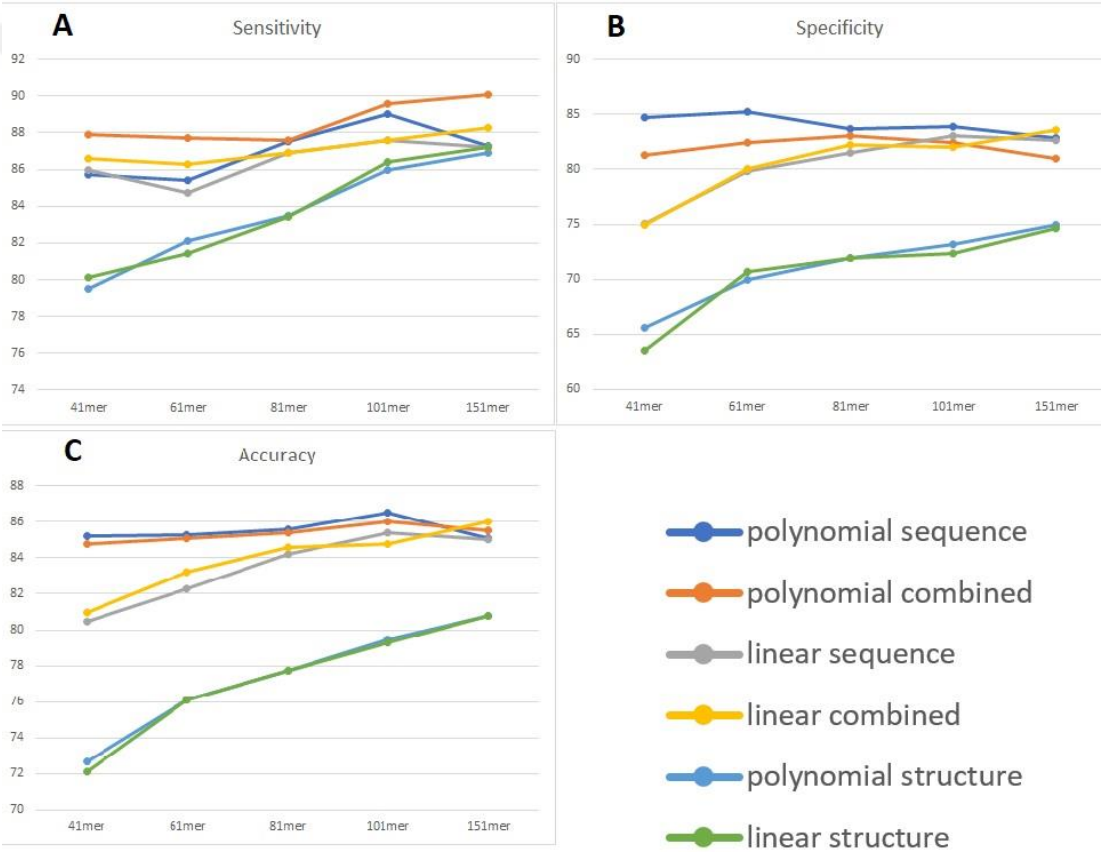

Figure 6. Support vector machine performance results on test data
(A: sensitivity, B: specificity and C: accuracy, polynomial: polynomial kernel function, linear: linear kernel function, sequence: sequence model, structure: structural model, combined: combined model)

Table 2. Support vector machine performance results on test data.
polynomial: polynomial kernel function, linear: linear kernel function, sequence: sequence model, structure: structural model, combined: combined model

		Length of flanking sequence (nt)				
		Accuracy				
SVM kernel function	**SVM model**	**41**	**61**	**81**	**101**	**151**
polynomial	sequence	85.2	85.3	85.6	86.5	85.1
polynomial	combined	84.8	85.1	85.4	86	85.5
linear	sequence	80.5	82.3	84.2	85.4	85
linear	combined	81	83.2	84.6	84.8	86
polynomial	structure	72.7	76.1	77.7	79.5	80.8
linear	structure	72.1	76.1	77.7	79.3	80.8
		Specificity				
SVM kernel function	**SVM model**	**41**	**61**	**81**	**101**	**151**
polynomial	sequence	84.7	85.3	83.7	83.9	82.9
polynomial	combined	81.3	82.4	83.1	82.4	81
linear	sequence	75.1	79.8	81.5	83.1	82.7
linear	combined	75	80.1	82.2	82	83.6
polynomial	structure	65.6	70	71.9	73.2	75
linear	structure	63.5	70.7	71.9	72.4	74.6
		Sensitivity				
SVM kernel function	**SVM model**	**41**	**61**	**81**	**101**	**151**
polynomial	sequence	85.7	85.4	87.5	89	87.3
polynomial	combined	87.9	87.7	87.6	89.6	90.1
linear	sequence	86	84.7	86.9	87.6	87.2
linear	combined	86.6	86.3	86.9	87.6	88.3
polynomial	structure	79.5	82.1	83.5	86	86.9
linear	structure	80.1	81.4	83.4	86.4	87.2

Table 3. SVM kernel parameters
(polynomial: polynomial kernel function, linear: linear kernel function, sequence: sequence model, structure: structural model, combined: combined model, C=cost value, D=degree for polynomial kernel)

		Length of flanking sequence (nt)				
SVM kernel	**SVM model**	41	61	81	101	151
polynomial	**sequence**	C=1, D=3	C=1, D=3	C=1, D=3	C=1, D=3	C=1, D=1
polynomial	**combined**	C=1, D=3	C=1, D=3	C=1, D=3	C=1, D=3	C=1, D=3
linear	**sequence**	C=0.01	C=100	C=0.01	C=0.01	C=0.01
linear	**combined**	C=1	C=0.01	C=0.1	C=1	C=0.01
polynomial	**structure**	C=100, D=1	C=100, D=1	C=10, D=1	C=0.1, D=1	C=1, D=1
linear	**structure**	C=0.01	C=100	C=0.1	C=0.001	C=0.01

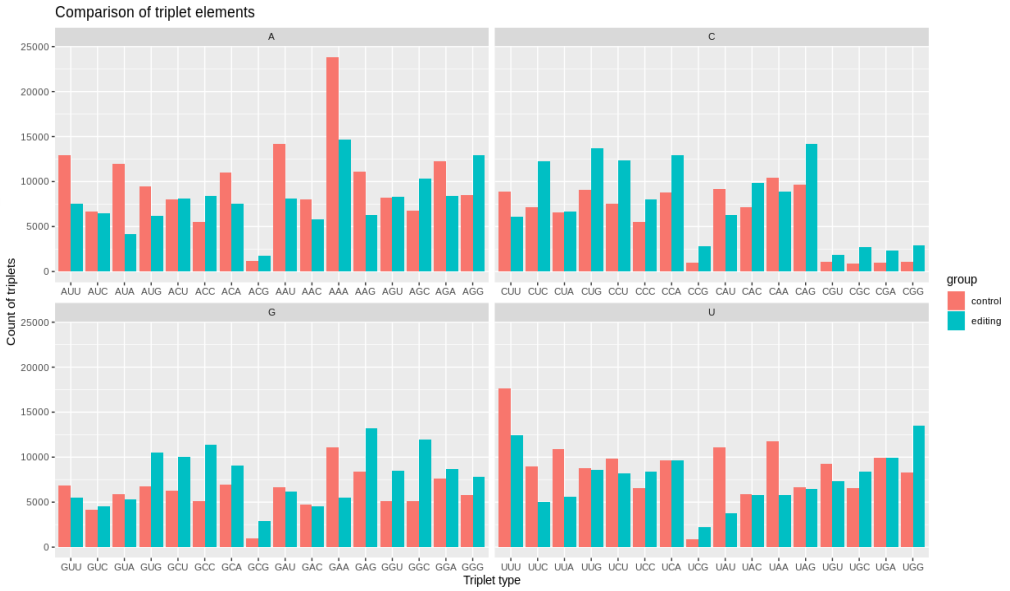

Figure 7. The frequencies of the triplet elements of editing dataset (positive) vs control (negative) for 64 types of triplets (sequence model)

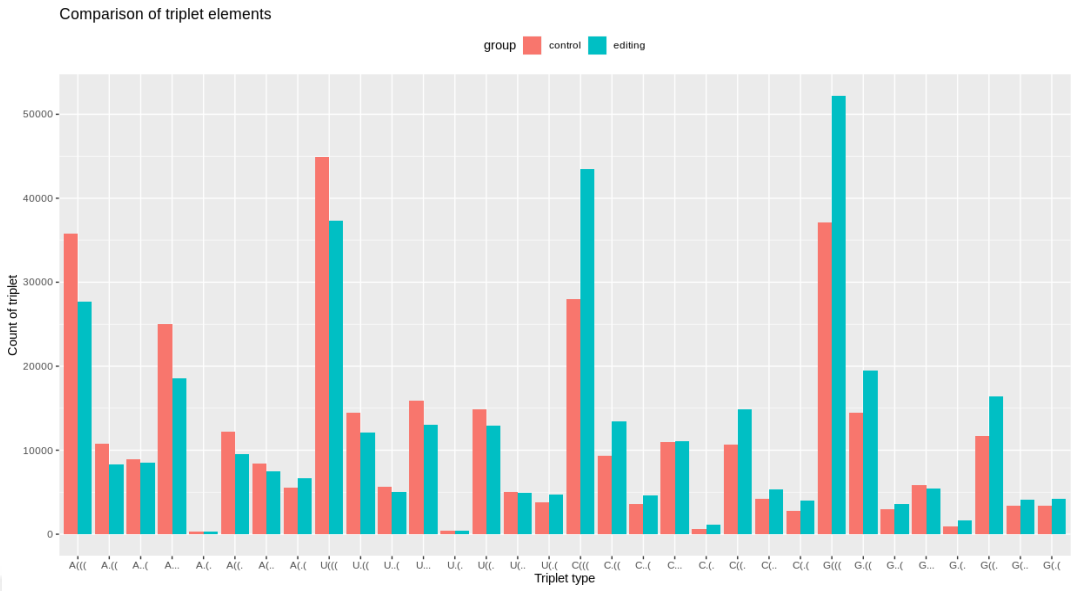

Figure 8. The frequencies of the triplet elements of editing dataset (positive) vs control (negative) for 32 types of structural triplets (structure-sequence model)

Table 4. F-scores for the features obtained from 101 nt flanking sequence (sequence and structural model)

Triplet type	F-score	Triplet type	F-score	Triplet type	F-score
GGC	0.55	CGA	0.1	C(((	0.67
GCC	0.46	AGA	0.099	G(((	0.48
AUA	0.32	AAA	0.098	G((.	0.18
UAU	0.3	CUU	0.097	A(((	0.17
UAA	0.3	ACA	0.097	G.((	0.16
UGG	0.27	AAC	0.068	C.((	0.15
GAA	0.27	CAC	0.066	C((.	0.14
UUA	0.26	CCC	0.06	U(((	0.099
AAU	0.26	GCA	0.051	A((.	0.069
AUU	0.23	UUU	0.044	A.((	0.064
CUG	0.22	CGU	0.04	U.((	0.046
AAG	0.21	GGG	0.04	G.(.	0.042
CAG	0.2	UCC	0.039	A...	0.041
GCU	0.19	UGC	0.039	C(.(	0.037
UUC	0.18	UGU	0.037	U((.	0.03
CUC	0.18	GUU	0.029	C.(.	0.029
CCU	0.18	UCU	0.025	C..(	0.023
GGU	0.18	CAA	0.023	C(..	0.022

Table 4. F-scores for the features obtained from 101 nt flanking sequence (continued)

CCA	0.17	ACG	0.022	A(.(	0.017
AGC	0.17	GGA	0.014	U...	0.017
GCG	0.17	GUA	0.0063	U(.(	0.014
CGC	0.16	GAU	0.0041	A(..	0.01
CCG	0.15	GUC	0.004	G..(	0.0094
CGG	0.14	AUC	0.0011	G(..	0.0088
AGG	0.14	GAC	0.001	G(.(	0.0085
GUG	0.14	UAG	0.00057	U..(	0.0045
GAG	0.14	CUA	0.00037	G...	0.0021
AUG	0.13	UAC	3.00E-04	A..(	0.0016
ACC	0.12	UUG	0.00025	U.(.	0.00036
UCG	0.11	ACU	0.00012	U(..	0.00027
CAU	0.1	UGA	4.00E-05	C...	0.00013
		AGU	1.80E-05	A.(.	9.30E-05

In SVMs, as with many supervised learning models, selecting the right features is crucial for improving performance, reducing computation time, and enhancing interpretability. Each feature's importance was assessed using its F-score and established high and low F-score thresholds. By excluding features below these thresholds, SVM were trained on the remaining features and evaluated for approach's effectiveness. Figure 9 illustrates the performance metrics for methods using sequence and structure-sequence data (with a polynomial kernel function and a sequence length of 101 nucleotides) across various feature counts. The findings show that SVM performance does not improve beyond using the top 40 features for sequence data or the top 7 features for the structure-sequence model.

