ID N. AIDR01/16900

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

CONSIGLIO NAZIONALE DELLE RICERCHE

IBIOM-CNR

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVII Cycle

Machine-Learning Approaches For Deciphering The Human Editome

Academic Tutors

Prof. Graziano Pesole

Prof. Ernesto Picardi

Candidate

Pietro Luca Mazzacuva

January, 2025

To my family and friends.

Acknowledgements

I would like to thank the National Research Centers: "High Performance Computing, Big Data and Quantum Computing" (Project no. CN_00000013) and "Gene Therapy and Drugs based on RNA Technology" (Project no. CN_00000041); and Extended Partnerships: MNESYS (Project no. PE_0000006) and Age-It (Project no. PE_00000015). This work was also supported by Life Science Hub Regione Puglia (LSH-Puglia, T4-AN-01 H93C22000560003) and INNOVA - Italian network of excellence for advanced diagnosis (PNC-EJ-2022-23683266 PNC-HLS-DA) and by ELIXIR-IT through the empowering project ELIXIRNextGenIT (Grant Code IR0000010). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from dbGaP under the accession number phs000424.v7.p2. The generation of the HEK293T ADAR1 knockout/overexpressing cell lines and their RNA-seq analysis was supported by the HI-TRON Kick-Start Seed Funding Program 2021 awarded to RP.

Abstract

Epitranscriptomic modifications have emerged as a key mechanism in the dynamic regulation of gene expression, as well as transcript function, stability, and structure. This set of post-transcriptional biochemical modifications, defined as the epitranscriptome, is ubiquitous in all three domains of life and pervasive in human transcriptome. The interconversion of adenine to inosine, known as A-to-I RNA editing, carried out by the ADAR enzymes operating on dsRNAs, is the main mechanism by which the cell directly modifies the information carried by transcripts without any alteration in the DNA. In recent years, it has become clear that A-to-I RNA editing has a plethora of functional downstream effects, playing a crucial role in neuroreceptor functionality, alternative splicing, proteomic diversification, circRNA biogenesis, gene expression, and cytosolic innate immunity through the MDA5-MAVS axis; moreover, A-to-I RNA editing dysregulation has been associated with several human disorders. NGS technologies, as well as RNA-seq technologies, allow the transcriptome-wide detection of A-to-I RNA editing as A-to-G mismatches by aligning RNA-seq reads to a reference genome; nevertheless, accurate discrimination of genuine A-to-I RNA editing events from potential genomic sources of variation, sequencing errors, and artifacts remains challenging. Although several empirical methods have been developed to identify A-to-I RNA editing in RNA-seq data, they rely on cumbersome filtering steps and annotations from public resources. Recently, to overcome the limitations of empirical methods, machine learning and deep learning-based algorithms have been used to identify A-to-I RNA editing in RNA-seq data. However, these approaches are strongly dependent on both the specific set of input features provided and the size and quality of the training data. To exploit millions of known A-to-I events stored in the REDIportal database, collected through a rigorous computational protocol, we propose a novel TCN-based binary classifier capable of accurately detecting A-to-I RNA editing events in RNA-seq data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Human Transcriptome - An Historic Perspective

The set of all the RNA transcripts of a cell is called the transcriptome. Initially, it was believed to be composed of only three main classes of RNAs involved, in various ways, in protein biosynthesis, a process referred to as translation [[1]]. In particular, such classes are: 1) messenger RNAs (mRNAs), responsible for carrying the genetic information to ribosomes where the translation takes place; 2) ribosomal RNAs (rRNAs), which constitute the ribosomes nucleic acid component in charge of the catalytic activity for concatenating amino acids, and 3) the transfer RNAs (tRNAs), responsible for carrying the amino acids to the ribosomes. Over the years it has been established that only a small portion of the transcriptome (about 2% in H. sapiens) is translated into proteins [[2], [3]], while most of the transcripts, called non-coding RNAs (ncRNAs), perform a myriad of structural or regulatory functions [[4]. In detail, the discovery of small interfering RNA (siRNA) [[5]] and microRNA (miRNA) [[6]] has radically changed the understanding of the transcriptome. These classes of small ncRNAs, by binding to other RNA molecules to form regions of double-stranded RNA (dsRNA), usually downregulate gene expression through the biological process known as RNA interference. Subsequently, other classes of ncRNAs have been identified, for instance, Piwi-interacting RNAs (piRNAs) which form complexes with Piwi proteins and are involved in the silencing of retrotransposons in germline cells, transcription initiation RNAs (tiRNAs), which play a key role in the process of cellular differentiation in embryonic development, [[7], [8]] long non-coding RNAs (lncRNAs), which are ncRNAs longer than 200 base pairs, for many of which, although there is still no precise picture of their functions, the evidence collected so far indicates that they are critical in epigenetic regulation pathways, signaling protein complexes scaffolding and gene expression in mammals [[9]]. To date,

many other classes of ncRNAs have been identified (e.g. circular RNAs) and are still under study. In multicellular eukaryotes, transcriptomes are further heavily diversified through the biological process called alternative splicing [[10]], by which a single genomic locus can originate multiple different transcripts and originate multiple different tissue- and cell-specific transcripts, also in response to specific conditions [[11], [12]]. Alternative splicing is undoubtedly a key component in the functional diversification of transcriptomes; indeed, in multicellular eukaryotes, it affects over 90% of the multi-exon genes [[13], [14]]. It should also be noted that alternative splicing does not only affect proteins coding genes but also involves approximately 30% of ncRNAs genes [[15]]; in particular lncRNAs genes have on average 2.3 expressed isoforms [[16]]. The composition of the transcriptome does not remain constant over time. Empirical evidence shows that on average only 5% - 10% of the genome is expressed constitutively [[17], [18], [19]]. However, along the various phases of the cell cycle or cell differentiation, almost all of the genome is transcribed. In humans, for example, data indicate that approximately 93% of the genome is transcribed in different tissues at different times [[20]]. Historically the pervasive transcription in humans, i.e. the observation that the vast majority of the human genome is transcribed, has been a major subject of scientific debate in the biological sciences since the early 2000s. The beginning of this debate can be traced back to 2002 when a study of the human transcriptome was conducted by targeting the non-repeated portions of human chromosomes 21 and 22 [[21]] using tiling array technology in several tumor cell lines; a type of microarray in which, like traditional microarrays, labeled DNA or RNA target molecules are hybridized to probes fixed on a solid surface, but instead of probing known genes that may be dispersed throughout the genome, as in traditional microarrays, they probed sequences known to reside in contiguous regions at high density. This study showed that the amount of genomic sequences transcribed was an order of magnitude higher than the exonic sequences. This first result was later confirmed by further studies conducted also using tiling arrays on the entire human genome [[17], [18], [22], [23]], which led in 2004 to the birth of the Encyclopedia Of DNA Element (ENCODE) project [[24]], whose aim was to identify and catalog the functional elements present in the human genome; where a functional element is any segment of the genome that codes or modulates the expression of a defined gene product, be it a polypeptide or a ncRNA. However, considering the limitations of microarrays, such as background noise generated by cross-hybridizations, signal saturation, and complicated signal normalization methods to compare between different experiments, the idea that a considerable portion of the detected ncRNAs could be the effect of transcription errors and artifacts persisted [[25]] up until the advent of RNA sequencing (RNA-seq) technologies as well as second-generation sequencing technologies, also called next generation sequencing (NGS), made it possible to

obtain sequencing data of the entire transcriptome, confirming the pervasive transcription of the human genome [[26]]. At present, the annotation of the functional elements of the human transcriptome corresponding to its lncRNA component, has been made easier by the advent of third-generation sequencing technologies that allow the sequencing of RNA molecules long up to tens of thousands of nucleotides [[27]].

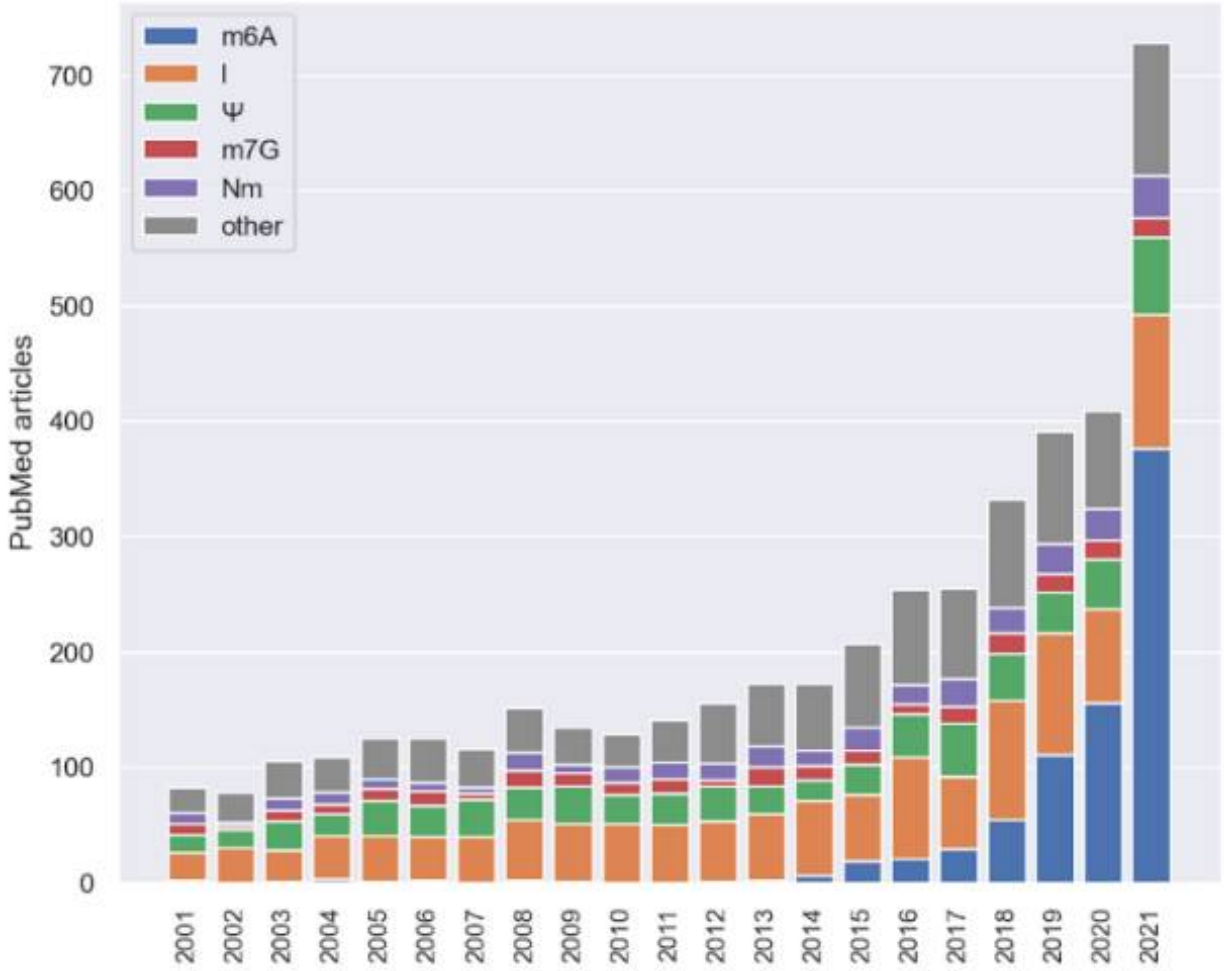

Figure 1.1: Number of scientific publications in epitranscriptomics research per year, in the PubMed database. Taken and modified from Arzumanian VA, Dolgalev GV, Kurbatov IY, Kiseleva OI, Poverennaya EV. Int. J. Mol. Sci. 2022, vol. 23, issue 13851.

The problem of understanding the human transcriptome, however, does not end with the identification of all the the expression product that compose it, since downstream of transcription lies the galaxy of functionally relevant post-transcriptional biochemical modifications of RNA molecules that are collectively defined as epitranscriptome [[28]]. Historically, the first RNA modification to be identified occurred with the isolation of pseudouridine in the 50s [[29]] up to the present day in which more than 170 modifications have been characterized

along all the known RNAs classes [[30]]. This was made possible thanks to the use of NGS technologies. The spectrum of these modifications ranges from simple methylation, as in the case of N6-Methyladenosine (m6A), to the complex synthesis of 5-methoxycarbonylmethyl-2-thiouridine, which is generated in a sequence of reactions catalyzed by different enzymes [[31]]. Empirical evidence has shown that the epitranscriptome is essential for the correct functioning of the various classes of RNAs. It is crucial for transcription [[32]], translation [[33]], transport [[34]], alternative splicing [[35]], the formation of secondary structures [[36]], and the immune response [[37]]. Some of these modifications always occur in precise positions of the RNA molecules, while others are strongly influenced by the precise cellular or environmental context and therefore present a certain dynamism [[31]]. The importance of the epitranscriptome is highlighted by the correlation between its malfunction and several diseases [[38], [39]], in particular several forms of cancer such as hepatocellular carcinoma [[40]] and Hodgkin's lymphoma [[41]]. In this regard, it is important to highlight how the effectiveness of mRNA vaccines produced against SARS-CoV2 is largely due to the use, in their preparation, of the N1-methyl pseudouridine modification which increased their translational efficiency [[42]]. Given the importance of understanding the epitranscriptome, the number of scientific publications on it has grown rapidly in recent years (Figure 1.1); although only a fraction of all RNA modifications have been thoroughly explored. In this regard, the most studied modified nucleotides are listed below:

- N6-methyladenosine:

  N6-methyladenosine (m6a) is produced by reversible methylation of the amino group on carbon six of adenosine. This methylation is catalyzed by multiprotein complexes with methyltransferase activity, called m6a writers, which mainly recognize the motif GGACU; for example, in humans, the METTL3, METTL14, and WTP complex are well characterized [[43], [44]]. The m6A modification is widely present in all eukaryotes and is essential for the stability of transcripts, their translocation, their translation, and alternative splicing [[45], [46]]. In eukaryotes, it is mainly present in the 5' Untranslated Region (5'UTR) and in the 3' Untranslated Region (3'UTR) of mRNAs, and is found also in tRNA, rRNA, circRNA, miRNA, snRNA, and lncRNA [[43]]. m6a is recognized downstream by a series of RNA binding proteins, called m6a readers, which have specific binding sites for m6a recognition, for example in humans the protein YTHDC1 is widely studied [[47]]. In humans, m6a is involved in several human diseases, for example, obesity, heart disease, and tumors [[47]]; therefore the regulation of m6A is the target of therapeutic strategies under development [[48]].