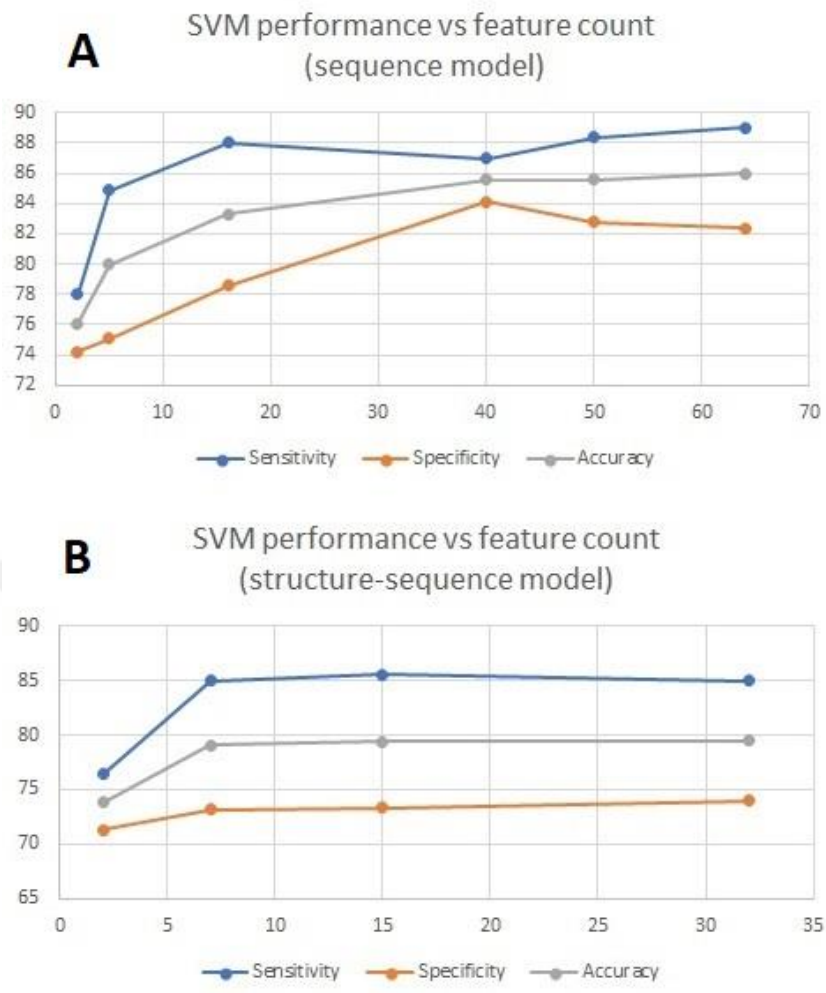

Figure 9. SVM performance metrics vs feature count

4.4 Validation Using Experimental Data

The SVM classifier was validated on experimentally verified data using the approach shown in Figure 10, consistent with the methods described in Ouyang's study (18). SVM classifier were applied to data from the Bahn study (69) which involved transcriptome sequencing (RNA-seq) of U87MG cells either after ADAR1 knockdown or control siRNA transfection. The study indicated that, in the cells transfected with control siRNA, expression levels of other RNA editing proteins, ADAR2 and ADAR3 (which are catalytically inactive), were significantly lower than ADAR1. Consequently, A-G editing events found exclusively in control siRNA transfected cells were classified as true editing events, while those only in ADAR1 knockdown cells or in both cell types were classified as false RNA editing events.

After identifying A-G substitutions and removing known SNPs, 1135 and 1340 A-G substitutions in the first and second replicates of the control siRNA transfected group, 164 and 174 substitutions in the first and second replicates of the ADAR1 knockdown group were identified respectively. To enhance the credibility of these editing sites, we only included A-G editing events that appeared in both replicates within each group. This approach yielded 566 A-G editing sites in the control group and 105 in the ADAR1 knockdown group. Notably, 50 of these sites were found in both the control and ADAR1 knockdown groups, as shown in Figure 10.

For the validation testing, the sequence model and polynomial kernel function based on data from 101 nucleotides surrounding the editing events were selected as the optimal running parameters due to their superior performance. The SVM was trained on 2500 randomly chosen A-to-I RNA editing events from the Rediportal database, along with 2500 sequences of equal length from the reference genome containing central adenines and not listed in Rediportal. The trained SVM classifier was then applied to editing sites from experimentally validated data. Of the 516 A-G editing sites identified solely in the control group (true editing sites), the SVM accurately predicted 479 as true A-G editing sites. Furthermore, of the 105 A-G editing sites identified in the ADAR1 knockdown group (false editing sites), SVM correctly identified 81 as false. This resulted in a sensitivity of 92,8%, a specificity of 77,1%, and an overall accuracy of 90,2%. SVM predictor was also evaluated using the dataset from Bahn's study (69). From the 4141 true editing sites (each with a minimum 20% editing level) identified in control siRNA knockdown cells, the SVM predictor correctly predicted 3635 as true editing sites, achieving a true positive rate (TPR) of 87.8%. However, we were unable to assess the SVM's performance on true negative sites, as such data was not available in the article.

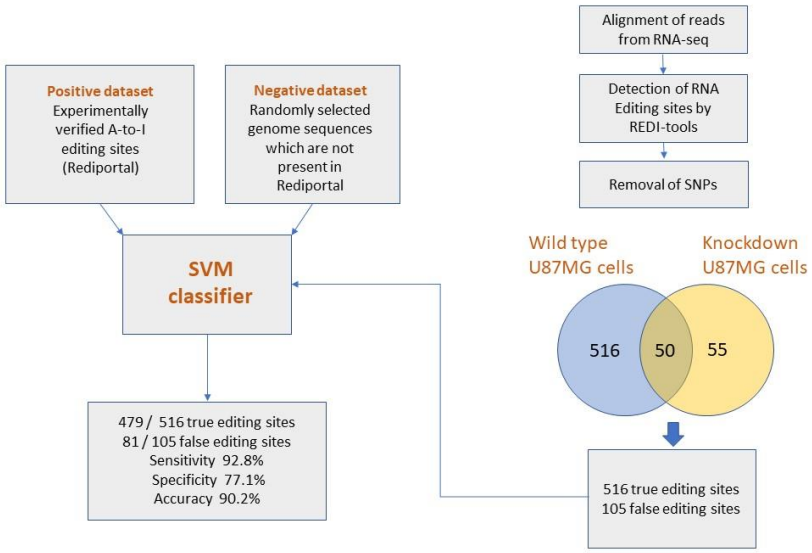

Figure 10. Support vector machine workflow
1) Positive and negative dataset is trained to generate the SVM classifier 2) SVM classifier is validated with experimental data: RNA editing sites from wildtype and knockdown U87MG cells was generated by Reditools and known SNPs were removed. The remaining editing sites were tested by SVM generated using sequence model, polynomial kernel function and 101nt sequences.

4.5 Comparison of RDDSVM With Other Tools on Validated A-To-I RNA Editing Sites

RDDSVM was evaluated against other machine learning classifiers that also utilize sequence data surrounding A-to-I RNA editing sites. Sun et al. (2016) employs string distances between edited and non-edited sites to create a similarity matrix, which an SVM uses to predict new editing sites. IonisePredict (Eggington, Greene, & Bass, 2011) considers the identity of the four bases both 5' and 3' of an adenosine, predicting based on the experimentally determined editing site preferences for human ADAR1 and ADAR2. AIRLINER (Nigita et al., 2015) explores short motifs that might characterize RNA editing signals and applies a logistic regression technique to predict RNA editing events.

To assess the predictive accuracy of our method relative to these classifiers, we used validated human A-to-I RNA editing sites from Peng et al. (2012). Of the 58 A-to-I editing sites identified by our computational pipeline, 57 were confirmed as true

editing sites by Sanger sequencing. One site was identified as a false positive. Of these 58 sites, our RDD-SVM correctly identified 53, achieving an accuracy of 91.4%, which surpasses the performance of the predictors (Table 5).

Table 5. Comparison of classifiers
* Results taken from (15). For Sun et al's method, results for binary classification and one-class classification methods are provided. For IosinePredict, combined correct prediction results for ADAR1 and ADAR2 protein preferences are used. Cutoffs of un-edited sites are 9.6% for ADAR1 and %21 for ADAR2. For AIRLINER, 0.5 is selected as the probability threshold.

CLASSIFIER	ACCURACY (%)
RDDSVM	91.4
Sun et al. (Binary classification) *	79.3
Sun et al. (One-class classification) *	87.9
InosinePredict (hADAR1+hADAR2) *	77.6
AIRLINER	72.4

4.6 Differential Editing Sites

Editing frequency data used in this study was obtained from (9) in which RNA editing data was generated from RNAseq results of 5,672 tumor samples and 564 non-tumor tissue samples from 17 TCGA cancer types. From those editing frequencies, RNA editing sites with differential editing levels (FDR < 0.05 and a mean editing level difference ≥ 5%) between tumor and non-tumor samples of 12 tumor types with existing matching non-tumor samples were selected (Materials and Methods). Most of the differential RNA editing sites showed over-editing in head and neck squamous cell carcinoma (HNSC), breast invasive carcinoma (BRCA), thyroid carcinoma (THCA), and lung adenocarcinoma tumors (LUAD), while differential editing sites in kidney chromophobe (KICH) and kidney renal papillary cell carcinoma (KIRP) tumors showed under-editing compared to matching normal samples (Figure 11 and Table 6).