- Pseudouridine:

  Pseudouridine (ψ), as mentioned previously, was the first modified base discovered and is the most abundant epitranscriptomics modification in the three domains of life [[49]], and is found in a wide range of RNA classes [[49]]. It is an isomer of uridine in which carbon 5 of uracil is linked to ribose by a carbon-carbon glycosyl bond, instead of the nitrogen-carbon glycosyl bond involving nitrogen 1 of uracil. In tRNAs, ψ is usually found at nucleotide position 55 and has been associated with increased stability of tRNAs in response to cellular stress [[50], [51]]; thus, it is crucial for the correct course of protein synthesis. The interconversion of uridine to ψ is catalyzed by a family of enzymes known as pseudouridine synthases. In humans, ψ is synthesized by a multiprotein complex composed of the enzymes RPUSD1,2,3,4, PUS10, PUS7L, and TRUB2 [[28]]. Unlike the case of m6a, pseudouridylation is an irreversible reaction since ψ is a thermodynamically much more stable isomer of uracil, given the higher activation energy for the cleavage of the carbon-carbon bond compared to the nitrogen-carbon bond [[28]]. Pseudouridylation has also been linked to human diseases, including mitochondrial dysfunction, anemia, dyskeratosis, and numerous tumors [[52], [53], [54], [55]].

- 7-Methylguanosine:

  7-Methylguanosine (m7G) is a common RNA modification in eukaryotes [[56]] where it forms part of the 5' cap, which is an important structure required for correct translation and to prevent degradation of the mRNAs by the innate immune system [[57]]. It consists of the methylation of the nitrogen in position 7 of guanine, by methyltransferases specific for guanosine, which in humans it is catalyzed by the enzyme METTL1. It has been found in rRNAs, tRNAs, mRNAs, and miRNAs [[56], [58]]in particular the methylation in position 46 in the variable region of tRNA is almost ubiquitous in all organisms [[56]]. Recently it has been shown that m7G can be present internally in trascripts in different condition-specific locations [[59]], proving m7g to be a more dynamic transcriptomic modification than previously hypothesized, and closely related to the regulation of translation [[59]]; in particular, although the issue is still a matter of debate, it appears to be essential to the biogenesis of the miRNA let-7[[28]].

- 2'-O-methyl ribose:

  2'-O-methyl ribose (2'-O-Me) is formed by methylation of ribose at the hydroxyl group of carbon 2', properly called 2'-hydroxyl. This methylation occurs in a nonspecific manner and therefore can occur in conjunction with any nitrogenous base. This epitranscriptomic modification, very common in eukaryotes and also present in the other two domains of life [[28], [60]], in humans is catalyzed by the methyltransferase fibrillarin (FBL) [[61]] which, by associating with the ncRNAs U3, U8, and U13, forms active riboprotein complexes called small nuclear ribonucleoproteins (snRNPs) which are located in the dense fibrillar component (DFC) of the nucleolus. 2'-O-Me modifications, such as m6A, are present in multiple classes of RNAs, such as snRNA, snoRNA, lncRNA, rRNA, tRNA, and mRNA [[62]]. In eukaryotes, 2'-O-methylation is frequently found on nucleotides in position 7 of tRNAs. In mRNAs, it is mainly found in the 5'UTR sequences and the 3'UTR 3' sequences, but it is also present in CDSs at the level of the AGUA motifs [[63]]. 2'O-methylation affects the structure and stability of RNAs, as well as their interactions with other cellular components [[64]]. For example, it has been shown that this modification prevents the formation of tertiary structures of RNA [[64]], with an effect on the epigenetic regulation of cells. 2'-O-Me modifications have been linked to several human diseases, such as asthma, Alzheimer's disease, Prader-Willi syndrome, and breast cancer [[[64], [65]].

- Inosine:

  Inosine (I) is the second most abundant modified base in the human transcriptome after m6A and is present in all three domains of life [[66]]. Inosine is produced by hydrolytic deamination of the amino group on the carbon in position 6 of adenine (Figure 1.2). The production of I from adenosine is called A-to-I RNA editing and represents the main process by which the cell modifies the information contained within RNAs, modifying their codons, without any alteration of the genetic information contained in the DNA. Indeed, I is recognized downstream by the cellular machinery as a guanosine, functionally resulting in a change from adenine to guanine in the nucleotide sequence during base pairing [[67]]. Inosine is synthesized by a family of proteins called adenosine deaminases acting on RNAs (ADARs) [[68], [69]], which have as substrate molecules of double-stranded RNA (dsRNAs) and are conserved in all animals [[70]]. Inosine, having a hybridization pattern different from adenosine, affects both the local structure of RNAs and their interactions with other RNAs or with RNA-binding proteins (RBD) [[71]]. In eukaryotes, several tRNAs have an I in the nucleotide at position 34 that should support the recognition of the wobble codon [[72]].

Of great interest are the changes from A to I that occur in mRNAs, indeed they influence the expression by modifying the stability of any secondary structure and intramolecular folds [[73], [74]]. If they occur on the CDS portions, they may modify the primary structure of the corresponding proteins [[71]]. In this regard, it is worth mentioning the well-known case of the cation-specific ion channel called Glutamate ionotropic receptor AMPA type subunit 2 (GluR-2), which represents one of the components of AMPA receptors, in which the change from A to I determines the replacement from a glutamine (Q) to an arginine (R), at the level of the codon encoding amino acid number 607 called Q/R site, of the corresponding gene (GRIA2). RNA editing at the Q/R site determines a drastic change in the functionality of GluR-2 which, following the transition from Q to R, becomes impermeable to the Ca2+ ions [[75]], and studies in mice have shown that this change in permeability is essential for the correct development of the brain [[76]].

The phenomenon of A-to-I RNA editing, catalyzed by the ADAR family of enzymes, is the biological subject of this thesis and will be discussed in the following paragraph.

1.2 A-to-I RNA editing - A General Overview

As previously mentioned, A-to-I RNA editing is the second most abundant epitranscriptomic modification. It alters the sequence of the transcripts and can be detected in sequencing data as a nucleotide substitution between the RNA molecule and the corresponding genomic locus of origin [[77]].

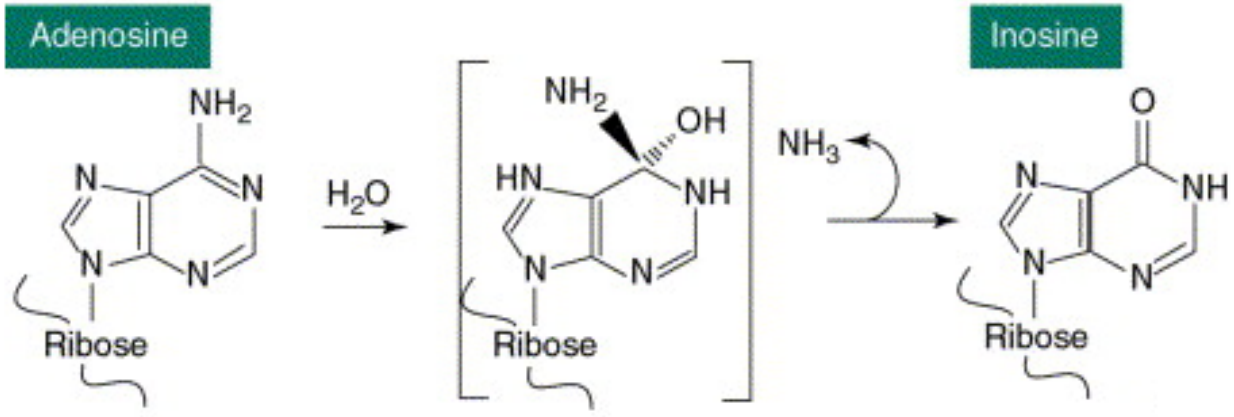

Figure 1.2: A-to-I RNA editing schematic reaction diagrams. Taken and modified from Gerber AP, Keller W. TIBS. vol 26, issue 6, pp. 376-384.

In humans, A-to-I RNA editing is carried out by three different members of the ADAR family [[68], [69]], namely the double-stranded RNA-specific adenosine deaminase (ADAR1), the double-stranded RNA-specific editase 1 (ADAR2, also known as ADARB1) and ADAR3 (also known as ADARB2) (Figure 1.3). ADAR1 is expressed in 2 isoforms (Figure 1.3), ADAR1p150 (1226 amino acids long and having two Z-DNA binding domains, three dsRNA binding domains, and a deaminase domain) and ADAR1p110 (931 amino acids long and having a Z-DNA binding domain, three dsRNA binding domains and a deaminase domain). ADAR2 is expressed in 3 isoforms (Figure 1.3), ADAR2a (701 amino acids long and having two dsRNA binding domains and a deaminase domain), ADAR2b (741 amino acids long and having two dsRNA binding domains and a deaminase domain that has 40 amino acid Alu insert in the C-terminal half of its sequence), and ADAR2R (674 amino acids long and having an R domain, two dsRNA binding domains and a deaminase domain). ADAR3 (Figure 1.3), instead, is expressed in a single isoform (793 amino acids long and has an R-domain, two dsRNA binding domains, and a deaminase domain). ADAR1 and ADAR2 are ubiquitously expressed in human tissues. ADAR1 is the most highly expressed and is responsible for most of the editing activity [[78]]. ADAR3, like ADAR2, is mainly expressed in the brain and is catalytically inactive [[78]].

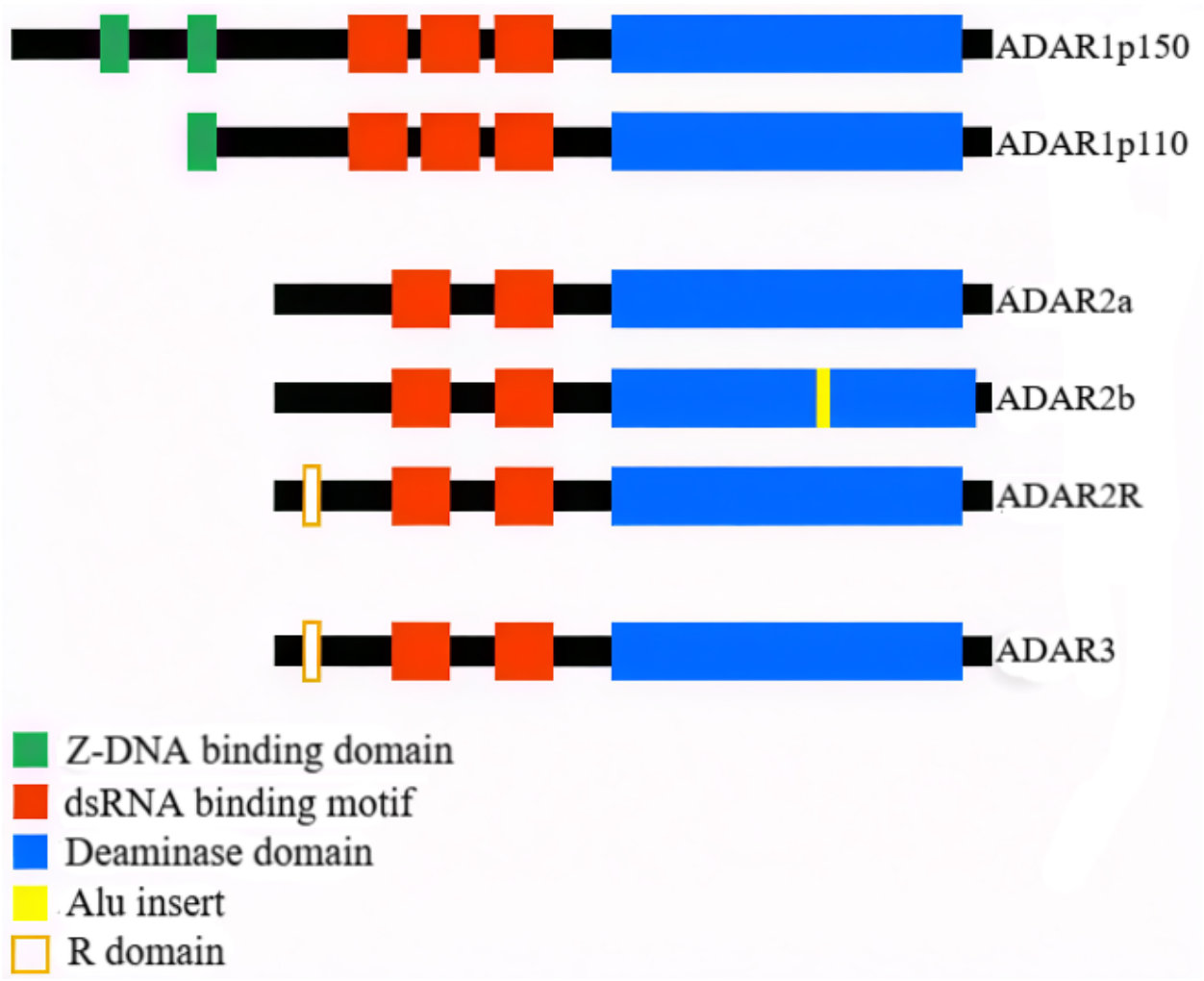

Figure 1.3: Schematic structure of H. sapiens ADAR enzymes (hADARs). Taken and modified from Cattenoz, Pierre B (2012). Characterization of Alu element expression and A-to-I RNA editing in mammals [Doctoral dissertation, The University of Queensland]. UQ eSpace.

All isoforms of the 3 enzymes recognize exclusively dsRNAs and edit at specific sequence motifs (Figure 1.4). The identification of inosine in the sequencing data is based on its guanosine-like behavior during base pairing. As a consequence, inosine is interpreted as guanosine by the cellular machinery [[67]], and A-to-I RNA editing changes appear as A-to-G substitutions (AG substitutions) [[73], [79], [80], [81]]. Inosines falling in CDSs can cause non-synonymous substitutions giving rise to new protein isoforms with potentially different functionalities. These edited sites are referred to as recoding [[77]] sites (Figure 1.6a). Inosines residing in ncRNAs or the UTRs of mRNAs, instead, can play multiple structural and regulatory roles [[77]]. The first empirical evidence for A-to-I RNA editing in humans came with the discovery of a handful of recoding sites [[82], [83]]; the aforementioned Q/R site of the GluR-2 receptor, which determines its impermeability to the Ca2+ ions (Figure 1.5),

and the I/V and N/S sites in the serotonin 5-hydroxytryptamine2c receptor (5-HT2cR), which reduces its affinity for the G31 protein [[84]]. Recoding sites were the first to be discovered due to their characteristic of being edited at high frequency, which makes them more easily identifiable events at low sequencing depth.

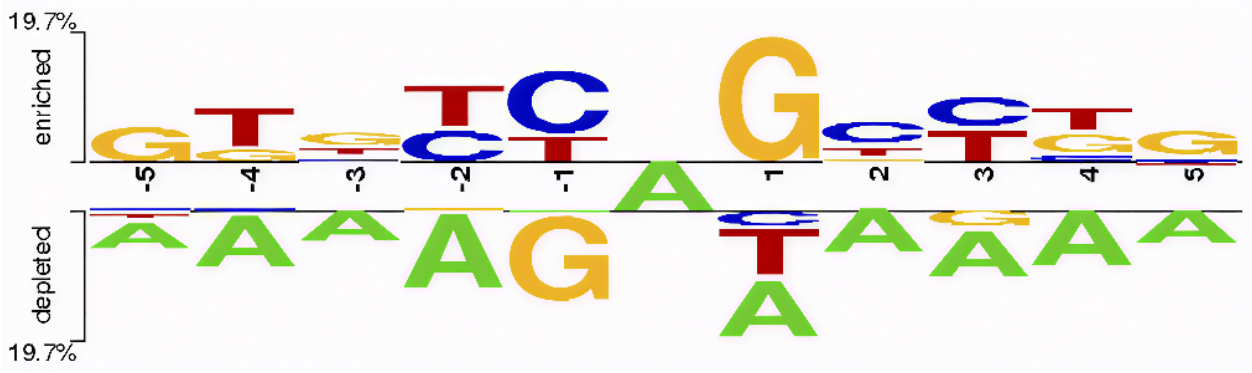

Figure 1.4: hADARs sequence preferences for base positions flanking (−5, +5) detected A-to-I editing sitesTaken and modified from Picardi E, Manzari C, Mastropasqua F, Aiello I, D'Erchia AM, Pesole G. Sci. Rep. 2015, vol 5.

The increase of human transcriptome data, coupled with DNA sequencing data, have subsequently led to the identification of more than a thousand of these recoding sites in humans [[85], [86]], which remain of great interest in the biological sciences, giving their functional downstream effects. Editing in recoding sites is mainly located in the nervous system tissues [[85], [86]], and in addition to the sites already mentioned in GluR-2 and 5-HT2c, it is important to mention the recoding sites in the voltage-gated potassium channels Shaker, in which A-to-I RNA editing influences their activation kinetics [[83]]. Although experimental evidence has highlighted a multiplicity of recoding sites edited by ADAR2 in the nervous system tissues [[87]], gene knockout studies in mice suggest that the only fundamental target of ADAR2 is the Q/R site of GluR-2 [[88]]. Indeed, mice lacking the Adar2 gene (the gene encoding ADAR2) suffer from severe epileptic seizures and die within 3 weeks from birth, while mice lacking the Adar2 gene with the Q607R mutation in the GRIA2 gene do not show this phenotype [[88], [89]]. Important editing recoding sites have also been observed in tissues outside the nervous system, in proteins ubiquitously expressed having functions that are not strictly related to the functioning of the nervous system, such as filamin (FLA), antizyme inhibitor 1 (AZIN1), and endonuclease 8-like 1 (NEIL1). FLA is an actin crosslinker expressed in most human tissues, whose transcript is edited mainly by ADAR2 [[90], [91]]. The product of the ADAR1-edited AZIN1 transcript is able to translocate from the cytoplasm to the nucleus and increases its affinity for the antizyme by inhibiting the degradation of ornithine decarboxylase (ODC), thus promoting cell proliferation [[92], [93]. The NEIL1

ADAR1-edited product causes a 30-fold reduction in the rate of removal of thymine glycol from duplex DNA [[92], [94]].

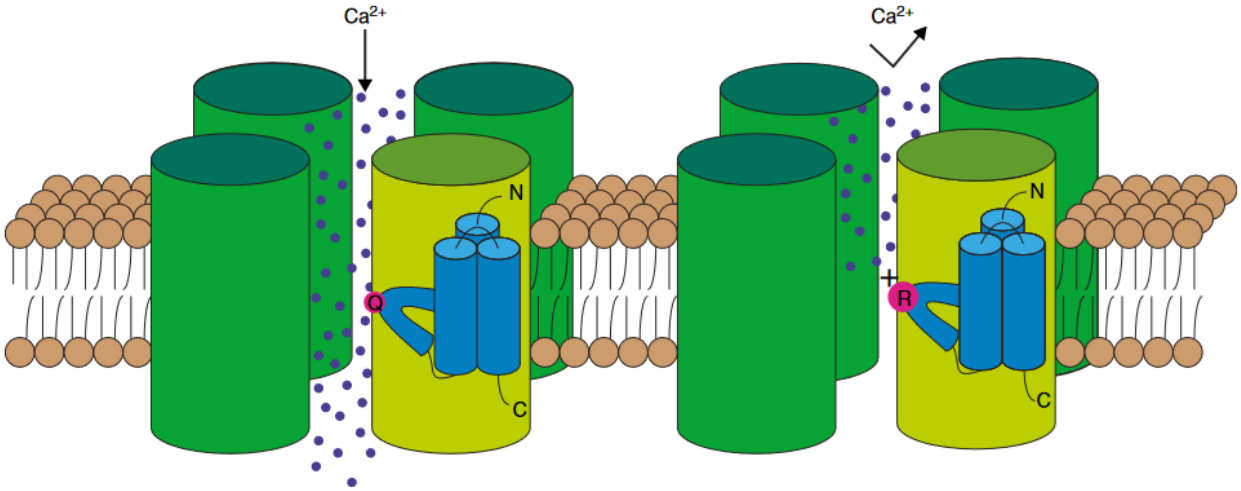

Figure 1.5: A-to-I functional effect on the GluR-2 Q/R recoding site. Taken and modified from Slotkin W, Nishikura K. Genome Med. 2013, vol 5.

To date, only a small fraction of the recoding sites identified in humans have been functionally characterized [[77]]. Interestingly, only a few tens of them are conserved in mammals [[95]]. Although the true large-scale impact of A-to-I RNA editing at recoding sites is still largely unknown, the effects of dysregulation of ADAR activity at some recoding sites in humans have been extensively characterized over the years [[76]]. In particular, anomalous decreases in ADAR2 enzymatic activity in the nervous system, corresponding to decreasing levels of editing at the Q/R site of GluR-2, induce increases in Ca2+ ion flux that have been correlated with several neurological and neurodegenerative diseases as well as with the progression of some forms of cancer [[76]]. In particular, using murine models, the correlation between the decrease in ADAR2 enzymatic activity with the onset of both amyotrophic lateral sclerosis (ALS) [[96]] and transient forebrain ischemia [[97]] has been demonstrated. In the first case, the mechanism was proposed by which the increase in cellular Ca2+ concentration induces excitotoxicity in motor neurons [[98], [99]], while in the second case, it was proposed that the increase in cellular Ca2+ ion concentration induces phosphorylation of N-methyl-D-aspartate (NMDA) receptors by the phosphorylase Cdk5, which leads to cell death of CA1 pyramidal neurons [[97], [100]]. The role of ADAR2 as a tumor suppressor of the brain tumor Glioblastoma (GLB) has also been demonstrated; in fact, ADAR2 overexpression in immortalized GLB cells (U118, U172, U87) determines a significant reduction in cell proliferation and migration [[101]], caused by a consistent decrease in the levels of Ca2+-dependent phosphorylation of protein kinase B (PKB, also called

Akt), whose activation by phosphorylation induces cell proliferation. Interestingly, in GLB, the decrease in ADAR2 enzymatic activity is associated with ADAR1 overexpression [[101]]; for which the hypothesis that ADAR1 sequesters ADAR2 forming inactive dimers has been proposed [[101]]. On the contrary, the correlation between the abnormal increase in ADAR1 and ADAR2 enzymatic activity with both hyperphagia and growth retardation has been demonstrated in mouse models [[102]], due to the increase in the levels of editing of the five recoding sites of 5-HT2cR (sites A and B target of ADAR1, site D target of ADAR2, sites C and E target of both ADAR1 and ADAR2) [[103]], which leads to a drastic decrease in serotonin efficacy. An abnormal increase in ADAR1 activity has been also correlated to the progression of human hepatocellular carcinoma (HCC) [[93], which is the fifth most common type of cancer. In particular, the excessive ADAR1-mediated editing of the AZIN1 Y/S recoding site (amino acid in position 367) in HCC cells promotes cell proliferation through potent inhibition of antizyme [[93]. Although still a matter of debate, some studies suggest that the increase in editing levels in the five recoding sites of 5-HT2cR may also play a role in the onset of depression, schizophrenia, and bipolar disorder [[104]]. Although A-to-I RNA editing at recoding sites has an undoubted biological relevance, the vast majority of edited sites fall within non-coding sequences and are normally edited at a low frequency. As previously mentioned, A-to-I RNA editing in non-coding regions triggers a plethora of important downstream effects. In particular, A-to-I RNA editing at the level of intronic sequences alters the alternative splicing process of pre-mRNAs, inducing the production of particular protein isoforms [[105]] rather than others (Figure 1.6b). Inosines in the 3'UTR regions very often modify the binding sites for miRNAs (Figure 1.6c), similarly, inosines within the seed regions of miRNAs modify the set of target mRNAs recognized by them (Figure 1.6d) [[106], [107]]; in this regard, recent studies have demonstrated how miRNAs are extensively edited in humans [[106]. Empirical evidence collected in the last few years suggests that A-to-I RNA editing plays a role in the biosynthesis of circRNAs [[108], a class of particularly stable RNAs recently discovered, whose biological functions are still largely unknown, in which inosines in the intronic regions flanking circRNAs prevent their pairing, blocking their formation and expression (Figure 1.6e). It has also been recently observed that A-to-I RNA editing of endogenous dsRNAs by ADAR1 inhibits the activation of the cytosolic response of the innate immune system [[109], [110], [111]], by disrupting endogenous dsRNAs base pairing. Although endogenous dsRNAs are produced by the cell, they are extremely similar to viral genomes [[112]] that activate the innate immune response through the activation of melanoma differentiation-associated protein 5 (MDA5), an RBP that binds dsRNAs hundreds of bases long and acts as a detector [[113]].