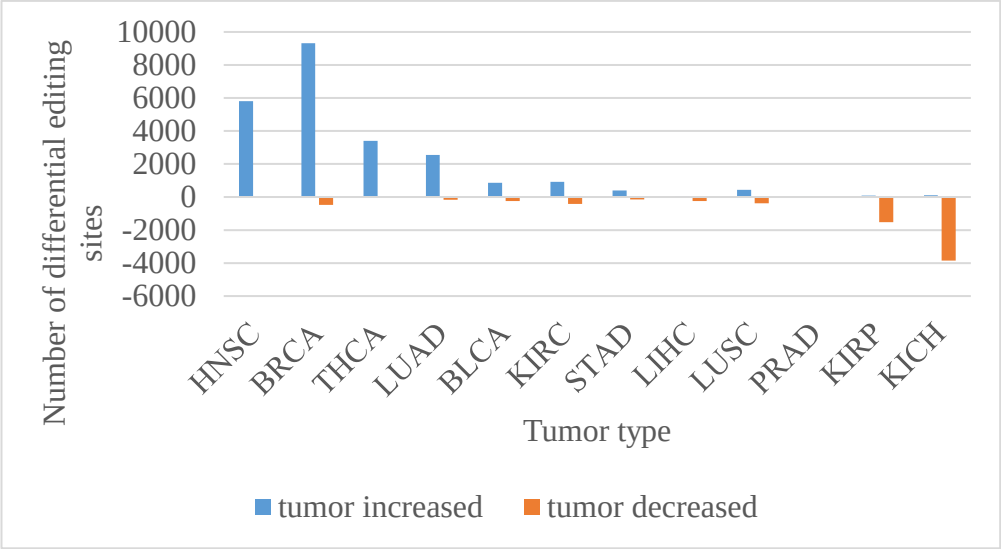

Figure 11. Distribution of differential editing sites in 12 TCGA cancer type

4.7 Features of Regulatory Editing Sites

Although in Han et al.'s study (9) they mostly focused on functional effects of clinically relevant non-synonymous RNA editing sites in coding regions of genes, most of the differential RNA editing events derived in this study are actually in non-coding regions of the genome especially in 3'UTR of genes. Previous studies showed that the editing of 3UTR of genes can alter RNA stability and expression (22,126,127). (128) has used correlation of RNA editing levels with RNA expression in lung adenocarcinoma to predict regulatory editing sites that may change RNA expression and found out that those sites were enriched in both apoptotic and innate immune pathways. My aim was to extend this approach to other cancer types and determine common candidates in different cancer types for evaluation of those candidates to identify new cancer specific prognostic RNA editing markers. Correlation analysis of RNA editing levels with RNA expression was performed in 4 major cancer types with increased editing frequencies in tumor samples: head and neck squamous cell carcinoma (HNSC), breast invasive carcinoma (BRCA), thyroid carcinoma (THCA), lung adenocarcinoma (LUAD) and determined regulatory editing sites in those cancers. Out of 9326 (BRCA), 3400 (THCA) and 2549 (LUAD) differential RNA editing sites, 1303 (BRCA), 460 (THCA), 370 (LUAD) sites were found to be statistically correlated with expression level of matching gene respectively (Table 7).

Only 6 sites were found to be correlated with expression out of 5805 sites in HNSC and those sites are not evaluated in our study due to low sample size.

Table 6. Details of RNA editing sites:
(1) Total RNA editing sites with >10x RNAseq coverage, (2) Informative RNA editing sites present in at least 5 pairs of normal and non-tumor samples, (3) Differential editing sites between tumor and non-tumor samples (FDR < 0.05) and a mean editing level difference ≥ 5% (4) RNA editing sites with increased editing in tumor samples, (5) RNA editing sites with decreased editing in tumor samples

	Sample size		**Number of RNA editing sites**				
	Non-tumor	**Tumor**	**Total (1)**	**Informative (2)**	**FDR < 0.05 and DIFF> 5% (3)**	**Increased in tumor (4)**	**Decreased in tumor (5)**
HNSC	42	426	35510	28926	5827	5805	22
BRCA	105	837	76555	64877	9803	9326	477
THCA	59	498	52701	48779	3432	3400	32
LUAD	58	488	54362	33935	2712	2549	163
BLCA	19	252	39270	24854	1101	861	240
KIRC	67	448	63717	38596	1330	917	413
STAD	33	285	26389	13122	547	401	146
LIHC	50	200	23540	18119	291	53	238
LUSC	17	220	36822	28482	799	426	373
PRAD	52	374	43078	37695	64	15	49
KIRP	30	198	36686	30925	1608	90	1518
KICH	25	66	22317	21400	3949	101	3848

Table 7. Details of editing sites in 3 types of cancer
Differential editing sites were defined as sites with a mean editing level difference ≥ 5% between tumor and non-tumor samples (FDR < 0.05). Regulatory sites are defined as RNA editing sites with editing frequencies statistically correlated to normalized RNA expression values using spearman correlation (FDR corrected p-value < 0.05- Benjamin- Hochberg)

Cancer Type	**BRCA**	**LUAD**	**THCA**
Tumor Samples	837	488	498
Differential sites	9326	2549	3400
Regulatory sites	1303	370	460
Genes	338	129	187
Average Sites / Gene	3.9	2.9	2.5
3UTR	1178 (90.4%)	334 (90.2%)	335 (72.8%)
Intronic	84 (6.4%)	8 (2.2%)	101 (22%)
ncRNA	38 (2.9%)	27 (7.3%)	23 (0.05%)
5UTR	2	-	-
Coding	1	1	1

Spearman coefficients of regulatory sites show different properties for BRCA, LUAD and THCA (Figure 12). For BRCA and LUAD, the coefficients of regulatory sites (spearman p-value <= 0.05) were mostly positive showing that increased editing is generally correlated with increased RNA expression. In THCA, the coefficients of regulatory sites were evenly distributed as positive and negative showing that increase in editing can either correlate with an increase or decrease in RNA expression.

In BRCA and LUAD regulatory sites had decreased editing frequency compared to non-regulatory sites (p values: BRCA: 6.3e-06, LUAD: 0,021) while in THCA, regulatory sites had increased editing compared to non-regulatory sites (p value: THCA: 0.037) (Figure 13).

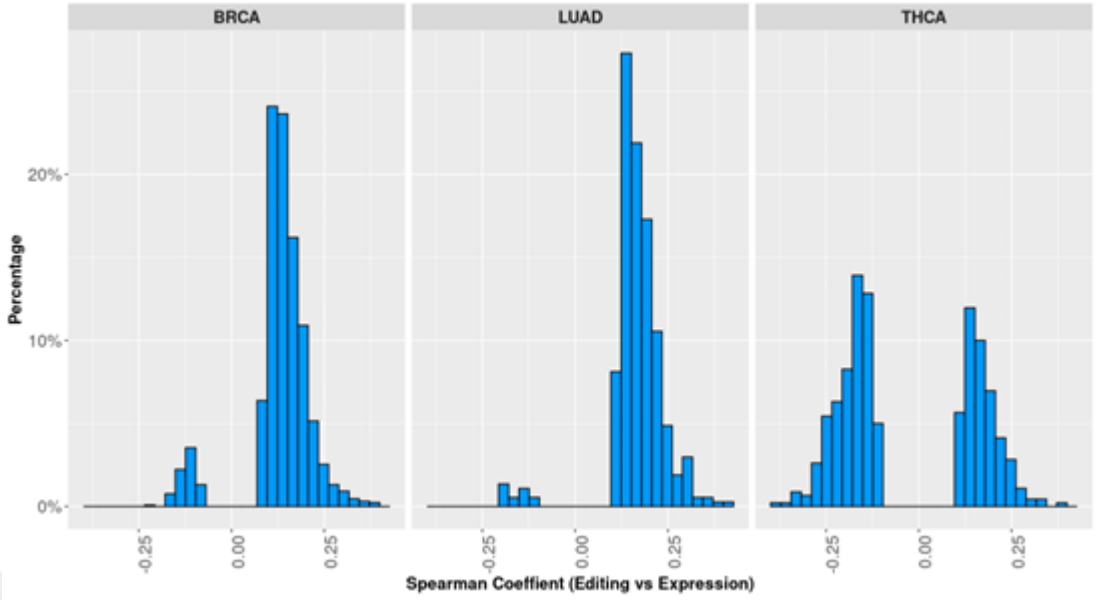

Figure 12. Distribution of spearman coefficients for regulatory sites in BRCA, LUAD and THCA

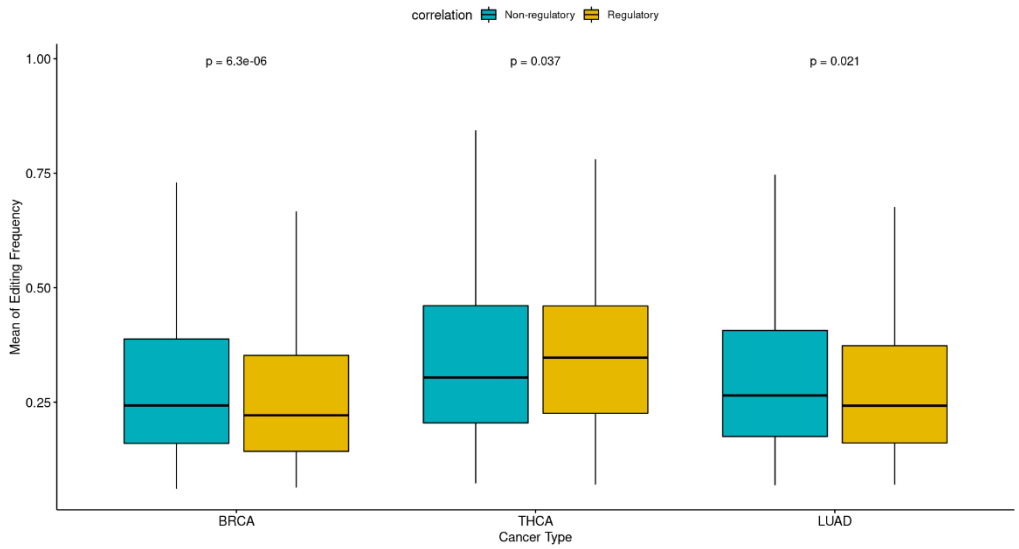

Figure 13. Comparison of mean editing frequencies between regulatory and non-regulatory sites

4.8 Regulatory Editing Sites Are Enriched In 3UTR Of Genes

Table 7 shows the details of regulatory editing sites in 3 types of cancer. The average number of sites per gene in BRCA, LUAD and THCA was 3.9, 2.9 and 2.5 respectively. Regulatory sites were mostly located in 3'UTR of the genes in all cancer

types although the number of sites in intronic regions were enriched (22%) in THCA compared to BRCA and LUAD. Chi-square test shows that the localization of regulatory and non-regulatory editing sites (sites not correlated to gene expression) were statistically different in BRCA and LUAD, while there was no difference in THCA (p-values: BRCA: 1.8e10, LUAD: 3.5 e10, THCA: 0.07). In BRCA and LUAD, the location of the regulatory sites shifted from introns to 3'UTR of genes compared to non-regulatory sites. This tendency for regulatory sites to be in 3'UTR of genes, may support the possibility of regulation gene expression by editing of miRNA binding and preventing the interaction of miRNAs with 3'UTRs of those mRNAs. The editing frequency patterns of the regulatory and non-regulatory sites in 3UTR region of genes and those were similar between 2 groups (Figure 14).

Figure 14. Comparison of editing frequency pattern of regulatory and non-regulatory sites in 3UTR of the genes.