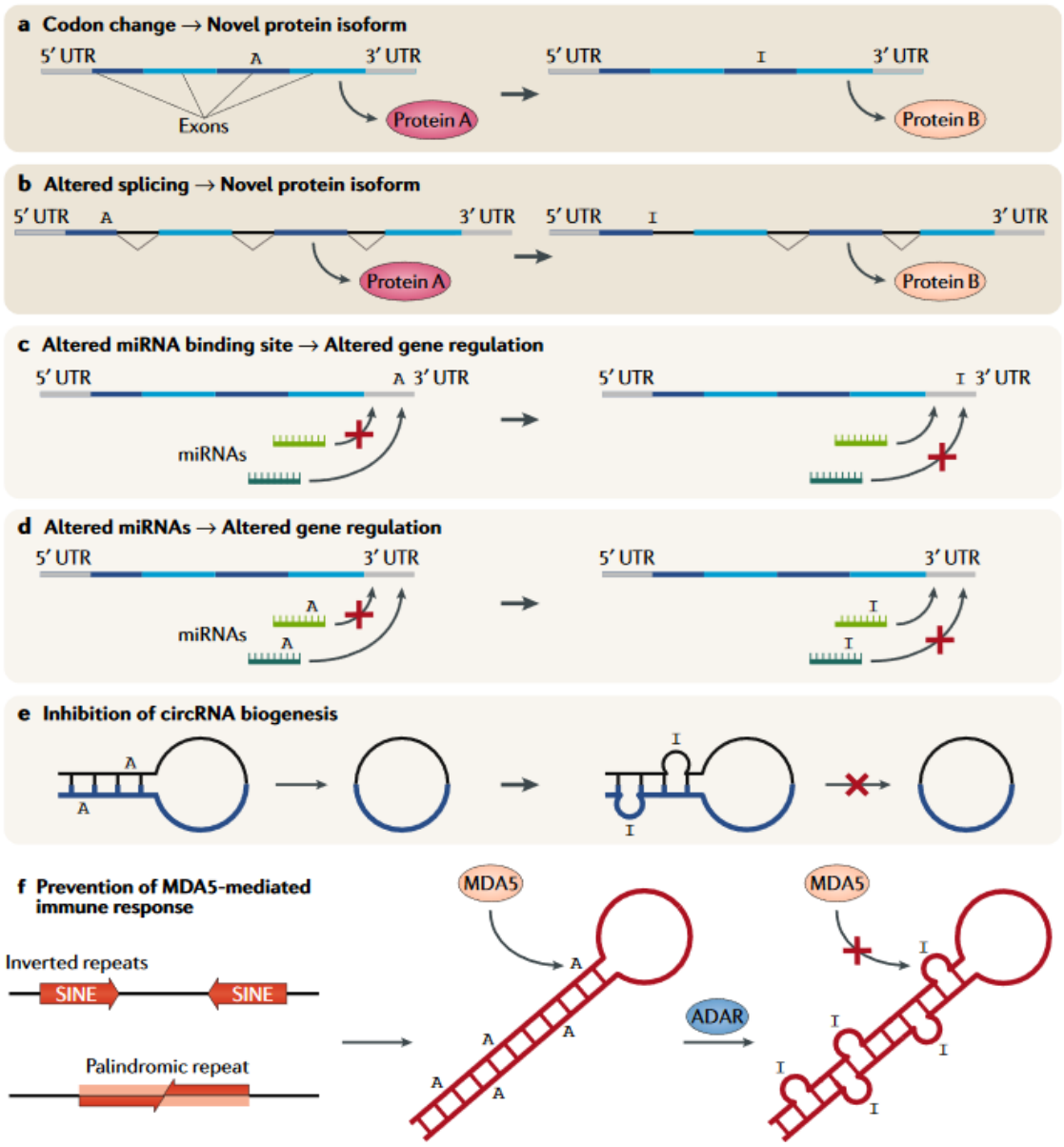

Figure 1.6: Characterized A-to-I RNA editing functional effects in human cells. (a) Amino acid substitution by A-to-i RNA editing on Recoding sites with subsequent production of novel protein isoforms. (b) Disruption of alternative splicing by A-to-i RNA editing on splice junctions, with subsequent production of novel protein isoforms. (c) Generation or disruption of miRNAs recognition sites by A-to-I RNA editing on target UTR regions.(d) Modification of miRNAs sets of target sequences by A-to-I RNA editing on the miRNAs themselves.(e) Disruption of circRNAs biogenesis by A-to-I RNA editing on the parent transcripts. (f) Prevention of MAVS-MDA5-mediated innate immune response activation by A-to-I RNA editing on endogenous dsRNAs. Taken and modified from Eisenberg E, Levanon EY. Nat. Rev. Genet. 2018, vol 18, pp. 473–490.

The binding of MDA5 to dsRNAs induces a conformational change that enables it to bind the mitochondrial antiviral signaling protein (MAVS) generating an active multiprotein complex MDA5-MAVS. The MDA5-MAVS complex increases the expression of type I interferon (IFN I) and inflammatory cytokines that induce the antiviral response. In human cells, the presence of dsRNA also determines the hyperactivation of RNA-activated kinase (PKR) that promotes translation arrest and cell death by apoptosis.RNA editing of endogenous dsRNAs mediated by ADAR1 has the effect of breaking their pairing such that they are no longer recognized by MDA5, preventing the unwanted activation of the innate immune system (Figure 1.6f). It has been experimentally demonstrated in murine cells that the lack of ADAR1 results in a lethal phenotype for the cells while the double ADAR1-MDA5 knockout results in a vital phenotype; this has also led to the hypothesis that the main role played by ADAR1 is to edit endogenous dsRNAs [[114], [115], [116]]. Over the years, the amount of evidence collected, however, clearly indicates that in humans the preferred targets of ADARs are dsRNAs, formed by inverted copies of retroelements in the introns or the UTRs [[77]]. In particular, 99% of the non-recoding sites in humans are found in the Alu repeat sequences [[117]], which constitute approximately 10% of the human genome. Alu sequences are mainly concentrated within gene-rich genomic regions; in particular, they are mostly present in the intronic portions of genes. Pairs of inverted copies of Alu sequences present within the same primary transcript can anneal forming double-stranded regions that become targets of ADARs [[77]]. Considering the abundance of non-recoding sites in Alu sequences, in which virtually all the adenines present are interconverted into inosines, editing activity on Alu sequences is commonly used to quantify transcriptome-wide ADARs activity, through the metric known as Alu Editing Index (AEI) [[118]]. In particular, AEI is calculated by the following equation:

$$AEI = 100 * \left(\frac{AG}{AA+AG}\right) \tag{1.1}$$

where AG is the total number of A-to-G mismatches observed in Alu sequences and AA is the number of adenines matching the corresponding adenines in the reference genome, observed in the Alu sequences. As for recoding sites, the dysregulation of editing levels across non-recoding sites has been associated with various autoimmune and metabolic diseases as well as some forms of cancer [[76]]. In particular, in patients affected by the autoimmune disease known as Aicardi-Goutières syndrome (AGS), have been identified nine ADAR1 mutations that are correlated with high levels of interferon-mediated immune response due to a lack of suppression of the MDA5/MAVS signaling pathway [[116]]. Furthermore, ADAR1 overex-

pression has been identified in patients affected by lupus erythematosus [[119]]; although the involvement of ADAR1 in this pathology has not yet been fully elucidated. Furthermore, it has been demonstrated in mice the correlation between the dysregulation in liver cells of the uric acid biosynthesis pathway and the reduction of the catalytic activity of ADAR2 at the non-recoding site of the miRNA miR-376a [[107]]. It is important to clarify that miR-376a is involved in the expression of genes involved in many metabolic pathways other than uric acid biosynthesis, therefore the implications of reduced ADAR2 activity on liver functionality have not yet been fully clarified [[76]]. In murine models, it has also been demonstrated that melanoma progression is related to the downregulation of ADAR1 expression [[120]]; in fact, it has been observed that ADAR1 behaves as a tumor suppressor in epidermal cells [[120]], promoting the RNA interference process by upregulating both the expression of the Dicer endonuclease [[120]], through edited let-7 miRNA, and Dicer enzymatic activity on miRNAs forming heterodimers with it [[120]]. While advances in sequencing technologies have shed light on the actual extent of A-to-I RNA editing in the human epitranscriptome, science is still far from the full understanding of this important biological process. In fact, it is still difficult to identify A-to-I RNA editing events as they are largely tissue-specific or condition-specific. The difficulty in identifying A-to-I RNA editing events is accompanied by the difficulties in characterizing their downstream functional effects; it is no coincidence that in humans the biological implications of A-to-I RNA editing at the level of the majority of recoding sites are still unknown [[77]], and the same is true for the majority of non-recoding sites, which moreover exceed the former by several orders of magnitude. However, considering that the involvement of A-to-I RNA editing in a multitude of cellular processes in humans has been widely demonstrated, the study of A-to-I RNA editing has assumed a leading role in the biological sciences research landscape. In this regard, it is interesting to report that in recent years ADARs enzymes have been successfully used in vivo to correct certain genetic mutations by editing specific target adenines using artificial oligoribonucleotides as a guide [[121], [122]], even though this technology is still characterized by a high degree of off-target editing [[123]]. In conclusion, it can be reasonably stated that research on the regulatory mechanisms of ADARs, on the molecular basis of their selectivity and the downstream effects of A-to-I RNA editing, represents a considerable prospect in the development of precision medicine.

1.3 RNA-seq - A Focus On Illumina Sequencing Technology

Following Sanger's discovery of the first DNA sequencing method in the late 1970s, there has been a progressive evolution in nucleic acid sequencing technologies, which have acquired ever-increasing productivity and efficiency. The first real revolution in the field of nucleic acid sequencing, after the Sanger method, occurred in the early 2000s with the advent of NGS technologies, which, thanks to complex automation systems and the ability to perform many sequencing reactions in parallel, have exponentially increased the sequencers' throughput. It is important to highlight how all commercially available NGS technologies require common preliminary steps for the appropriate preparation of samples, consisting of the extraction and purification of nucleic acids, the fragmentation of the extracted nucleic acids, the selection of fragments according to the adopted sequencing technology, the ligation of two different adapters at both ends (5' and 3') of the fragments, and the amplification of the fragments by PCR. If primers complementary only to one of the two adapters are used during the sequencing, the procedure is called single-end sequencing. Single-end sequencing is characterized by a loss of accuracy at the 3' ends due to the polymerase loss of processivity. Today, to solve this problem, the sequencing procedure relies on two primers, each one complementary to one of the two adapters, so that the fragments are sequenced one time from one end to the other and the second time in the opposite direction; thus, this technique is called paired-end sequencing. To date, there are four main NGS technologies that differ from each other in terms of maximum fragment length, specific sequencing method adopted, and sequencing accuracy; these technologies are the Roche 454 technology, the ThermoFisher SOLiD technology, the ThermoFisher Ion Torrent/Proton technology, and the Illumina technology. Although Roche 454 and ThermoFisher platforms have been discarded, Illumina technology is currently the most widely used sequencing technology, because it has the lowest unit cost and provides the highest output volume in terms of reads produced. Illumina sequencing (Figure 1.7) is performed on a glass support, the flow cell, on which are fixed two types of oligonucleotides of known sequence, used as primers for a series of amplification reactions, called bridge amplification, that generates clusters of clones for each fragment, and represents the first of two processes on which Illumina sequencing is based. In Illumina sequencing, the adapters mentioned above, to which the fragments to be sequenced are ligated, are synthesized so that they are complementary to the oligonucleotides present on the flow cell.

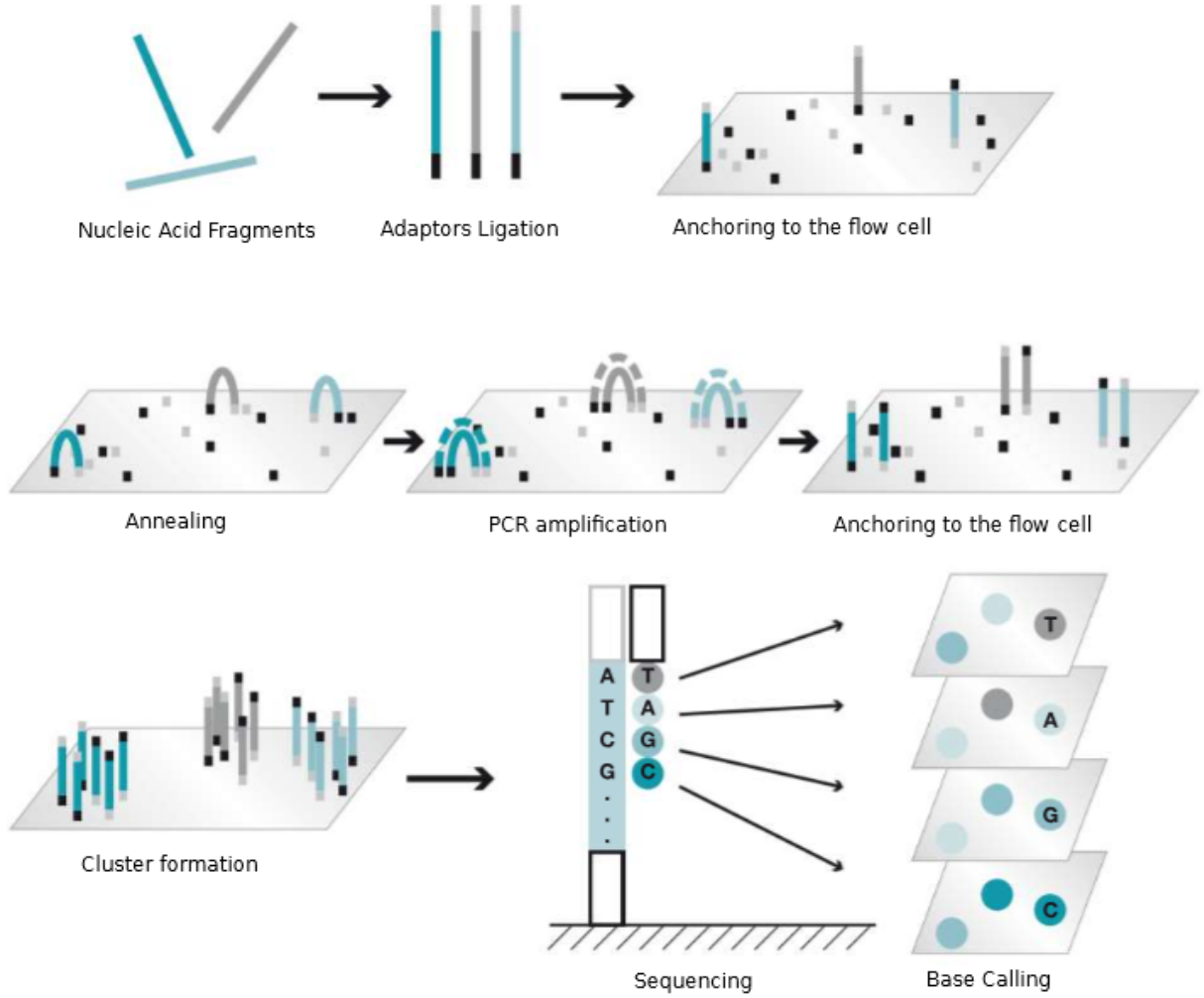

Figure 1.7: Illumina sequencing schematics. Taken and modified from Helmer Citterich M, Ferrè F, Pavesi G, Pesole G, Romualdi C. "Piattaforme di sequenziamento degli acidi nucleici." Fondamenti di bioinformatica, edited by Epitesto Srl, Zanichelli editore S.p.A, 2018, pp. 95-112.

The bridge amplification begins with the denaturation of the fragments to be sequenced to allow base pairing between the adapters and the oligonucleotides present on the flow cell, followed by the elongation of the oligonucleotides by the polymerase, using the fragments as guide strands for synthesis. Once the first amplification is complete, a second denaturation phase occurs with the washing out of the fragments not anchored to the flow cell. From this point on, only the fragments anchored to the flow cell, obtained by elongating one of the two oligonucleotides, will be amplified, by annealing (base pairing between two nucleotides sequences) their free end to an oligonucleotide, different from the one that acted as a primer for the first amplification, generating structures that resemble actual bridges between the

oligonucleotides anchored to the flow cell; hence the name bridge amplification. The bridge amplification reaction then continues with alternating phases of denaturation, annealing, and amplification for a certain number of cycles. The second process of Illumina sequencing consists of the actual sequencing which represents a variant of the Sanger method, in which unlike the latter, reversible fluorescent terminators are used. During the various sequencing cycles, dNTPs are added in equimolar quantities, and for each of the clonal clusters, mentioned above, the incorporation of a single species of dNTP is accompanied by the emission of fluorescence at a certain wavelength. Fluorescence emissions are recorded by the sequencer's detection apparatus, and the readings of the detection apparatus are then converted into base sequences using an algorithm called base caller. Each sequencing cycle ends with the removal of the terminators using specific enzymes and the washing out of excess reagents from the flow cell which is then ready for the next cycle. The set of sequencing platforms based on Illumina technology allows the sequencing of fragments of length between 50 and 250 bp (in some cases up to 300) with an error rate smaller than 1% that is concentrated at the 3' end of the fragments due to loss of efficiency of the polymerase. Errors due to insertion or deletion are rare and are around 0.01%, with a peak in GC-rich regions, therefore the vast majority of errors consist of base substitutions, with peaks in homopolymeric regions or regions containing inverted repeats. It should be noted that systematic errors due to PCR artifacts add up to Illumina technology's instrumental errors; so the correct sample preparation is a crucial element for the success of Illumina sequencing. Nonetheless, each base yielded by the base caller is assigned a quality score that measures the probability the base has been wrongly called. This quality score Q, called Phred quality score, is calculated using the following equation:

$$Q_i = -10 log_{10} P_i \qquad (1.2)$$

where Qi is the Phred quality score of base i and Pi is the estimated probability that base i has been called incorrectly. According to the previous equation, the larger the Q, the smaller the error probability. For example, a Phred quality score of 20 indicates an error probability of 1%, a Phred quality score of 30 indicates an error probability of 0. 1%, etc. The term RNA-Seq, an abbreviation of RNA sequencing, instead, refers to a set of techniques that, using NGS platforms, can identify, and possibly quantify, RNAs present in samples obtained from cells of interest; mainly used to characterize their transcriptomes. In general, RNA-seq techniques consist of a series of biochemical reactions with the final synthesis of DNA fragments complementary to the extracted RNAs, called cDNA.

RNA sequencing is generally performed using cDNA prepared according to the following steps:

- Isolation of transcripts by DNA degradation using DNase and monitoring of RNA degradation to guarantee an adequate quantity of initial total RNA, followed by a possible enrichment for RNAs of interest and depletion of unwanted RNAs; the mRNA enrichment is often used by which they are selected exploiting the 3' polyadenylated tails by pairing them with oligoT (oligonucleotides composed solely of thymines) on a solid layer, washing away all other RNA classes.
- Synthesis of the first cDNA strand by retrotranscription using reverse transcriptases and specific or random primers.
- Synthesis of the second cDNA strand and generation of the cDNA library by PCR amplification.

It should be noted that during the creation of the second strands of cDNA, the location on the forward or reverse strand of the transcript is lost, therefore to overcome this problem it is usual to synthesize the second strands of cDNAs by using deoxyuridine triphosphates (dUTPs) instead of the common deoxythymidine triphosphates (dTTPs) to allow selective degradation of the second strands by enzymes known as uracil-N-glycosylase and subsequent amplification by PCR of the first strands only. This protocol is called stranded RNA-seq, similarly, the protocol that does not involve the degradation of the first strand is called unstranded RNA-seq. In conclusion, considering both advantages in terms of cost and output volume, as well as its high degree of accuracy, all the datasets used in this work were created starting from RNA-seq data obtained using Illumina technology.

1.4 Machine Learning-Based Tools for A-to-I RNA Editing Detection

As previously mentioned, NGS and RNA-seq technologies made it possible to sequence, with high accuracy, the whole transcriptome providing vast amounts of data to be analyzed. As previously stated A-to-I RNA editing events identification remains a difficult task [[124], [125]]. To date, many bioinformatics pipelines based on empirical methods make it possible to identify bona-fide A-to-I editing events in NGS data. These approaches, however, rely primarily on high-quality RNAseq data and WGS/WES data to reduce the noise originating from genomic sources of variation, alignment errors, and artifacts[[124], [125], [126]]. Even if bioinformatics approaches are undoubtedly reliable, the abundance of transcriptomic data made available by NGS, unlocked the possibility of using Machine Learning (ML) or Deep Learning (DL) algorithms as powerful and useful alternative tools to identify A-to-I RNA editing events, overcoming the classical empirical methods shortcomings in terms of time-consuming and computationally expensive filtering steps [[127]], although their efficacy is highly dependent on the specific input features adopted as well as training data size and quality [[128]]. The current landscape of already published ML- or DL-based algorithms, to detect A-to-I RNA editing events, can be organized into two main categories:

- Sequence-based:

  This class of methods relies on the specific nucleotide sequence containing the putative editing site to discriminate between true or false editing sites. This class of methods is composed of very fast and efficient algorithms, that bypass the alignment to reference genome steps and are well-suited for large-scale analysis conducted within reasonably short computational times. They adopt various strategies for feature selection and embeddings and some of them are still updated with new data made available. On the other hand, their principal disadvantage is that using only nucleotide sequences, they are not able to fully capture the real complexity of transcriptomics information, so the simplified view of the underlying biological phenomena leads to overlooking crucial factors such as the dynamicity of A-to-I RNA editing and its differential regulation among different tissues, limiting their insight capacity. The principal tools of this class are:

– PAI:

PAI is a support vector machine (SVM)-based binary classifier [[129]]. This approach is based on encoding RNA sequences of at least 51 nucleotides, centered in the investigated site, using the pseudo dinucleotide composition (PseDNC) method, belonging to the class of pseudo-K-tuple nucleotide composition (PseKNC) methods. It makes use of six RNA physicochemical properties to represent the RNA sequence of various lengths. PAI was trained via the jackknife test, which is a reliable cross-validation method, and finally tested on an independent benchmark dataset. PAI achieved 79.51% accuracy, 85.60% sensitivity, 73.11% specificity, and a Matthews correlation coefficient (MCC) of 0.60 on the validation dataset. On the independent dataset, PAI identified 247 out of 300 A-to-I editing sites with a sensitivity of 82.33%. PAI was made available as a web server, requiring input sequences in FASTA format.

– iRNA-AI:

iRNA-AI like PAI is an SVM-based method but unlike PAI it has been specifically developed to identify A-to-I editing sites in the human transcriptome [[130]]. iRNA-AI uses sequences of 51 nucleotides, centered in the investigated site, in which each nucleotide is encoded as a vector of 3 physicochemical properties, and was trained and tested using datasets obtained from the DARNED database containing 333216 putative A-to-I RNA editing sites. iRNA-AI achieved a sensitivity of 86.18%, a specificity of 95.23%, an accuracy of 90.71%, and an MCC of 0.82. iRNA-AI was further tested on a balanced independent dataset containing 3243 experimentally confirmed A-to-I editing sites, achieving a sensitivity of 84.19%, a specificity of 89,36%, an accuracy of 93.81%, and an MCC of 80%. iRNA-Ai is also available as a web server requiring sequences in FASTA format.

– iRNA-3typeA:

  iRNA-3typeA is another SVM-based algorithm designed for the analysis of the human and murine transcriptome, but unlike the previous ones is not specific for A-to-I RNA editing, but can also identify N1-methyladenosine (m1A) sites and m6A sites [[131]]. iRNA-3typeA, like PAI, adopts a feature encoding belonging to the PseKNC class. In particular, it uses a single nucleotide pseudo composition based on 6 physicochemical properties, to encode nucleotide sequences of at least 41 nucleotides, centered in the investigated site. iRNA-3typeA was trained, validated, via jackknife-test, and finally tested on thousands of known epitranscriptomics sites both human and murine. Regarding A-to-I RNA editing on the human sites, iRNA-3typeA achieved a sensitivity of 86.18%, a specificity of 95.23%, an accuracy of 90.71%, and an MCC of 82.0% on the test set. Also, iRNA-3typeA is available as a web server accepting RNA sequences in FASTA format.

– iMRM:

  iMRM is an XGboost-based ML algorithm capable of identifying multiple epitranscriptomics modifications, like iRNA-3typeA, in Mus musculus and Saccharomyces cerevisiae, and Homo sapiens as well [[132]. In particular, iMRM can identify m6A sites, 5-methylcytosine (m5C) sites, m1A sites, ψ sites, and A-to-I RNA editing sites. iMRM adopts different feature extraction and feature encoding strategies for each of the aforementioned epitranscriptomics modifications. In the case of A-to-I RNA editing, iNRM encodes RNA sequences of 51 nucleotides, centered in the investigated site, as a feature vector in R470 using a combination of the single nucleotide one hot encoding method, the k-tuple nucleotide composition encoding method and the dinucleotide physicochemical properties encoding method. In the case of the A-to-I RNA editing, training and validations set were built using about 3000 positive and negative examples, and the model was trained both via 10-fold cross-validation and the jackknife test, achieving an Area Under Curve score (AUC) of 0.98 and an accuracy of 91.7%. Like the previous tools, iMRM is available as a web server requiring RNA sequences of 51 nucleotides provided in FASTA format.

– EPAI-NC:

EPAI-NC is another SVM-based method specific for A-to-I RNA editing [[133]]. The model was trained using features based on nucleotide frequency and n-gapped patterns describing RNA sequences from 51 up to 250 nucleotides. EPAI-NC was trained similarly to the previous SVM-based methods and achieved an accuracy of 93.90%, a sensitivity of 96.80%, a specificity of 90.80%, and an MCC of 87.90% on the validation set. The model's performance was further tested on an independent dataset identifying correctly 253 out of 300 known A-to-I RNA editing sites and achieving an accuracy of 84.33%. For user convenience, a user-friendly web server for EPAI-NC was developed, requiring RNA sequences in FASTA format. However, at the time of writing, it seems in dismission.

– EditPredict:

EditPredict is a Convolutional Neural Network (CNN) based binary classifier developed to identify real A-to-I RNA editing sites in RNA sequences of 101 nucleotides, centered in the investigated site [[134]]. EditPredict operates as a pattern recognition algorithm aimed at identifying specific motifs in the A-to-I RNA editing sites' flanking regions. It serves as a DL tool to predict A-to-I RNA editing sites using information retrieved from the human reference genome, encoding each nucleic acid sequence with a matrix of features in which each nucleotide is encoded as a one-hot encoding vector. EditPredict was trained on about 4.67 million publicly available A-to-I RNA editing sites taken from the REDIportal database v1 [[86]] and millions of randomly selected non-edited sites, from the human genome at least 200 bp far from the known A-to-I RNA editing sites. EditPredict achieved values between 95% and 97% of accuracy, precision, recall, and F1-score on the validation set. The model was further tested on an independent dataset based on A-to-I RNA editing sites and not-edited sites identified in RNA-seq experiments on the Jurkat cell line using Jurkat ADAR1 knock-out samples as negative control. EditPredict achieved an accuracy of 99.5% on Alu sequences and an accuracy of 93.6% on not-Alu sequences. EditPredict is available as a web server, requiring different lists of genomic coordinates in text files for Alu and non-Alu sites. EditPredict also offers the possibility to work with pre-trained models specific for GRCH37 or GRCH38 assembly as well as tissue-specific ones.

– ATTIC:

ATTIC is a stacked-ensemble-based ML tool designed to identify A-to-I RNA editing sites in H. sapiens, M. musculus, and D. melanogaster [[135]]. ATTIC is built upon 14 ML algorithms and uses 37 feature encoding methods to efficiently encode RNA sequences of 41 up to 51 nucleotides; many of these encoding strategies are common to the previously discussed ML-based tools. ATTIC was trained on data derived from the training datasets of PAI, PAI-SAE, EPAI-NC, PRESa2i, and iAI-DSAE, using two different cross-validation strategies, specifically, 10-fold cross-validation and leave-one-out testing. ATTIC achieved, depending on the specific stack adopted, values between about 82% and about 99% of accuracy, between about 85% and about 99% of AUC score, between about 83% and about 98% of recall, between about 79% and about 100% of precision, between about 79% and about 99% of F1-score, about 55% - 98% Cohen Kappa and between about 57% and about 9% of MCC. ATTIC is available as a web server requiring RNA sequences provided in FASTA format.