4.9 Most Genes Have Multiple Editing Sites In Tumors

In all 3 cancer type, most of the candidate genes regulated by RNA editing, had multiple regulatory RNA editing sites. Figure 15 shows the heatmap of top 10 genes from all 3 cancer type with the most regulatory RNA editing sites (total 24 genes - some genes are shared for multiple cancer type, Table 8). While a great number of genes (eg: PSMB2, PGAM5, SOAT1, CCNYL1 etc) were edited in all 3 cancer types,

the number of regulatory sites in those genes in THCA were much less compared to BRCA and LUAD. Some of the genes were (eg: UGGT1, APOL1, MRPS16, FAM20B, SYAP) were edited in multiple sites in BRCA and LUAD and those genes were not edited in THCA. DAPK2, TG and ZYB11B were only edited in THCA. Clustering of the editing sites shows that the editing pattern of BRCA and LUAD are much similar to each other compared to THCA.

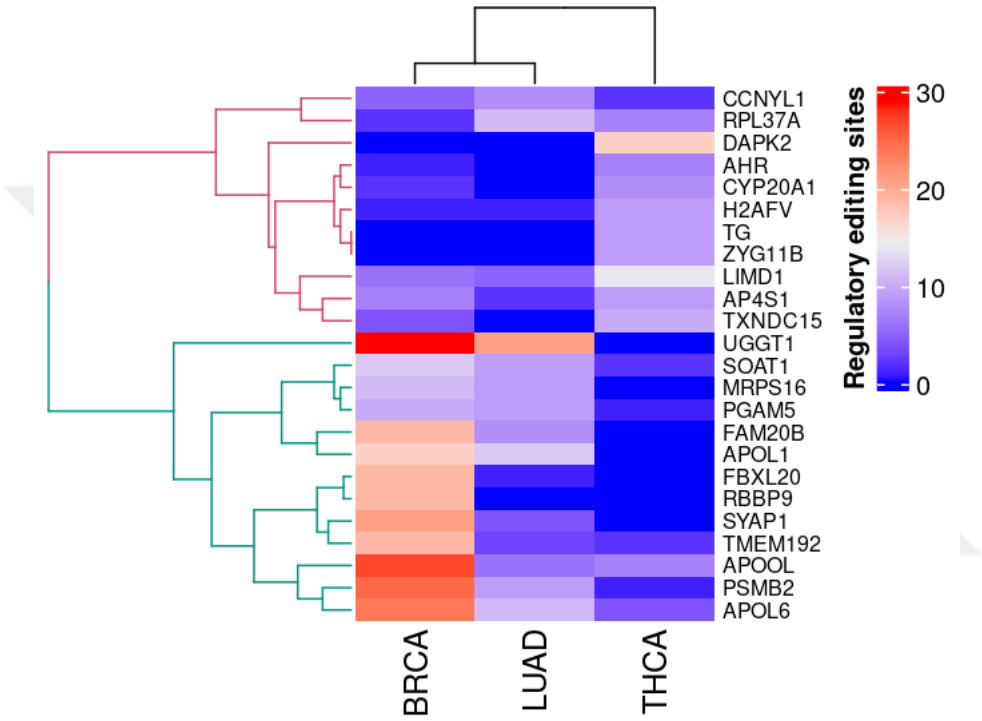

Figure 15. Heatmap of top 10 genes from 3 cancer type with the highest number of regulatory RNA editing sites

Table 8. Genes with the highest number of regulatory editing sites in BRCA, LUAD and THCA.

GENE	BRCA	LUAD	THCA	LOCATION	PROTEIN NAME
UGGT1	29	21	0	3UTR	UDP-glucose glycoprotein glucosyltransferase 1
APOL1	17	12	0	3UTR	Apolipoprotein L1
MRPS16	11	9	0	ncRNA	Mitochondrial Ribosomal Protein S16
FAM20B	19	8	0	3UTR	FAM20B Glycosaminoglycan Xylosylkinase
SYAP1	21	4	0	3UTR	Synapse Associated Protein 1
FBXL20	19	1	0	3UTR	F-Box And Leucine Rich Repeat Protein 20
RBBP9	19	0	0	3UTR	RB Binding Protein 9, Serine Hydrolase
PSMB2	25	9	1	3UTR	Proteasome 20S Subunit Beta 2
PGAM5	10	9	1	3UTR	Mitochondrial Serine/Threonine P. Phosphatase
SOAT1	12	9	2	3UTR	Sterol O-Acyltransferase 1
CCNYL1	5	8	2	3UTR	Cyclin Y Like 1
TMEM192	19	3	2	3UTR	Transmembrane Protein 192
APOL6	24	11	4	3UTR	Apolipoprotein L6
RPL37A	2	11	7	3UTR	Ribosomal Protein L37a
APOOL	27	6	7	3UTR	Apolipoprotein O Like
AHR	1	0	7	3UTR	Aryl Hydrocarbon Receptor
CYP20A1	2	0	8	3UTR	Cytochrome P450 Family 20 Subfamily A Member 1
AP4S1	7	2	9	3UTR	Adaptor Related Protein Complex 4 Subunit Sigma 1
H2AFV	1	1	9	INTRONIC	H2A.Z Variant Histone 2
TG	0	0	9	INTRONIC	Thyroglobulin
ZYG11B	0	0	9	3UTR	Zyg-11 Family Member B, Cell Cycle Regulator
TXNDC15	4	0	10	3UTR	Thioredoxin Domain Containing 15
LIMD1	6	5	14	ncRNA	LIM Domain Containing 1
DAPK2	0	0	17	INTRONIC	Death Associated Protein Kinase 2

Gene set enrichment analysis (GSEA) results of those 24 genes showed that apoptosis regulation was the most significant term with overlapped genes PSMB2 and DAPK2 (Table 9).

Table 9. Gene set enrichment analysis (GSEA) results (Enrichr – Bioplanet 2019) for 24 genes mentioned in Figure 15

Term	p-value	q-value	Genes
Apoptosis regulation	0.00392	0.15555	PSMB2, DAPK2
TSH regulation of gene expression	0.00599	0.15555	TG, AHR
ER quality control compartment (ERQC)	0.00718	0.15555	UGGT1
DCC role in regulating apoptosis	0.01194	0.19399	DAPK2
N-glycan trimming in the ER and calnexin/calreticulin cycle	0.01549	0.20141	UGGT1
G-protein-independent seven transmembrane receptor signaling	0.02257	0.20236	SOAT1
ER-associated degradation (ERAD) pathway	0.02257	0.20236	UGGT1
Proteasome complex	0.02842	0.20236	PSMB2
Steroid biosynthesis	0.03076	0.20236	SOAT1
Apoptosis	0.03379	0.20236	PSMB2, DAPK2

4.10 RNA Editing Is Crucial for Survivability of The Patients In LUAD

The genes with the highest editing-expression correlation in 3 types of cancer patients were also evaluated. Table 10 shows top 10 genes which has the highest mean of spearman coefficients. In BRCA and LUAD, highly correlated genes are positively correlated, and they are negatively correlated in THCA. In BRCA and LUAD, EIF2AK2 and APOL1 are shared amongst highly correlated genes. We evaluated the effect of the editing frequency of those highly correlated regulatory RNA editing sites on the survival of the cancer patients. For a specific RNA editing site, the patients which had editing at that site was grouped into two groups: High or low editing as described in Materials and Methods section. Most of the regulatory sites were shown to be critically more important for survivability of the patients in LUAD (61.7% - 29/47) compared to very few in BRCA (7.8% - 4/51) and THCA (3.1% - 1/32). Figure 16 shows survival plots of 2 examples of RNA editing sites (PSMB2 Chr1:36068196 and ERO1L Chr14:53109810) with high correlation of survivability in LUAD.

Table 10. Details of 10 genes with the highest editing-expression correlation in BRCA, LUAD and THCA

Cancer Type	Gene name	Mean Spearman's ρ	Total sites	Sites correlated with survival
BRCA	DDX58	0.3	6	1
	APOL1	0.275	17	2
	EIF2AK2	0.27	7	0
	ZNF83	0.264	1	0
	FHL2	0.255	1	0
	GM2A	0.25	3	0
	PPM1K	0.249	6	0
	CCRN4L	0.245	1	0
	PIGM	0.243	8	1
	PVR	0.241	1	0
	Total		**51**	**4**
LUAD	ERO1L	0.332	3	3
	MYO19	0.302	1	0
	APOL1	0.298	12	11
	CDC27	0.287	1	1
	EIF2AK2	0.28	3	1
	GPC6	0.26	1	1
	TMPO	0.247	5	2
	CCRN4L	0.232	1	0
	RPL37A	0.228	11	3
	PSMB2	0.224	9	7
	Total		**47**	**29**
THCA	RIC8B	-0.392	1	0
	GCSH	-0.352	1	0
	METTL7A	-0.299	1	0
	TG	-0.27	11	1
	TXNDC15	-0.258	10	0
	SLC25A29	-0.256	4	0
	NDUFC2	-0.254	1	0
	SNED1	-0.244	1	0
	IFNAR2	-0.242	1	0
	NEIL1	-0.239	1	0
	Total		**32**	**1**

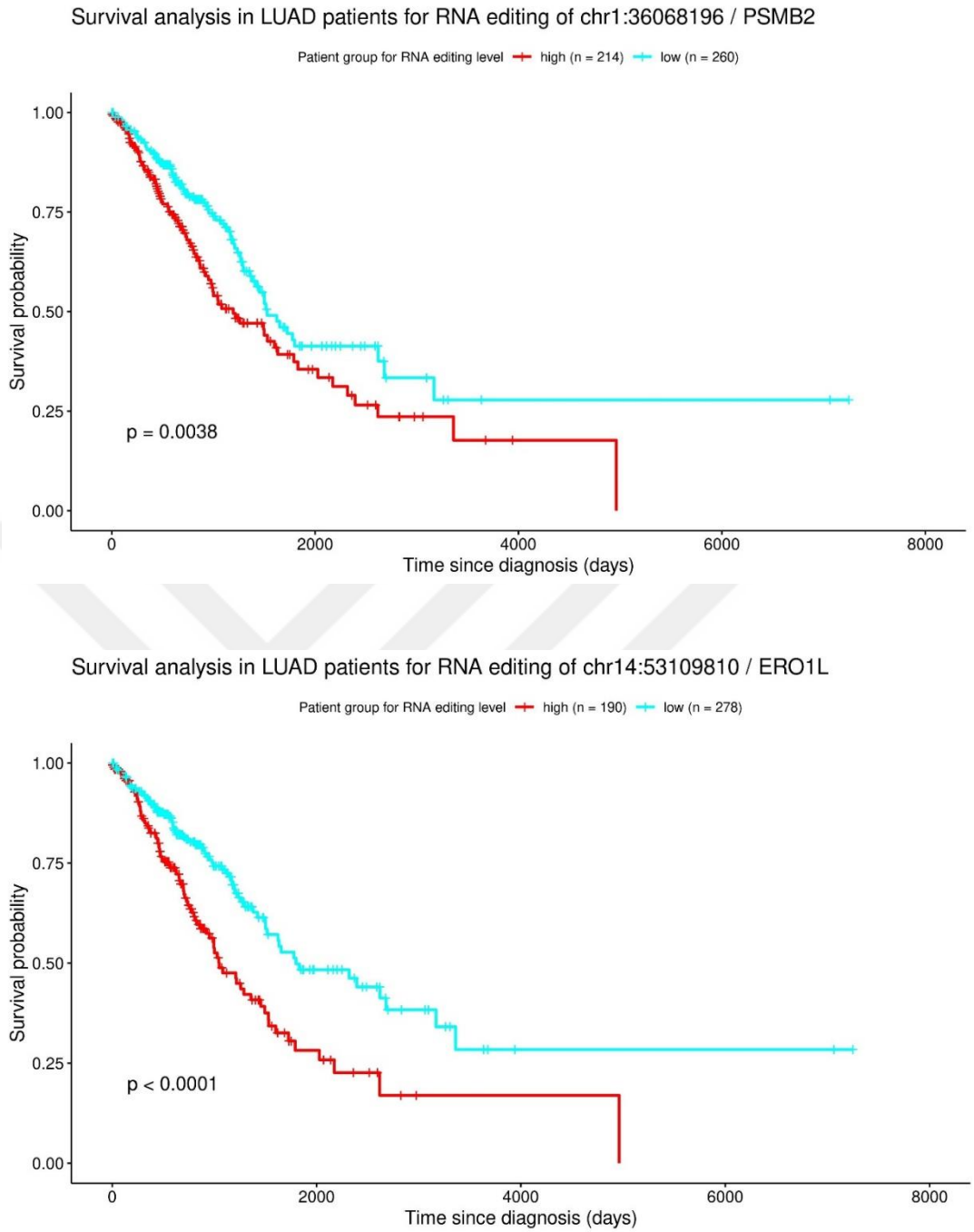

Figure 16. Survival Analysis of LUAD patients for RNA editing of PSMB2 Chr1:36068196 and ERO1L Chr14:53109810