– PAI-SAE:

PAI-SAE is a sparse auto-encoder-based binary classifier developed to detect A-to-I editing sites conceptually derived from PAI [[136]]. PAI-SAE uses feature vectors derived from both the dinucleotide-based auto-cross covariance (DACC) approach, which utilizes ten RNA dinucleotides' physicochemical properties and the pseudo dinucleotide composition approach used in PAI. The PAI-SAE encoding method represents an extension of PAI nucleotide sequence representation, utilizing additional statistics-based features capable of providing a more precise data representation. PAI-SAE was trained via jackknife test validation on the same benchmark dataset as PAI. PAI-SAE achieved 81.97% of accuracy, 87.20% of sensitivity, 76.47% of specificity, and 64.14% of MCC. Unlike PAI, PAI-SAE was not further validated on an independent dataset, so there aren't indications of its performance on real data.

– RDDSVM:

RDDSVM is an SVM-based binary classifier aimed at A-to-I RNA editing detection [[137]] in 101 nucleotide windows centered in investigated sites. The first model is based on two feature encoding methods. The first method, called trinucleotides secondary structure dot-bracket notation, encodes each nucleotide in the 101 nucleotides window as a feature vector in R32 obtained by all the possible base pairing combinations of the triplet centered in it. The second method, called trinucleotide sequence encoding, encodes each nucleotide in the 101 nucleotides window as a feature vector in R64 obtained by all the possible trinucleotide sequences centered in it. RDDSVM was trained using 5000 A-to-I RNA editing sites from the REDIportal database as positive examples and 5000 randomly chosen sites from the reference genome not present in the REDIportal database as negative examples. The training was carried out randomly splitting the dataset into training and test sets with a 50/50 ratio with an equal number of positives and negatives in the training and test sets. RDDSVM achieved a sensitivity of about 90%, a specificity of about 85%, and an accuracy of about 87%. RDDSVM was further tested on an independent dataset including 516 true editing sites and 105 false editing sites, obtained from U87MG wild-type and ADAR1 knock-out (as negative control) cell lines RNA-seq experiments. RDDSVM achieved a sensitivity of about 93%, a specificity of about 77%, and an accuracy of about 90% on the U87MG independent dataset. RDDSVM is available as a docker image and a stand-alone R script through its GitHub repository. It requires text data for its execution.

- Alignment-based

  This class of A-to-I RNA editing detection tools comprises methods that require the alignment of RNA-seq reads to a reference genome as a preliminary and unavoidable step. These algorithms operate at least on BAM files, but in most cases require additional files, such as Variant Call Files (VCFs), RepeatMasker files (rmsk), and so on. Alignment-based algorithms are, in general, far more computationally demanding than the previous ones, as they take into account great numbers of variables and alignment-derived parameters, such that in some cases a single BAM file analysis can last several days, however, they investigate the real state of the transcriptome recorded using RNA-seq experiments and not its transposition at a genomic level. These tools are conceived to leverage the abundance of accurate data made available by the progress in NGS technologies in the last decade, such that they can learn more complex patterns to provide more reliable predictions on various epitranscriptomics modifications, taking into account the biological context implicit in the transcriptomic data. The main alignment-based ML or DL algorithms to predict A-to-I RNA editing from BAM files are:

  – DeepRed:

    DeepRed is a Multi-Layer Perceptron (MLP) based stacked ensemble created to differentiate between true A-to-I RNA editing events and genomic sources of variation in H. sapiens [[138]]. It is structured as an ensemble of hundreds of MLPs clustered in units of forty MLPs, in which each one of them is structured in two sub-units. The first sub-unit takes sequences of 101 nucleotides, each one encoded as a one-hot encoded vector, and the second sub-unit takes sequences of 41 nucleotides also encoded via the one-hot encoding method. The training dataset was obtained from 64 human RNA-seq experiments using both the separate samples method and the pooled samples method. The units were trained separately via 5-fold cross-validation and then put together using the averaging method. DeepRed achieved an average AUC score of 98.1% on the training set and 97.9% on the validation set. DeepRed was further tested on an independent dataset composed of experimentally validated 48 true A-to-I RNA editing events and 47 SNPs, derived from U87 cell-line RNA-seq samples, achieving an accuracy of 93,23%, a sensitivity of 79.66% and a specificity of 99.25%. DeepRed is freely available through a GitHub repository.

– RDDpred:

RDDpred is a random forest (RF) RDD-based binary classifier designed to discriminate true RNA-editing events from artifacts in RNA-seq data [[139]]. RDDpred operates on single genomic coordinates encoded by a feature vector of 15 alignment-derived parameters, which implicitly consider also the surrounding genomic coordinates. RDDpred was trained on a dataset consisting of about 150 thousand true A-to-I RNA editing events, that matched the rigorously annotated editing sites shared by RADAR and DARNED databases, and hundreds of thousands of artifact examples obtained with the MES method. RDDpred was further tested on two publicly available independent RNA-editing datasets; Peng's dataset and Bahn's dataset. RDDpred achieved 90% of accuracy on Peng's dataset and 95% of accuracy on Bahn's dataset. RDDpred is freely available on the GitHub repository as a docker image requiring, other than the BAM files to be examined, the Reference Genome used to produce the alignment in FASTA. It is also highly suggested to provide a Single Nucleotide Polymorphism Database (dbSNP) file.

– RNA-editing-pred:

RNA-editing-pred [[140]] is a dual-module algorithm based on an RF and a Bidirectional Long-Short Term Memory (Bi-LSTM), specific for the differentiation between A-to-I RNA editing events and genomic sources of variation. RNA-editing-pred uses RNA sequences of 101 nucleotides for the Bi-LSTM module and RNA sequences from 50 up to 200 for the RF. Each nucleotide is encoded as a one-hot encoding vector of a variable dependent on both the RNA sequence and its predicted secondary structure in PAIRS format using a linear fold DL algorithm. RNA-editing-pred was trained on three datasets derived from three different species; H. sapiens, M. musculus, and T. trachurus. Both H. sapiens and M. musculus datasets have been collected from the REDIportal database instead T. trachurus dataset has been collected from the Darwin Tree of Life portal. The RF classifier achieved an accuracy of 75% for H. sapiens, 74% for M. musculus, and below 65% for T. trachurus. for all the possible windows of length showing a high degree of robustness. The Bi-LSTM achieved 94.8% accuracy for H. sapiens, 84.6% accuracy for M. musculus, and below 72.4% accuracy for T. trachurus, showing both greater generalization capability and lower consistency over the three datasets concerning the RF module. RNA-editing-pred is also freely available in its GitHub repository.

– RED-ML:

RED-ML is a logistic regression-based ML tool for the A-to-I RNA editing detection from RNA-seq data [[141]]. It takes RNA-seq alignment' BAM files, the Reference Genome used for the alignment in FASTA format, and a dbSNP annotation file; used to rule out matching genomic variants when present. RED-ML utilizes three broad classes of features for its analysis. The first class consists of basic read features, such as the number of supporting reads for a candidate site and the putative editing frequency. The second class includes features related to sequencing artifacts and misalignments, such as mapping qualities of supporting reads, the relative position of the candidate site in the mapped reads, indications of strand bias, and whether the candidate site falls into simple repeat regions. The third class of features is based on known properties of A-to-I RNA editing, such as the different A-to-G substitution pattern, whether the investigated site is in an Alu region or not, and its sequence context, whether the investigated site is in a region-compatible with ADARs' preferred motif or not. In total, RED-ML encodes each sequence of three nucleotides, centered in the investigated sites with 28 features. RED-ML was trained on 1475 experimentally validated positive sites and 2925 negative sites, 1200 of them sampled from dbSNP138, with an 80/20 ratio for the training set and test set. RED-ML achieved an AUC score of 98% and an AUPR score of 94% on the test set. RED-ML was further tested on three independent sets of putative positive A-to-I RNA editing sites, about 50000 from one CH24T cell line RNA-seq experiment, about 40000 from one CH62T cell line RNA-se sample, and about 20000 from one HeLa cell line RNA-seq sample, achieving in all three cases a true positive rate $> 90\%$. RED-ML is freely available on its GitHub repository as a docker image. It requires the BAM files to be examined, the Reference Genome used to produce the alignment in FASTA format, and a dbSNP file. It's highly suggested to use dbSNP138 being the tool optimized on that specific dbSNP release.

– DEMINING:

Demining is a CNN-based framework constructed to differentiate true A-to-I RNA editing events from genomic sources of variations, specifically designed for H. sapiens [[142]]. Its DL module (DeepDDR) takes as input RNA sequences of 51 nucleotides, encoded as 16x51 matrices, whose columns indicate the preprocessed RNA-seq base counts for each nucleotide and the rows correspond to the 12 possible base substitutions plus the 4 cases in which there are no substitutions. DeepDDR was trained on a dataset composed of 122.872 true A-to-I RNA editing events and an equal number of single nucleotide polymorphisms (SNPs), derived from 403 RNA-seq samples taken from the 1000 Genomes Project as well as their corresponding WGS samples. DeepDDR achieved an AUC score of 99.4% and an AUPRC (area under the precision-recall curve) score of 99.4% on the test set. DeepRDD was further tested on two independent datasets; one taken from the 1000 Genomes Project and another derived from M. musculus bone marrow RNA-seq experiments on ADAR1 wild-type cells and ADAR1 knock-out cells as a negative control. DeepRDD achieved an AUC score of 99.3% and an AUPRC score of 97.9% on the 1000 Genomes Project independent dataset, instead, it reached an AUC score of 95.9% on the M. musculus dataset. Furthermore, DeepRDD correctly identified about 7600 previously reported real editing events in 19 RNA-seq samples from acute myeloid leukemia (AML). DEMINING is available for download, under an open-source license, from the GitHub repository or Zenodo repository.

– TAE-ML:

TAE-ML is a RF based binary classifier designed to discriminate true A-to-I RNA editing sites among the various Single Nucleotide Variants (SNVs) present in VCF files [[143]]. TAE-ML is based on 600 Decision Trees in which the maximum features, the number of random features considered in the node splitting, is set to the square root of the total number of features. TAE-ML adopts various filters to rule out sequencing artifacts and utilizes RNA sequences of 101 nucleotides. in which each nucleotide is encoded as a feature vector of 33 alignment-based parameters. TAE-ML achieved 39.1% precision, 90.9% recall, 92.5% accuracy, and 54.6% F1-score on Hela cells, 55.3% of precision, 89.9% recall, 90.9% of accuracy, and 68.5% of F1-score on CH24T cells and 57.8% of precision, 84.4% of recall, 93.4% of accuracy and 68.6% of F1-score on CH62T cells. Unfortunately, the TAE-ML algorithm is not publicly available.

It's worth mentioning that in the last few years, with the advent of third-generation sequencing technologies, a new class of DL-based tools to identify epitranscriptomics modifications in RNA-seq data has emerged. To date, the new class of algorithms specific for third-generation sequencing data, in particular Oxford Nanopore Technology (ONT), has given the opportunity to directly identify inosine in transcripts. However, it is still affected by higher error rates than the previously mentioned platforms. Most likely as the advancement in third-generation sequencing technologies will provide high-quality data, like the ones provided by the second-generation, this class of tools will become the dominant one showing advantages of the third-generation sequencing over NGS. In conclusion, since the detection of A-to-I editing is still challenging and ML or DL approaches are promising computational alternatives, this work aimed to develop a DL algorithm based on Temporal Convolutional Networks, which represents an efficient and low-demanding class of Artificial Neural Networks to analyze temporal sequences data.

1.5 Temporal Convolutional Networks - A Technical Outlook

In the context of Deep Learning, the modeling of sequences, in particular temporal sequences, is generally accomplished through recurrent and recursive architectures given their ability to retain, in their memory cells, information regarding chronologies, in principle, indefinitely long. However, several studies concerning the modeling of temporal sequences using convolutional architectures suggest that these can achieve performances equal to or superior to recurrent architectures in the fields of speech synthesis [[144]], natural language modeling [[145], [146]], and translation [[147], [148]]. In detail, results obtained on various benchmarks commonly used for the evaluation of recurring architectures (text8, MuseData, Nottingham, JSB Chorales, Penn Treebank Corpus, etc.) the Temporal Convolutional Networks (TCN) have not only provided better performances compared to the canonical Long Short- Term Memory (LSTM) and Gated Recurrent Unit (GRU), thus paving the way for the use of TCNs in sequence modeling in different application fields, but have also allowed a simpler and clearer formalization of the problem of temporal sequence analysis[[149]].