4.11 Testing of the Effect of Regulatory Editing Sites For mRNA – miRNA Interaction

One of the possible ways of RNA editing to regulate mRNA expression is to alter the sequence of miRNA binding sites in 3UTRs of genes. miRNASNP-v3 (129) predicts the target gain or loss effect of SNPs in 3'UTRs using prediction tools TargetScan (130) and miRmap (131). It was previously reported that combination of these two tools have the best prediction performance in human (132). Since RNA editing sites can practically act as SNPs, miRNASNP-v3 online prediction module was used to test the effect of regulatory editing sites on miRNA-mRNA interactions. Regulatory and non-regulatory RNA editing sites were tested whether they resulted in any difference in miRNA target site properties. There was no significant difference in the number of gained or lost miRNA target sites by the effect of RNA editing in all three types of cancer (Table 11). Minimum free energies (MFE - kCal/mol) of the miRNA-mRNA duplexes of the miRNA sites gained or lost were also not different in BRCA and THCA. Only in LUAD, MFE of the miRNA-mRNA duplexes for miRNA sites gained were lower in regulatory sites gained (p-value=0.006). This shows that miRNA target sites gained in LUAD were less stable for regulatory sites compared to non-regulatory sites.

Although the number of miRNA target sites altered by RNA editing were not statistically different between regulatory and non-regulatory sites in all three types of cancer, individual cases show promising results for the effect of RNA editing on miRNA-mRNA interactions. Table 12 shows the results of miRNASNP-v3 prediction for RNA editing sites in PSMB2 gene in all 3 cancer types. PSMB2 is one of highly edited genes in BRCA and LUAD with only one editing point THCA. In all 3 cancer types, the number of miRNA sites lost is higher than the number of miRNA sites gained by RNA editing resulting in lower number of miRNA target sites in PSMB2 3UTR. This is consistent with the fact that increased RNA editing can be correlated with increased RNA expression in PSMB2 (positive mean spearman correlation) in all 3 cancer types. One can speculate that higher mean spearman correlation in LUAD is a result of the high difference between the number of miRNA sites lost and gained in

LUAD resulting in lower number of miRNA target sites compared to BRCA and THCA. The expression of PSMB2 is also higher in tumor samples compared to normal samples in all 3 cancer types consistent with the fact that PSMB2 editing is increased in tumors samples compared to normal samples.

Table 11. miRNASNP-v3 prediction results for regulatory and non-regulatory sites

BRCA	**Regulatory sites (n=1303)**	**Non-regulatory sites (n=3290)**	**p-value**
miRNA sites gained (mean) by editing	5.2 ± 4.1	4.9 ± 3.8	0.28
miRNA sites lost (mean) by editing	4.9 ± 4.1	4.8 ± 4	0.65
MFE of miRNA – mRNA duplex for miRNA sites gained (mean)	-15.9 ± 4	-15.9 ± 4	0.43
MFE of miRNA – mRNA duplex for miRNA sites lost (mean)	-13.8 ± 3.7	-13.7 ± 3.6	0.84
LUAD	(n=370)	(n=1009)	
miRNA sites gained (mean) by editing	5.3 ± 4.1	4.9 ± 3.8	0.14
miRNA sites lost (mean) by editing	5 ± 4.1	5.1 ± 4.2	0.82
MFE of miRNA – mRNA duplex for miRNA sites gained (mean)	-15.7 ± 4.2	-16.3 ± 4	0.006
MFE of miRNA – mRNA duplex for miRNA sites lost (mean)	-13.8 ± 3.7	-13.9 ± 3.6	0.28
THCA	(n=460)	(n=1264)	
miRNA sites gained (mean) by editing	4.6 ± 3.5	4.5 ± 3.5	0.97
miRNA sites lost (mean) by editing	5.4 ± 4.8	5.2 ± 4.4	0.55
MFE of miRNA – mRNA duplex for miRNA sites gained (mean)	-16.3 ± 4.1	-16.2 ± 4.2	0.54
MFE of miRNA – mRNA duplex for miRNA sites lost (mean)	-14.2 ± 3.5	-14.1 ± 3.7	0.39

Table 12. Details of regulatory sites in PSMB2 in 3 types of cancer
(1) Means of spearman coefficients (editing vs expression) of the RNA editing sites in PSMB2 (2) predicted miRNA target site (miRNASNP-v3) gained or lost by RNA editing (3) The ratio of expression of PSMB2 in tumor / normal (TPM)(129)

Cancer type	**BRCA**	**LUAD**	**THCA**
Number of regulatory editing sites	25	9	1
Mean Spearman's ρ (1)	0.16	0.23	0.12
miRNA sites gained (2)	108	34	3
miRNA sites lost (2)	110	52	5
PSMB2 expression ratio (tumor/normal) (3)	1.33-fold	1.52-fold	1.21-fold

4.12 BRCA And LUAD Has A Similar RNA Editing Pattern

31 regulatory sites are shared for BRCA, LUAD and THCA (Figure 17). Figure 18 shows the heatmap of correlation coefficients those 31 shared sites. Similar to the the clustering of the number of editing sites in 3 types of cancer (Figure 15), RNA editing frequency clustering shows that regulatory site editing pattern in BRCA and LUAD are more similar to each other compared to THCA.

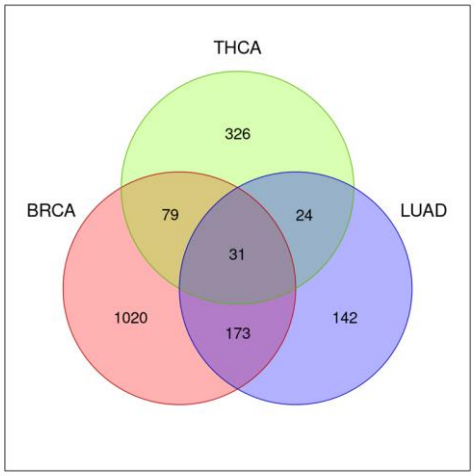

Figure 17. Distribution of regulatory sites shared in 3 cancer cell type

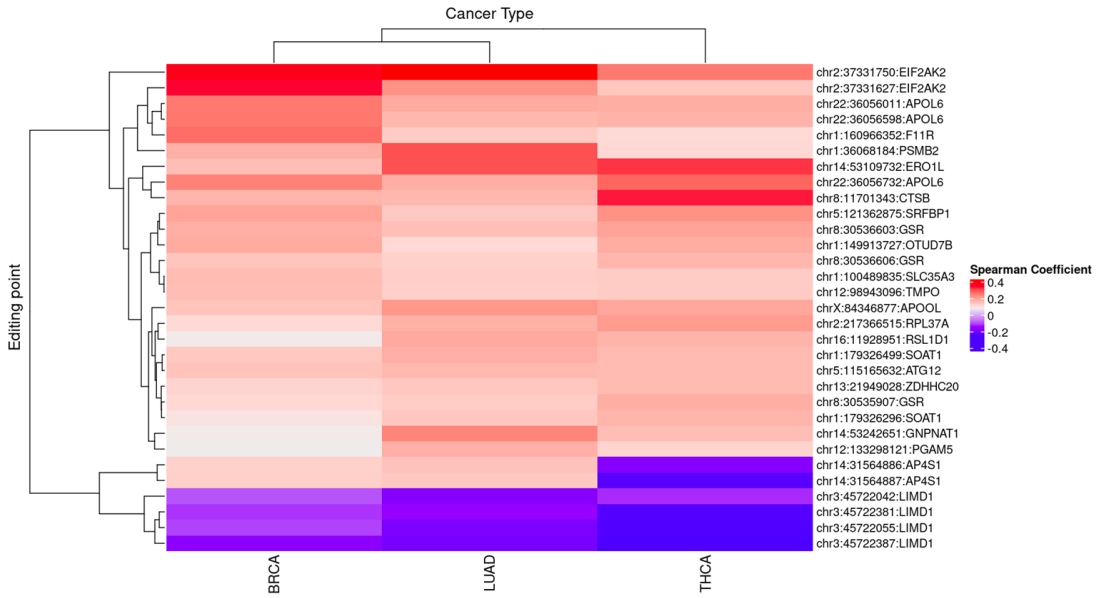

Figure 18. Heatmap of 31 regulatory sites shared in BRCA, LUAD and THCA.

5 DISCUSSION

RNA-seq is increasingly utilized to explore RNA editing sites across the transcriptome. Yet, the computational processing of RNA-seq data often leads to mapping errors, creating numerous false-positive RNA editing sites. To address this, machine learning (ML) techniques have been employed to leverage data from confirmed RNA editing sites to forecast new ones. Several research teams have formulated ML-based approaches for this task, including (15–18).

In the first part of my thesis, I developed a relatively straightforward method that utilizes sequence and structure information from experimentally validated A-to-I RNA editing sites for new predictions. The structural model incorporates the dot-bracket notation for RNA structure along with triplet sequence data. Alternatively, the second approach uses only triplet sequence data, omitting structural information. Both methods' outputs were trained using a support vector machine (SVM) to create a classifier. Interestingly, the sequence model outperformed the structural model. This outcome was unexpected given that both the adjacent structure and the primary sequence are considered crucial for editing specificity and efficiency, as noted by (133). This discrepancy might be due to the influence of distant elements on RNA editing specificity. For instance, structural elements located more than 150 nucleotides away from the edited adenosine in the Gabra-3 gene have been shown to affect both the efficiency and specificity of editing, as discussed by (134). Daniel and colleagues identified editing inducer elements (EIEs) that generally regulate editing at specific sites. Since the study only included structural data up to 150 nucleotides from the editing site, the effects of these distant elements were not observed on the predictions.

During the optimization and validation of the SVM using experimental data, the sensitivity of our predictions was higher than the specificity. This indicated that there were fewer false negatives in our predictions. This issue may stem from how we created the negative dataset for the SVM predictor. Sequences with a central adenine from the reference genome that were not listed in the Rediportal database were randomly selected for the negative dataset, similar to approaches used in other studies

(15,17). Consequently, a small fraction of the sequences in the negative dataset might be unknown, real RNA editing sites, which could affect the classifier's specificity. This challenge is often unavoidable due to the inherent difficulties in creating a negative dataset for such machine learning problems. Additionally, although Bahn's study suggests that ADAR2 expression is significantly lower than ADAR1, it is still possible that ADAR2 remains active in small amounts, possibly creating genuine editing sites even in ADAR1 knockdown cells, which could further impact the SVM's specificity during validation. A double knockdown of both ADAR1 and ADAR2 might yield more accurate results for our method, but such a study is not currently available for testing.