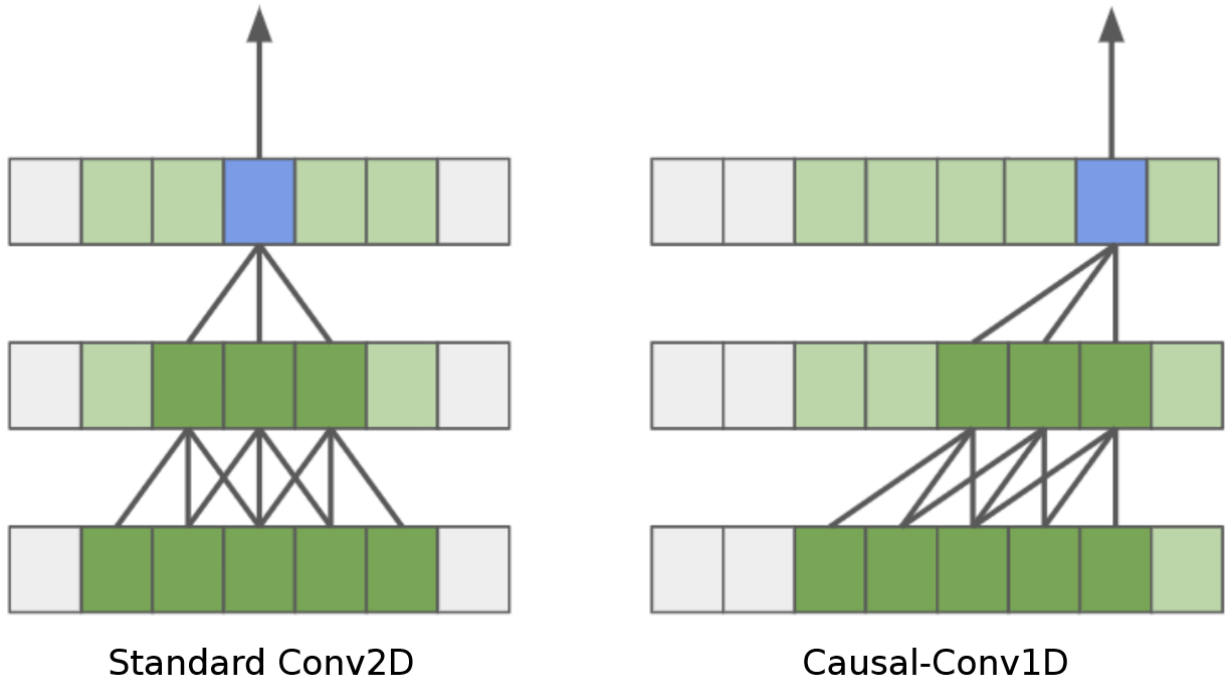

Figure 1.8: Schematic difference between Standard Conv2D and Causal-Conv-1D. Taken and modified from Kondratyuk D, Yuan L, Li Y, Li Zhang, Tan M, Brown M, Gong B. IEEE/CVF, 2021, pp. 16015-16025.

In general, convolutional neural networks (CNN), conceived by LeCun in the 1980s [[150]] in the field of image recognition, have almost immediately been applied to the analysis of phoneme sequences [[151]], but only in recent years have they been extensively applied in the field of Natural Language Processing (NLP) regarding the analysis of word semantics [[152]], classification of sentences or entire documents [[147],[153], [154]] and translation [[145], [146]]. The broad sector of natural language modeling, as previously mentioned, has long been the almost exclusive prerogative of Recurrent Neural Networks (RNN), which, although difficult to train [[155]], have the aforementioned hidden memory vector which, in theory, contains a representation of everything that has been seen by the recurrent layers during the analysis. Over the years, given the popularity of recurrent networks, several studies have been conducted to evaluate their actual performance, also focusing on the development of different recurrent architectures to fix one of their main limitations, which is the vanishing gradient phenomenon. Indeed, canonical RNNs, in the application fields in which they are normally used [[155], [156]] are not able to deal with long-distance correlations within data so, because of that, LSTM was theorized and implemented. Several deep investigations of recurrent-based architectures elucidated their behavior and performance highlighting, on the one hand, that other types of non-recurrent neural architectures provided better results and, on the other hand, that different variants of LSTMs either did not provide significant improvements compared to canonical LSTMs [[157]] or even compromised their effectiveness [[158]]. To improve the performance of recurrent architectures, hybrid architectures have been proposed that combine LSTMs with CNNs, such as Convolutional LSTMs [[159]], in which dense layers are replaced with convolutional layers, Quasi-RNNs [[160]] in which recurrent layers and convolutional layers are alternated, and Dilated-RNNs [[161]] in which the principle of dilated convolutions is incorporated into the recurrent layers, which have brought an improvement in the performance of recurrent architectures. The latter paved the way for the use of purely convolutional architectures instead of RNNs, in particular TCNs, that can capture long-range dependencies without being subject to information loss while proceeding along the temporal axis, known as catastrophic forgetting. The distinctive feature of TCNs, compared to classical CNNs, can be summarized in the use of hierarchies of causal dilated convolutions that allow, in an efficient and computationally parsimonious way, both to expand the receptive field, to reconstruct dependencies over very long chronologies, and to prevent information leakage from future to past [[162], [163]]. Going into technical details, a generic input sequence can be described as a vector $X \in R^{t+1}$ — $X = [x_0, x_1, x_2, \ldots, x_t]$ where x_0 corresponds to the input at time 0, x_1 to the input at time 1, and so on. To predict the output vector $Y \in R^{t+1}$ — $Y = [y_0, y_1, y_2, \ldots, y_t]$, corresponding to the input vector X, under the condition that the output y_i, at time i depends only on the vector X_i=[x_0, x_1, x_2,

..., x_i] $\forall$ i $\in$ [0, t], subset of the vector X corresponding to the inputs observed up to time i, the objective becomes to model through a neural network any probability distribution described by a function f: X $\rightarrow$ Y such that y_t = f([x_0, x_1, x_2, ..., x_t]) [[149]] minimizing the corresponding loss function L(Y, f(X)). This formalization is generally carried out employing autoregressive (AR) models in which the input x_{t+1} at time t+1 becomes the output y_{t+1} at time t+1 according to the Yule-Walker equation [[164]]:

$$X_{t+1} = \sum_{i=1}^{p} \varphi_i B^i X_{t+1} + \varepsilon_t \tag{1.3}$$

Where p is the number of model parameters, $\phi_1, \ldots, \phi_p$ are the model parameters, $B_i, \ldots, B_p$ are the backshift operators that move the summation to the left, ϵt is the noise at time t and X

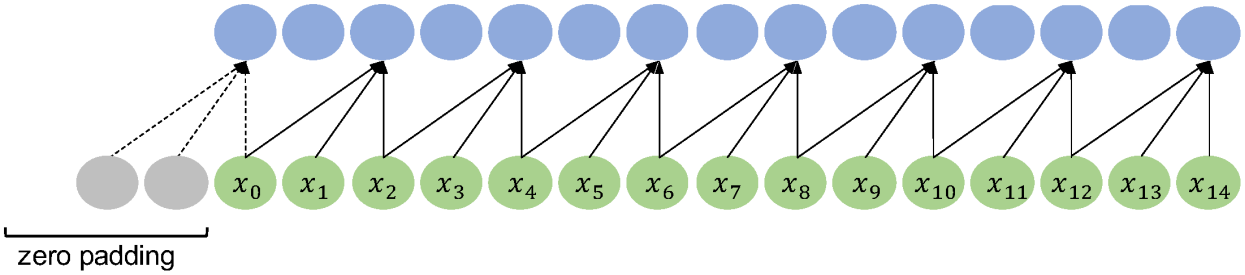

Figure 1.9: Temporal dimension width maintenance by means of zero padding in TCNs. Taken and modified from Ha D-A, Liao C-H, Tan K-S, Yuan S-M. Mathematics, 2021, vol 9, issue 24

In detail, Causal-Conv1Ds keep the width of the time windows unchanged by convolving the generic input xi at the time i in a single operation using kernels in RNxM, where N are the input channels and M is the width of the kernel, and usage of zero padding of length P = (S - 1)(T + 1) + M - S, where T + 1 is the length of the input sequence, M is the width of the kernel, and S is the convolution stride. Unlike Conv2D in Causal-Conv1D, zero padding is placed only to the left of the input x_0 at time 0 to ensure that the generic output $y_i=f(x_i)$ at time i depends only on the input vector $X_i=[x_0,x_1,x_2,...,x_i]$ (Figure 1.9). The general operating mechanics differences between Causal-Conv1D and standard Conv2D are shown schematically above (Figure 1.8). However, the use of Causal-Conv1D does not represent in itself the characterizing element of TCNs, since this class of convolutions has already been used in the Time-Delay Neural Networks (TDNNs) proposed at the end of the 80s [[151]] which have the disadvantage of requiring large-size kernels, or large numbers of layers, to effectively capture long-range dependencies along the temporal axis, making them impractical given their high cost in computational resources. Classical Conv1D-based CNNs have receptive fields, the portion of the input convolved by the neurons, of length r = 1 + n(k -1), where n is the number of layers and k is the kernel width. As can be easily deduced from the previous equation, the full coverage of the input sequence leads to the linear increase in both kernel size and number of layers, concerning the increment in the depth of the time dimension. The real innovation brought by TCNs consists in the use of dilation in convolutional layers [[144], [165]], which allows an exponential growth of the receptive field as a function of the number of layers [[165]], thus solving the above-mentioned TDNNs disadvantage (Figure 1.10). Specifically in the field of convolutional architectures, dilation refers to the distance between individual input elements that are convolved together. In other words, dilation consists of inserting fixed steps in the input reading such that individual portions of the convolved input do not necessarily have to be adjacent to each other. Dilation-based convolutions are called Dilated Convolutions; and in particular, Dilated Causal 1-Dimension Convolutions are the building blocks of TCNs. Using dilations the width of a TCN receptive field r is mathematically described by the following equation [[166]]:

$$r = 1 + \sum_{i=1}^{n} (k - 1)b^{i} = 1 + (k - 1)\frac{b^{n}-1}{b-1} \tag{1.4}$$

Where n is the number of Causal-Conv1D layers, k is the kernel width, and b is the exponentiation base, generally equal to 2; the factor bi is called dilation rate and represents the length of the dilation.

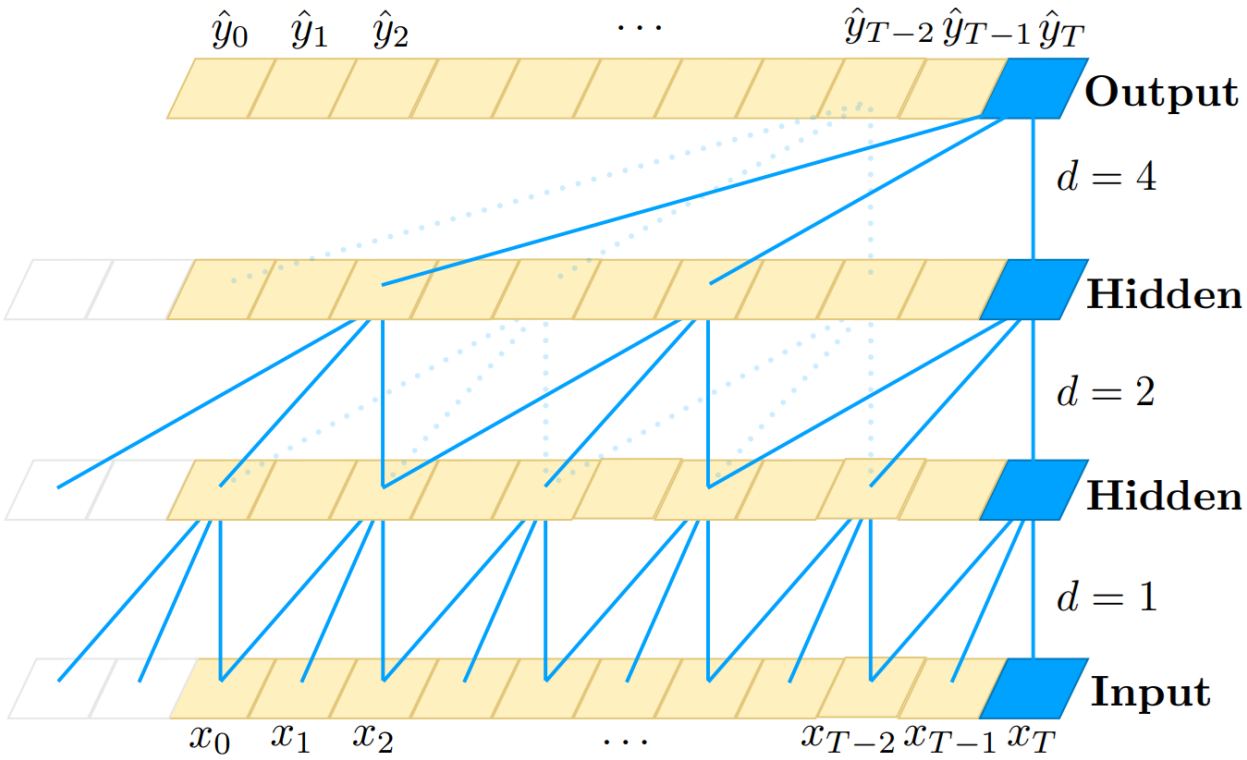

Figure 1.10: Dilated Convolutions Schematics. Taken and modified from Bai S, Kolter JZ, Koltun V. arXiv 2018, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling".