During the SVM predictor validation using experimental data, REDItools was utilized to identify A-to-I editing sites and excluded known SNPs. This process yielded 566 A-G editing sites in the control group and 105 A-G editing sites in the ADAR1 knockdown group, indicating that RNA editing is diminished following ADAR1 knockdown. Of the 516 A-G editing sites exclusive to the control group (true editing sites), the SVM successfully predicted 479 as true A-G editing sites. In the ADAR1 knockdown group, where 105 A-G editing sites were identified (considered false editing sites), the SVM accurately identified 81 of these sites as false, resulting in a sensitivity of 92.8%, specificity of 77.1%, and overall accuracy of 90.2%. Additionally, when testing the SVM predictor with the dataset from Bahn's study, out of 4141 true editing sites identified in control siRNA knockdown cells, the SVM correctly predicted 3635 as true editing sites, achieving a true positive rate (TPR) of 87.8%.

In second part of the thesis, the correlation of RNA editing and expression of genes were investigated in tumor samples obtained from (9) in which RNA editing data was generated from RNAseq results of TCGA tumor and non-tumor tissue samples. Out of the RNA editing sites with increased editing between tumor and non-tumor samples, in nearly %25 of those sites, the editing frequencies were correlated with expression of the genes in BRCA, THCA and LUAD (regulatory sites).

In BRCA and LUAD, the regulatory sites were enriched in 3'UTR of genes compared to intronic regions in non-regulatory sites. As a result of this tendency for regulatory sites to be in 3'UTR of genes, I hypothesized that this may support the possibility of regulation of gene expression by editing of miRNA binding and preventing the interaction of miRNAs with 3'UTRs of those mRNAs. I tested if the editing of regulatory and non-regulatory RNA editing sites in BRCA, LUAD and THCA had any affect in miRNA – mRNA interactions. Although there was no overall significant difference in the number of gained or lost miRNA target sites by RNA editing in all three types of cancer, in some cases (e.g.: PSMB2) the number of miRNA target sites were significantly decreased suggesting that RNA editing may influence the expression of the gene.

Another mechanism for regulation of gene expression by ADAR1 is the regulation of RNA stability by human antigen R (HuR). HuR selectively binds to single-stranded AU-rich RNA sequences and stabilizes RNA transcripts in normal (135) and cancer cells (136). ADAR1 is shown to recruit and bind HuR protein increasing mRNA stability. Wang and colleagues (6) showed that, in ADAR1 knockdowns, out of 775 genes whose expression levels decreased, the genes containing HuR binding sites showed significantly higher expression than others. Although I was not able to evaluate for ADAR1-HuR binding, we believe this mechanism may be another reasonable option for regulation of carcinogenesis in our samples and experimental studies are crucial for the testing of those possible regulatory mechanisms.

In this study, the impact of the editing of regulatory sites on the prognosis of the cancer patients was also evaluated. Interestingly, out of 3 cancer types, the RNA editing sites in LUAD tumor samples were mainly vital for the prognosis of the patient. This is consistent with other studies which indicated the clinical importance of A-to-I RNA editing profiles in lung adenocarcinoma (137–139). Out of the genes with the highest editing/expression correlation in LUAD patients, EROL1, PSMB2 and APOL1 had multiple editing points which many of them are found to be correlated with prognosis of the patients. ERO1L is an oxidoreductase involved in disulfide bond formation in the endoplasmic reticulum and plays an important role in ER stress-

induced apoptosis. Expression of ERO1L is shown to be a potential biomarker in LUAD (140) and it promotes the proliferation and metastasis of lung adenocarcinoma via the Wnt2/β-catenin signaling pathway (141). In this study, we showed that RNA editing of ERO1L can act as a potential biomarker in LUAD.

PSMB2 gene encodes for proteasome subunit beta type 2, a crucial subunit of 26S proteasome complex which plays a key role in the in removing misfolded or damaged proteins that could impair cellular functions. It is shown that PSMB2 knockdown suppressed proteasome activity and cell proliferation and promoted apoptosis, in gastric cancer cells (142) showing the importance of PSMB2 in carcinogenesis. Like ERO1L, to our knowledge this is first study to show that RNA editing of PSMB2 can also be a potential biomarker for LUAD.

APOL1 gene encodes apolipoprotein L1 (apoL1), an apolipoprotein component of high-density lipoprotein (HDL) which is synthesized in liver, pancreas, kidney, and brain. Lin and colleagues (143) showed that APOL1 was elevated in human pancreatic cancer tissues compared with that in adjacent tissues and was associated with poor prognosis. Their results showed that knockdown of APOL1 significantly inhibited the proliferation and promoted apoptosis in pancreatic cancer. Sharpnack and colleagues (128) have performed a similar methodology to our study and identified APOL1 as a potentially novel cancer-related gene, and discovered that editing of 3′-UTR of APOL1 may change the RNA secondary structure altering the binding of RNA binding proteins (RBPs) and miRNAs.

EROL1, PSMB2 and APOL1 are considered as oncogenes. Our analysis showed that, for all those genes, the survival of the LUAD patients were better in the low editing groups (hence low expression) which is consistent with the fact that increased editing of those genes is correlated with increased gene expression even though the actual mechanism is not clear. We believe the quantification of those single editing events in EROL1, PSMB2 and APOL1 (and many more new candidates) can be used as prognostic biomarkers in tumor tissue of LUAD patients and may be a useful tool as a therapeutic guide.

6 CONCLUSION

In the first part of my thesis study, I have demonstrated that a novel and relatively straightforward machine learning approach, utilizing triplet sequence information from established RNA editing sites, can effectively predict potential RNA editing sites in RNAseq data. This method's simplicity, compared to other similar studies, results in less computational demand while maintaining high performance. I have developed an open-source software named RDDSVM that facilitates the prediction of candidate A-to-I RNA editing sites. RDDSVM is user-friendly, designed for users with limited bioinformatics expertise, and is compatible with major operating systems. In the second part of my thesis, I identified novel prognostic RNA editing markers in the EROL1 and PSMB2 genes in LUAD patients. The quantification of these specific editing events can serve as predictive indicators in the tumor tissue of LUAD patients, potentially providing guidance for their therapy.

7 REFERENCES

1. Zhou ZY, Hu Y, Li A, Li YJ, Zhao H, Wang SQ, et al. Genome wide analyses uncover allele-specific RNA editing in human and mouse. Nucleic Acids Res. 2018;46(17)::8888-8897.
2. Mehedi M, Hoenen T, Robertson S, Ricklefs S, Dolan MA, Taylor T, et al. Ebola Virus RNA Editing Depends on the Primary Editing Site Sequence and an Upstream Secondary Structure. PLoS Pathog. 2013;9(10):e1003677.
3. Randau L, Stanley BJ, Kohlway A, Mechta S, Xiong Y, SöLLL D. A cytidine deaminase edits c to u in transfer RNAs in archaea. Science (80-). 2009;324(5927):657–9.
4. Bar-Yaacov D, Mordret E, Towers R, Biniashvili T, Soyris C, Schwartz S, et al. RNA editing in bacteria recodes multiple proteins and regulates an evolutionarily conserved toxin-antitoxin system. Genome Res. 2017;27(10):1696–1703.
5. Jepson JEC, Reenan RA. RNA editing in regulating gene expression in the brain. Biochim Biophys Acta - Gene Regul Mech. 2008;1779(8):459–70.
6. Wang IX, So E, Devlin JL, Zhao Y, Wu M, Cheung VG. ADAR Regulates RNA Editing, Transcript Stability, and Gene Expression [Internet]. Vol. 5, Cell Reports. The Authors; 2013. 849–860 p. Available from: http://dx.doi.org/10.1016/j.celrep.2013.10.002
7. Laurencikiene J, Källman AM, Fong N, Bentley DL, Öhman M. RNA editing and alternative splicing: The importance of co-transcriptional coordination. EMBO Rep. 2006;7(3):303–307.
8. Brümmer A, Yang Y, Chan TW, Xiao X. Structure-mediated modulation of mRNA abundance by A-to-I editing. Nat Commun. 2017;8(1):1255.
9. Han L, Diao L, Yu S, Xu X, Li J, Zhang R, et al. The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. Cancer Cell. 2015;28(4):515–28.
10. Paz N, Levanon EY, Amariglio N, Heimberger AB, Ram Z, Constantini S, et al. Altered adenosine-to-inosine RNA editing in human cancer. Genome Res. 2007;17(11):1586–95.
11. Maas S, Kawahara Y, Tamburro KM, Nishikura K. A-to-I RNA editing and human disease. RNA Biol. 2006;3(1):1–9.
12. Kleinman CL, Majewski J. Comment on “Widespread RNA and DNA sequence differences in the human transcriptome.” Science (80-). 2012;335(6074):1302.
13. Sakurai M, Ueda H, Yano T, Okada S, Terajima H, Mitsuyama T, et al. A biochemical landscape of A-to-I RNA editing in the human brain transcriptome. Genome Res. 2014;24(3):522–534.
14. Piskol R, Peng Z, Wang J, Li JB. Lack of evidence for existence of noncanonical RNA editing. Nat Biotechnol. 2013;31(1):19–20.
15. Sun J, De Marinis Y, Osmark P, Singh P, Bagge A, Valtat B, et al. Discriminative prediction of A-To-I RNA editing events from DNA sequence. PLoS One. 2016;11(10):e0164962.
16. Kim M su, Hur B, Kim S. RDDpred: A condition-specific RNA-editing prediction model from RNA-seq data. BMC Genomics. 2016;17(5).
17. Chen W, Feng P, Yang H, Ding H, Lin H, Chou KC. IRNA-AI: Identifying the adenosine to inosine

editing sites in RNA sequences. Oncotarget. 2017;8(3):4208–17.
18. Ouyang Z, Liu F, Zhao C, Ren C, An G, Mei C, et al. Accurate identification of RNA editing sites from primitive sequence with deep neural networks. Sci Rep. 2018;8:6005.
19. Xue C, Li F, He T, Liu GP, Li Y, Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics. 2005;6:310.
20. Das SG, Chakraborty HJ, Datta A. PremipreD: Precursor miRNA prediction by support vector machine approach. Trends Bioinforma. 2018;11(1):17–24.
21. Chow LC and YL and CHL and THMC and RKK. Recoding RNA editing of antizyme inhibitor 1 predisposes to hepatocellular carcinoma. Nat Med. 2013;19(2):209–16.
22. Zhang L, Yang C-S, Varelas X, Monti S. Altered RNA editing in 3' UTR perturbs microRNA-mediated regulation of oncogenes and tumor-suppressors. Sci Rep. 2016 Mar;6:23226.
23. Salameh A, Lee AK, Cardó-Vila M, Nunes DN, Efstathiou E, Staquicini FI, et al. PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. Proc Natl Acad Sci U S A. 2015;112(27):8403–8.
24. Hajji K, Sedmík J, Cherian A, Amoruso D, Keegan LP, O'connell MA. ADAR2 enzymes: efficient site-specific RNA editors with gene therapy aspirations. Rna. 2022;28(10):1281–97.
25. Melcher T, Maas S, Herb A, Sprengel R, Seeburg PH, Higuchi M. A mammalian RNA editing enzyme. Vol. 379, Nature. 1996. p. 460–4.
26. Nishikura K. A-to-I editing of coding and non-coding RNAs by ADARs. Nat Rev Mol Cell Biol. 2016;17(2):83–96.
27. Bhakta S, Tsukahara T. C-to-U RNA Editing: A Site Directed RNA Editing Tool for Restoration of Genetic Code. Genes (Basel). 2022;13(9).
28. Warren CJ, Santiago ML, Pyeon D. APOBEC3: Friend or Foe in Human Papillomavirus Infection and Oncogenesis? Annu Rev Virol [Internet]. 2022;9(Volume 9, 2022):375–95. Available from: https://www.annualreviews.org/content/journals/10.1146/annurev-virology-092920-030354
29. Pecori R, Di Giorgio S, Paulo Lorenzo J, Nina Papavasiliou F. Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination. Nat Rev Genet [Internet]. 2022;23(8):505–18. Available from: https://doi.org/10.1038/s41576-022-00459-8
30. Blanc V, Xie Y, Kennedy S, Riordan JD, Rubin DC, Madison BB, et al. Apobec1 complementation factor (A1CF) and RBM47 interact in tissue-specific regulation of C to U RNA editing in mouse intestine and liver. Rna. 2019;25(1):70–81.
31. Revathidevi S, Murugan AK, Nakaoka H, Inoue I, Munirajan AK. APOBEC: A molecular driver in cervical cancer pathogenesis. Cancer Lett [Internet]. 2021;496:104–16. Available from: https://www.sciencedirect.com/science/article/pii/S0304383520305036
32. Oakes E, Anderson A, Cohen-Gadol A, Hundley HA. Adenosine deaminase that acts on RNA 3 (adar3) binding to glutamate receptor subunit B Pre-mRNA Inhibits RNA editing in glioblastoma. J Biol Chem. 2017;292(10):4326–35.
33. Chen CX, Cho DS, Wang Q, Lai F, Carter K, Nishikura K. A third member of the RNA-specific

adenosine deaminase gene binding domains. Rna. 2000;6(6):755–67.

34. Thomas JM, Beal PA. How do ADARs bind RNA? New protein-RNA structures illuminate substrate recognition by the RNA editing ADARs. BioEssays. 2017;39(4):1–17.
35. Christofi T, Zaravinos A. RNA editing in the forefront of epitranscriptomics and human health. J Transl Med [Internet]. 2019;17(1):1–15. Available from: https://doi.org/10.1186/s12967-019-2071-4
36. Raghava Kurup R, Oakes EK, Manning AC, Mukherjee P, Vadlamani P, Hundley HA. RNA binding by ADAR3 inhibits adenosine-to-inosine editing and promotes expression of immune response protein MAVS. J Biol Chem [Internet]. 2022;298(9):102267. Available from: https://doi.org/10.1016/j.jbc.2022.102267
37. Pfaller CK, George CX, Samuel CE. Adenosine Deaminases Acting on RNA (ADARs) and Viral Infections. Annu Rev Virol. 2021;8:239–64.
38. Doherty EE, Karki A, Wilcox XE, Mendoza HG, Manjunath A, Matos VJ, et al. ADAR activation by inducing a syn conformation at guanosine adjacent to an editing site. Nucleic Acids Res. 2022;50(19):10857–68.
39. Jacobsen CS, Salvador P, Yung JF, Kragness S, Mendoza HG, Mandel G, et al. Library Screening Reveals Sequence Motifs That Enable ADAR2 Editing at Recalcitrant Sites. ACS Chem Biol [Internet]. 2023 Oct 20;18(10):2188–99. Available from: https://doi.org/10.1021/acschembio.3c00107
40. Gabay O, Shoshan Y, Kopel E, Ben-Zvi U, Mann TD, Bressler N, et al. Landscape of adenosine-to-inosine RNA recoding across human tissues. Nat Commun [Internet]. 2022;13(1):1184. Available from: https://doi.org/10.1038/s41467-022-28841-4
41. Nishikura K. A-to-I editing of coding and non-coding RNAs by ADARs. Nat Rev Mol Cell Biol. 2016;17(2):83–96.
42. Wang F, Cao H, Xia Q, Liu Z, Wang M, Gao F, et al. Lessons from discovery of true ADAR RNA editing sites in a human cell line. BMC Biol [Internet]. 2023;21(1):160. Available from: https://doi.org/10.1186/s12915-023-01651-w
43. Li JB, Levanon EY, Yoon J-K, Aach J, Xie B, LeProust E, et al. Genome-Wide Identification of Human RNA Editing Sites by Parallel DNA Capturing and Sequencing. Science (80-) [Internet]. 2009;29;324(5931):1210–3. Available from: https://doi.org/10.1126/science.1170995
44. Cuddleston WH, Fan X, Sloofman L, Liang L, Mossotto E, Moore K, et al. Spatiotemporal and genetic regulation of A-to-I editing throughout human brain development. Cell Rep [Internet]. 2022;41(5):111585. Available from: https://doi.org/10.1016/j.celrep.2022.111585
45. Bass BL. RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem. 2002;71(I):817–46.
46. Tang SJ, Shen H, An O, Hong H, Li J, Song Y, et al. Cis- and trans-regulations of pre-mRNA splicing by RNA editing enzymes influence cancer development. Nat Commun [Internet]. 2020;11(1):799. Available from: https://doi.org/10.1038/s41467-020-14621-5
47. Zhang Q, Xiu B, Zhang L, Chen M, Chi W, Li L, et al. Immunosuppressive lncRNA LINC00624

promotes tumor progression and therapy resistance through ADAR1 stabilization. J Immunother cancer. 2022 Oct;10(10).

48. Wong T-L, Loh J-J, Lu S, Yan HHN, Siu HC, Xi R, et al. ADAR1-mediated RNA editing of SCD1 drives drug resistance and self-renewal in gastric cancer. Nat Commun [Internet]. 2023;14(1):2861. Available from: https://doi.org/10.1038/s41467-023-38581-8
49. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321–49.
50. Yu Z, Luo R, Li Y, Li X, Yang Z, Peng J, et al. ADAR1 inhibits adipogenesis and obesity by interacting with Dicer to promote the maturation of miR-155-5P. J Cell Sci. 2022 Mar;135(5).
51. Kapoor U, Licht K, Amman F, Martin D, Dieterich C, Jantsch MF. ADAR-deficiency perturbs landscape in mouse tissues the global splicing Running Title : A-to-I RNA editing impacts pre-mRNA splicing Authors :
52. Gatsiou A, Stellos K. RNA modifications in cardiovascular health and disease. Nat Rev Cardiol [Internet]. 2023;20(5):325–46. Available from: https://doi.org/10.1038/s41569-022-00804-8
53. Liang Z, Goradia A, Walkley CR, Heraud-Farlow JE. Generation of a new Adar1p150 (-/-) mouse demonstrates isoform-specific roles in embryonic development and adult homeostasis. RNA. 2023 Sep;29(9):1325–38.
54. Mladenova D, Barry G, Konen LM, Pineda SS, Guennewig B, Avesson L, et al. Adar3 Is Involved in Learning and Memory in Mice. Front Neurosci. 2018;12:243.
55. Nakahama T, Kawahara Y. The RNA-editing enzyme ADAR1: a regulatory hub that tunes multiple dsRNA-sensing pathways. Int Immunol. 2023 Mar;35(3):123–33.
56. Karki R, Kanneganti T-D. ADAR1 and ZBP1 in innate immunity, cell death, and disease. Trends Immunol. 2023 Mar;44(3):201–16.
57. Hur S. Double-Stranded RNA Sensors and Modulators in Innate Immunity. Annu Rev Immunol. 2019 Apr;37:349–75.
58. Guo X, Liu S, Sheng Y, Zenati M, Billiar T, Herbert A, et al. ADAR1 Zα domain P195A mutation activates the MDA5-dependent RNA-sensing signaling pathway in brain without decreasing overall RNA editing. Cell Rep. 2023 Jul;42(7):112733.
59. Tang Q, Rigby RE, Young GR, Hvidt AK, Davis T, Tan TK, et al. Adenosine-to-inosine editing of endogenous Z-form RNA by the deaminase ADAR1 prevents spontaneous MAVS-dependent type I interferon responses. Immunity. 2021 Sep;54(9):1961-1975.e5.
60. Kleinova R, Rajendra V, Leuchtenberger AF, Lo Giudice C, Vesely C, Kapoor U, et al. The ADAR1 editome reveals drivers of editing-specificity for ADAR1-isoforms. Nucleic Acids Res. 2023 May;51(9):4191–207.
61. Riella C V, McNulty M, Ribas GT, Tattersfield CF, Perez-Gill C, Eichinger F, et al. ADAR regulates APOL1 via A-to-I RNA editing by inhibition of MDA5 activation in a paradoxical biological circuit. Proc Natl Acad Sci U S A. 2022 Nov;119(44):e2210150119.
62. Liddicoat BJ, Piskol R, Chalk AM, Ramaswami G, Higuchi M, Hartner JC, et al. RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. Science. 2015

Sep;349(6252):1115–20.
63. Orsolic I, Carrier A, Esteller M. Genetic and epigenetic defects of the RNA modification machinery in cancer. Trends Genet. 2023 Jan;39(1):74–88.
64. Wu S, Fan Z, Kim P, Huang L, Zhou X. The Integrative Studies on the Functional A-to-I RNA Editing Events in Human Cancers. Genomics Proteomics Bioinformatics. 2023 Jun;21(3):619–31.
65. Baker AR, Slack FJ. ADAR1 and its implications in cancer development and treatment. Trends Genet. 2022 Aug;38(8):821–30.
66. Lotsof ER, Krajewski AE, Anderson-Steele B, Rogers J, Zhang L, Yeo J, et al. NEIL1 Recoding due to RNA Editing Impacts Lesion-Specific Recognition and Excision. J Am Chem Soc. 2022 Aug;144(32):14578–89.
67. Xing Y, Nakahama T, Wu Y, Inoue M, Kim JI, Todo H, et al. RNA editing of AZIN1 coding sites is catalyzed by ADAR1 p150 after splicing. J Biol Chem. 2023 Jul;299(7):104840.
68. Luo J, Gong L, Yang Y, Zhang Y, Liu Q, Bai L, et al. Enhanced mitophagy driven by ADAR1-GLI1 editing supports the self-renewal of cancer stem cells in HCC. Hepatology. 2024 Jan;79(1):61–78.
69. Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res. 2012;22(1):142–50.
70. Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res. 2014 Mar;24(3):365–76.
71. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Jan;17:13.
72. Lee J-H, Ang JK, Xiao X. Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. RNA. 2013 Jun;19(6):725–32.
73. Chhangawala S, Rudy G, Mason CE, Rosenfeld JA. The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biol. 2015 Jun;16(1):131.
74. Carmi S, Borukhov I, Levanon EY. Identification of widespread ultra-edited human RNAs. PLoS Genet. 2011 Oct;7(10):e1002317.
75. Cho H, Davis J, Li X, Smith KS, Battle A, Montgomery SB. High-resolution transcriptome analysis with long-read RNA sequencing. PLoS One. 2014;9(9):e108095.
76. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009 Oct;37(18):e123.
77. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, et al. Widespread RNA and DNA Sequence Differences in the Human Transcriptome. Science (80-) [Internet]. 2011 Jul 1;333(6038):53–8. Available from: https://doi.org/10.1126/science.1207018
78. John D, Weirick T, Dimmeler S, Uchida S. RNAEditor: easy detection of RNA editing events and the introduction of editing islands. Brief Bioinform. 2017 Nov;18(6):993–1001.
79. Picardi E, Pesole G. REDItools: High-throughput RNA editing detection made easy.

Bioinformatics. 2013;29(14):1813–4.

80. Flati T, Gioiosa S, Spallanzani N, Tagliaferri I, Diroma MA, Pesole G, et al. HPC-REDItools: a novel HPC-aware tool for improved large scale RNA-editing analysis. BMC Bioinformatics. 2020 Aug;21(Suppl 10):353.
81. Zhang Q. Analysis of RNA Editing Sites from RNA-Seq Data Using GIREMI. Methods Mol Biol. 2018;1751:101–8.
82. Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res. 1999 Aug;9(8):677–9.
83. Liu Z, Quinones-Valdez G, Fu T, Huang E, Choudhury M, Reese F, et al. L-GIREMI uncovers RNA editing sites in long-read RNA-seq. Genome Biol [Internet]. 2023;24(1):171. Available from: https://doi.org/10.1186/s13059-023-03012-w
84. Wang Z, Lian J, Li Q, Zhang P, Zhou Y, Zhan X, et al. RES-Scanner: a software package for genome-wide identification of RNA-editing sites. Gigascience [Internet]. 2016 Dec 1;5(1):s13742-016-0143–4. Available from: https://doi.org/10.1186/s13742-016-0143-4
85. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics [Internet]. 2009 Jul 15;25(14):1754–60. Available from: https://doi.org/10.1093/bioinformatics/btp324
86. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.
87. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature [Internet]. 2015;526(7571):68–74. Available from: https://doi.org/10.1038/nature15393
88. The International HapMap Project. Nature. 2003 Dec;426(6968):789–96.
89. Piechotta M, Wyler E, Ohler U, Landthaler M, Dieterich C. JACUSA: site-specific identification of RNA editing events from replicate sequencing data. BMC Bioinformatics [Internet]. 2017;18(1):7. Available from: https://doi.org/10.1186/s12859-016-1432-8
90. Zhang F, Lu Y, Yan S, Xing Q, Tian W. SPRINT: an SNP-free toolkit for identifying RNA editing sites. Bioinformatics. 2017 Nov;33(22):3538–48.
91. Deo RC. Machine Learning in Medicine. Circulation. 2015 Nov;132(20):1920–30.
92. Lo Vercio L, Amador K, Bannister JJ, Crites S, Gutierrez A, MacDonald ME, et al. Supervised machine learning tools: a tutorial for clinicians. J Neural Eng [Internet]. 2020;17(6):62001. Available from: https://dx.doi.org/10.1088/1741-2552/abbff2
93. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol. 2020 Feb;9(2):14.
94. Pisner DA, Schnyer DM. Chapter 6 - Support vector machine. In: Mechelli A, Vieira SBT-ML, editors. Academic Press; 2020. p. 101–21. Available from: https://www.sciencedirect.com/science/article/pii/B9780128157398000067
95. Breiman L. Random Forests. Mach Learn [Internet]. 2001;45(1):5–32. Available from:

https://doi.org/10.1023/A:1010933404324
96. Breiman L. Bagging predictors. Mach Learn [Internet]. 1996;24(2):123–40. Available from: https://doi.org/10.1007/BF00058655
97. Ishwaran H, Kogalur U, Gorodeski E, Minn A, Lauer M. High-Dimensional Variable Selection for Survival Data. J Am Stat Assoc. 2010 Mar 1;105:205–17.
98. Wickramasinghe I, Kalutarage H. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation Indika. Soft Comput. 2021 Feb 1;24.
99. Popescu M-C, Balas V, Perescu-Popescu L, Mastorakis N. Multilayer perceptron and neural networks. WSEAS Trans Circuits Syst. 2009 Jul 1;8.
100. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. Crit Care. 2005 Feb;9(1):112–8.
101. Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res [Internet]. 2014 Jan 1;42(D1):D109–13. Available from: https://doi.org/10.1093/nar/gkt996
102. Kiran A, Baranov P V. DARNED: a DAtabase of RNa EDiting in humans. Bioinformatics [Internet]. 2010 Jul 15;26(14):1772–6. Available from: https://doi.org/10.1093/bioinformatics/btq285
103. Xiong H, Liu D, Li Q, Lei M, Xu L, Wu L, et al. RED-ML: a novel, effective RNA editing detection method based on machine learning. Gigascience. 2017 May;6(5):1–8.
104. Alon S, Erew M, Eisenberg E. DREAM: a webserver for the identification of editing sites in mature miRNAs using deep sequencing data. Bioinformatics [Internet]. 2015 Aug 1;31(15):2568–70. Available from: https://doi.org/10.1093/bioinformatics/btv187
105. Nigita G, Alaimo S, Ferro A, Giugno R, Pulvirenti A. Knowledge in the investigation of A-to-I RNA editing signals. Front Bioeng Biotechnol. 2015;
106. Yao L, Wang H, Song Y, Dai Z, Yu H, Yin M, et al. Large-scale prediction of ADAR-mediated effective human A-to-I RNA editing. Brief Bioinform [Internet]. 2019 Jan 18;20(1):102–9. Available from: https://doi.org/10.1093/bib/bbx092
107. Picardi E, D'Erchia AM, Giudice C Lo, Pesole G. REDIportal: A comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res. 2017;45(D1)::D750-D757.
108. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie Chem Mon. 1994;25:167–188.
109. Lehmann KA, Bass BL. The importance of internal loops within RNA substrates of ADAR1. J Mol Biol. 1999;
110. Tian N, Yang Y, Sachsenmaier N, Muggenhumer D, Bi J, Waldsich C, et al. A structural determinant required for RNA editing. Nucleic Acids Res. 2011;
111. Wong SK, Sato S, Lazinski DW. Substrate recognition by ADAR1 and ADAR2. RNA. 2001;
112. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24(12):1565–7.
113. e1071 [Internet]. Available from: https://cran.r-project.org/web/packages/e1071/index.html
114. libSVM [Internet]. Available from: https://www.csie.ntu.edu.tw/~cjlin/libsvm/

115. Chen YW, Lin CJ. Combining SVMs with various feature selection strategies. In: Studies in Fuzziness and Soft Computing. 2006. p. 315–24.
116. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–2079.
117. Picard [Internet]. Available from: https://broadinstitute.github.io/picard
118. SnpNexus [Internet]. Available from: https://www.snp-nexus.org/v4/
119. RDDSVM [Internet]. Available from: https://github.com/huseyintac/RDDSVM
120. Peng Z, Cheng Y, Tan BCM, Kang L, Tian Z, Zhu Y, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol. 2012;
121. Airliner [Internet]. Available from: http://alpha.dmi.unict.it/airliner/
122. InosinePredict [Internet]. Available from: http://hci-bio-app.hci.utah.edu:8081/Bass/InosinePredict
123. TCGAbiolinks [Internet]. Available from: https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html
124. Annotatr [Internet]. Available from: https://www.bioconductor.org/packages/release/bioc/html/annotatr.html
125. miRNASNP v-3 [Internet]. Available from: http://bioinfo.life.hust.edu.cn/miRNASNP/#!/tools
126. Stellos K, Gatsiou A, Stamatelopoulos K, Perisic Matic L, John D, Lunella FF, et al. Adenosine-to-inosine RNA editing controls cathepsin S expression in atherosclerosis by enabling HuR-mediated post-transcriptional regulation. Nat Med. 2016 Oct;22(10):1140–50.
127. Yang C-C, Chen Y-T, Chang Y-F, Liu H, Kuo Y-P, Shih C-T, et al. ADAR1-mediated 3' UTR editing and expression control of antiapoptosis genes fine-tunes cellular apoptosis response. Cell Death Dis. 2017 May;8(5):e2833.
128. Sharpnack MF, Chen B, Aran D, Kosti I, Sharpnack DD, Carbone DP, et al. Global Transcriptome Analysis of RNA Abundance Regulation by ADAR in Lung Adenocarcinoma. EBioMedicine [Internet]. 2018;27:167–75. Available from: https://doi.org/10.1016/j.ebiom.2017.12.005
129. Liu CJ, Fu X, Xia M, Zhang Q, Gu Z, Guo AY. MiRNASNP-v3: A comprehensive database for SNPs and disease-related variations in miRNAs and miRNA targets. Nucleic Acids Res. 2021;49(D1):D1276–81.
130. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 2015;4(AUGUST2015):1–38.
131. Vejnar CE, Zdobnov EM. MiRmap: Comprehensive prediction of microRNA target repression strength. Nucleic Acids Res. 2012;40(22):11673–83.
132. Faiza M, Tanveer K, Fatihi S, Wang Y, Raza K. Comprehensive overview and assessment of miRNA target prediction tools in human and drosophila melanogaster. 2017;
133. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. Nat Commun. 2011;2:319.
134. Daniel C, Veno MT, Ekdahl Y, Kjems J, Öhman M. A distant cis acting intronic element induces site-selective RNA editing. Nucleic Acids Res. 2012;40(19):9876–9886.

135. Meisner NC, Hackermüller J, Uhl V, Aszódi A, Jaritz M, Auer M. mRNA openers and closers: Modulating AU-rich element-controlled mRNA stability by a molecular switch in mRNA secondary structure. ChemBioChem. 2004;5(10):1432–47.
136. López De Silanes I, Fan J, Yang X, Zonderman AB, Potapova O, Pizer ES, et al. Role of the RNA-binding protein HuR in colon carcinogenesis. Oncogene. 2003;22(46):7146–54.
137. Shi S, Chen S, Wang M, Guo B, He Y, Chen H. Clinical relevance of RNA editing profiles in lung adenocarcinoma. Front Genet. 2023;14(March):1–12.
138. Amin EM, Liu Y, Deng S, Tan KS, Chudgar N, Mayo MW, et al. The RNA-editing enzyme ADAR promotes lung adenocarcinoma migration and invasion by stabilizing FAK. Sci Signal. 2017 Sep;10(497).
139. Wang C, Huang M, Chen C, Li Y, Qin N, Ma Z, et al. Identification of A-to-I RNA editing profiles and their clinical relevance in lung adenocarcinoma. Sci China Life Sci. 2022;65(1):19–32.
140. Liu L, Wang C, Li S, Qu Y, Xue P, Ma Z, et al. ERO1L Is a Novel and Potential Biomarker in Lung Adenocarcinoma and Shapes the Immune-Suppressive Tumor Microenvironment. Front Immunol. 2021;12(July):1–14.
141. Xie J, Liao G, Feng Z, Liu B, Li X, Qiu M. ERO1L promotes the proliferation and metastasis of lung adenocarcinoma via the Wnt2/β-catenin signaling pathway. Mol Carcinog. 2022 Oct;61(10):897–909.
142. Liu Z, Yu C, Chen Z, Zhao C, Ye L, Li C. PSMB2 knockdown suppressed proteasome activity and cell proliferation, promoted apoptosis, and blocked NRF1 activation in gastric cancer cells. Cytotechnology [Internet]. 2022;74(4):491–502. Available from: https://doi.org/10.1007/s10616-022-00538-y
143. Lin J, Xu Z, Xie J, Deng X, Jiang L, Chen H, et al. Oncogene APOL1 promotes proliferation and inhibits apoptosis via activating NOTCH1 signaling pathway in pancreatic cancer. Cell Death Dis [Internet]. 2021;12(8). Available from: http://dx.doi.org/10.1038/s41419-021-03985-1

8 CURRICULUM VITAE