Therefore, considering the exponential dependence between the receptive field and the number of layers, in the case of dilated convolutions, it becomes possible to entirely cover large time windows, with limited kernel widths and much fewer layers compared to TDNNs. Given the equation 1.4 the minimum number n of convolutional layers required to cover the entirety of the temporal dimension is given by the following equation:

$$n = log_b\left(\frac{(l-1)(b-1)}{k-1} + 1\right) \tag{1.5}$$

where b is the receptive field exponentiation base, l is the input sequence length and k is the kernel width. The presence of dilation also modifies the causal convolution zero padding length P, that, in this case, is described by the following equation:

$$P = 1 + (S - 1)(T + 1) + D(M - 1) - S \tag{1.6}$$

where T + 1 is the length of the input sequence, M is the width of the kernel, S is the convolution stride and D is the dilation rate. Although the use of Dilated Causal-Conv1D

guarantees effective coverage of the entire time window, TCNs need to have deep architectures to be able to abstract complex patterns; in other words, they must possess a high number of layers to achieve high generalization capability. Indeed, it has been empirically demonstrated that the pattern recognition capacity of an artificial neural network (ANN) is directly proportional to the number of layers that compose it, however, as much as a generic ANN is deep, as much it can be affected by the problems known as exploding gradient or vanishing gradient. The exploding gradient is characterized by a large increase in the gradient norm which leads to large weight updates, on the contrary, the vanishing gradient is characterized by an increase close to zero in the gradient norm such that the weights of the network remain almost unchanged; in both cases the network loses the ability to learn. While the exploding gradient is normally solved by applying functions to the gradient that guarantee it fits within threshold values, a technique known as gradient clipping, the vanishing gradient is elegantly solved by organizing the network into Residual Units (also called Residual Blocks) [[149]] having Shortcut Connections (also called Residual Connections) within (Figure 1.11). A Residual Block [[167]] is the building block of a specific class of neural networks, called Residual Networks (ResNets), which can be roughly described as a sequence of Residual Units stacked on top of each other. In detail, the Residual Unit is itself a neural architecture that models a composite function σ [[149]] such as:

$$\sigma = G(x + F(x)) \tag{1.7}$$

where G is a generic activation function, F is also a composite function that models a series of transformations to the input and x is a generic input tensor. It must be noted that the use of Residual Blocks, on the one hand, ensures that the flow of information reaches the deepest layers, preventing the vanishing gradient, but on the other hand, allows the network to learn the changes made to the data by a certain transformation rather than the transformation itself, further contributing to the stability of deep networks.

Taking into account that the receptive fields of TCNs depend exponentially on their depth, i.e. their number of layers, it is evident how their stabilization using Residual Blocks, becomes an indispensable tool, especially in the presence of limited kernel widths. However, while in classical ResNets, the output tensor of the composite function F is added directly to the input tensor of the Residual Block, in TCNs the aforementioned input tensor x and output tensor y can have different dimensions, i.e. $x \in R^{HxKxN}$, $y \in R^{HxKxM}$ — $N \neq M$ (where H is the number of inputs, K is the temporal dimension, N are the input channels and M are output channels), making the sum of these two tensors not feasible. To overcome the problem of the possible discrepancy between the dimensions of the input and output tensors, usually in TCNs, Residual Blocks may have a 1X1 Conv1D layer such as to ensure, through

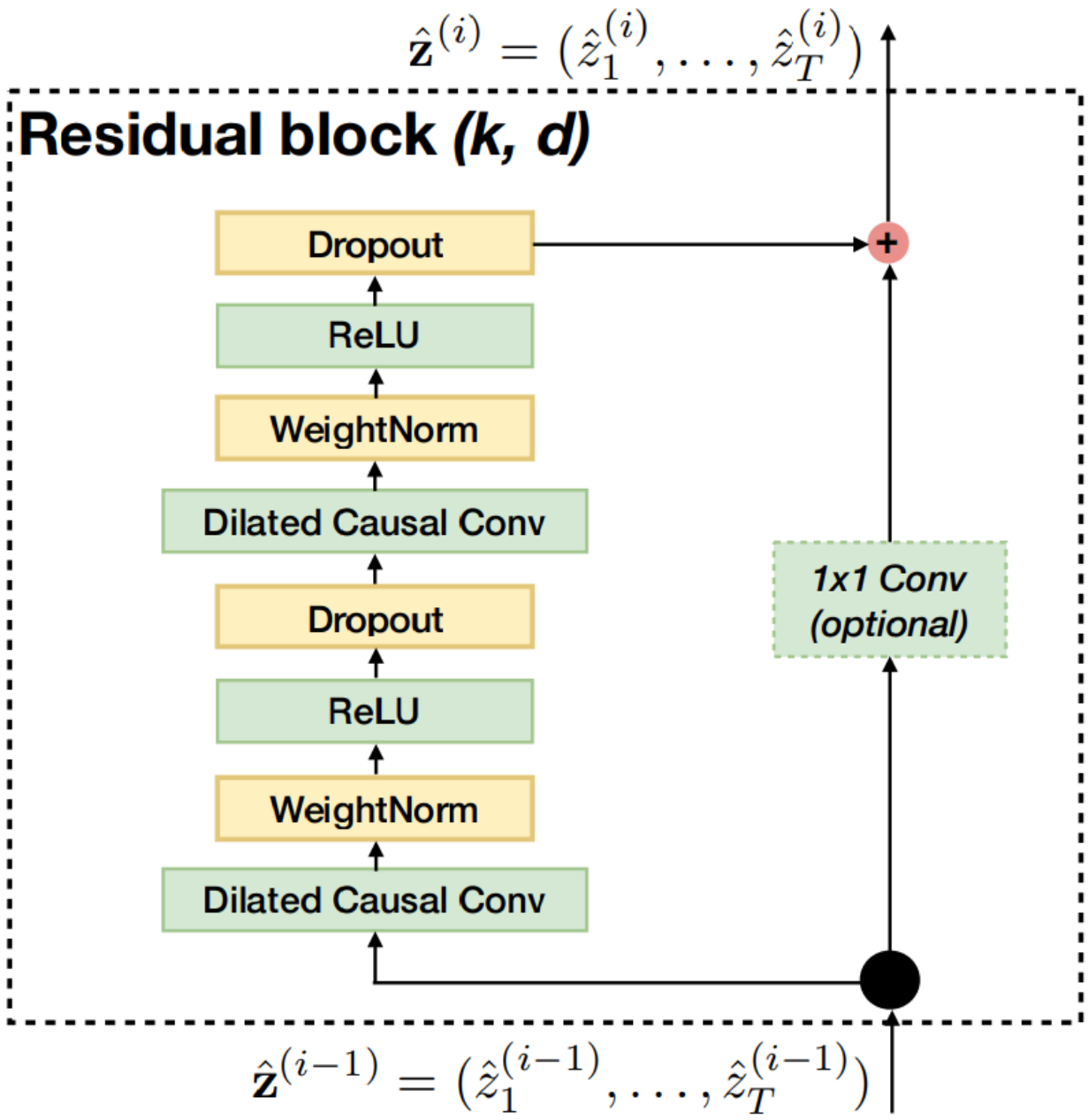

Figure 1.11: Residual Block Schematics. Taken and modified from Bai S, Kolter JZ, Koltun V. arXiv 2018, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling".

upsampling or downsampling, the input tensor is projected into the same space as the output tensor, or vice versa. In summary, in RNNs, predictions along the time axis are performed sequentially while the convolution operations of TCNs allow multiple time steps processing at the same time, both in the training phase and in the inference phase. In RNNs, it is possible to modulate the size of the memory cell only by varying the size of the memory

vector. In the case of TCNs, it is possible to modulate the receptive field in different ways; by modifying the number of layers, the width of the kernel, or the dilation rate [[149]]. In other words, TCNs allow a finer control of the used memory. Empirical data [[149]] demonstrate that the use of Residual Blocks, a typical element of TCNs, allows more effective control of the gradient compared to both classical recurrent architectures and LSTMs and GRUs. During training for very long time sequences, backpropagation for LSTMs and GRUs tends to require a lot of memory consumption to retain partial results from multiple cell gates, while for TCNs, having shared filters on each layer, backpropagation tends to require less memory. Like RNNs, TCNs can process variable-length inputs, making them very flexible drop-in replacements for sequential data. Considering the biological problem of A-to-I RNA editing detection can be modeled as the analysis of a temporal sequence of 2n+1 time steps, in which the value corresponding to time t=n+1 (i.e. the central inosine) is dependent on the values at t<n+1 (i.g. guanosine depletion on the 5' side of the edited site as shown in Figure 1.4) and affects the values at t>n+1 (i.g. guanosines enrichment on the 3' side of the edited site as shown in Figure 1.4), and the advantages of TCN architectures over RNNs, especially in terms of maintaining stable gradients and low memory consumption during the backpropagation phase, as previously stated, the object of this work is the development of TCN-based binary classifier for the identification of potential A-to-I RNA editing sites in RNA-seq data.

1.6 Aim of The Thesis

A-to-I RNA editing is pervasive in human transcriptomes [[168], [87]] and it is carried out by the ADAR enzymes operating on double-stranded RNAs [[69]]. Depending on the location of adenosines along RNAs, i.e. CDS, intron, or UTRs, A-to-I RNA editing may have various functional downstream effects [[77]]. It can change the neuroreceptors' functionality, affect alternative splicing, increase the proteomic complexity, inhibit the biogenesis of circRNAs and modulate gene expression through edited miRNAs or edited miRNA binding sites in target RNAs [[77], [169]]. ADAR-mediated changes in repetitive regions, especially those in opposite orientations formed by Alu repeats [[170]], can suppress the type I interferon signaling through the MDA5-MAVS axis [[171], [110], [109]], tuning the cytosolic innate immunity. A-to-I events in CDS regions, although limited in number, play a crucial role in cell physiology, and their dysregulation has been associated with several human disorders [[172], [76]]. The advancements in NGS technologies, coupled with the advent of RNA-seq technology, have unlocked the profiling of A-to-I RNA editing in the human transcriptome. Nonetheless, the accurate detection of RNA editing events in RNA-seq data remains challenging [[124], [125]]. To date, several empirical methods to identify A-to-I RNA editing in RNA-seq data have been developed. They need high-quality RNAseq data coupled with WGS/WES data to rule out potential genomic sources of variation as well as sequencing errors and artifacts [[124], [125], [126]]. These methods are based on multiple filtering steps and require reliable annotations from public resources. Consequently, they are time- and resource-consuming requiring complex computational procedures [[124], [125], [126], [173]]. Recently, various methods based on Machine- and Deep-Learning algorithms, have been released to overcome empirical methods' limitations in terms of execution time and hardware resource consumption [[127]]. The generalization capability of these methods is strongly dependent on both the set of input features and the size and quality of the training data [[128]]. Given the vast amount of known A-to-I events stored in the REDIportal database [[168], [174]], collected from more than 9000 GTEx [[78]] RNA-seq experiments through a rigorous computational protocol [[175]], the aim of this thesis consists in the development of a DL-based binary classifier able to discriminate genuine A-to-I RNA editing events in Illumina RNA-seq data, exploiting a TCN architecture.

Chapter 2

Results and Discussion

2.1 Feature Selection, Feature Extraction, and TCN Training

Unlike most of the machine (ML) and deep learning (DL) tools for detecting A-to-I changes [[134], [141], [139]], our approach employs RNAseq base counts encoded and normalized as frequencies on windows covering genomic regions flanking putative events to capture both the dynamic nature of RNA editing and the sequence-specific motif of ADARs, as well as its tissue and cell type specificity [[176], [177]].

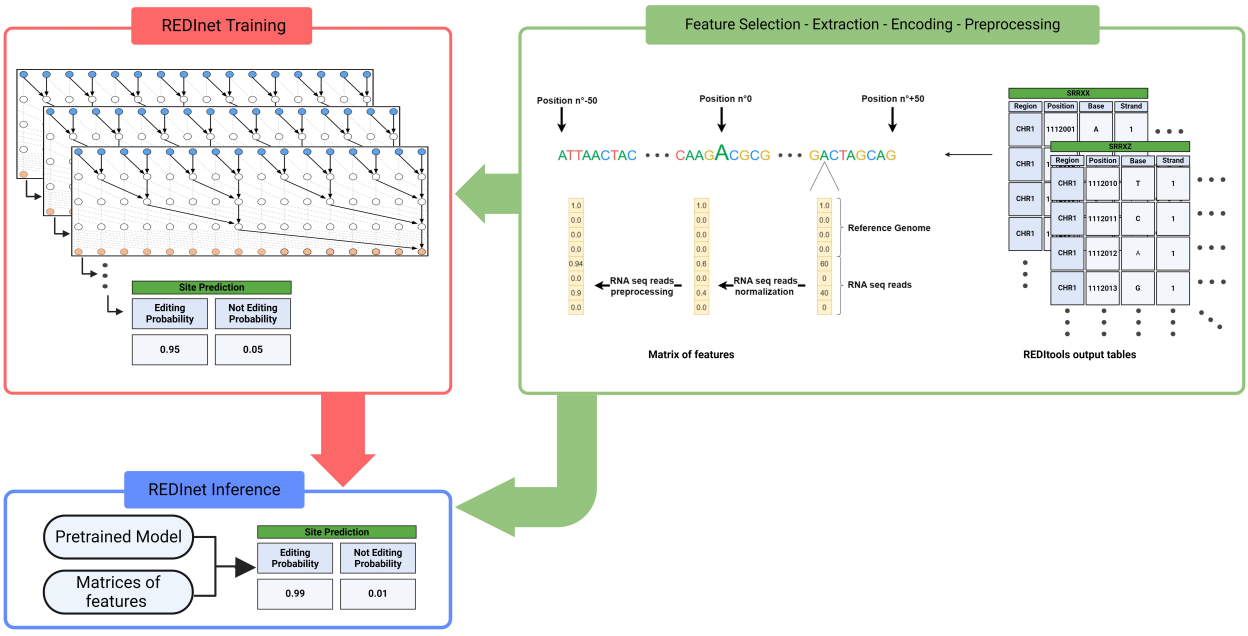

Figure 2.1: Overview of the pipeline used to train and test the TNC-based binary classifier.

Through the REDItools protocol [[168]] we have processed 8404 RNA-seq data (Figure 2.1) from 55 human body sites of the GTEx project in combination with their WGS/WES data. We collected about 12.5 million unique examples for training, validation, and testing purposes (Table 2.1). Positive sites, comprising bona fide A-to-I changes already stored in the REDIportal database, displayed an unimodal distribution of editing levels with values < 0.2 for the vast majority of events (Figure 2.2 , blue line), as expected for bulk human tissues [[168], [177]].

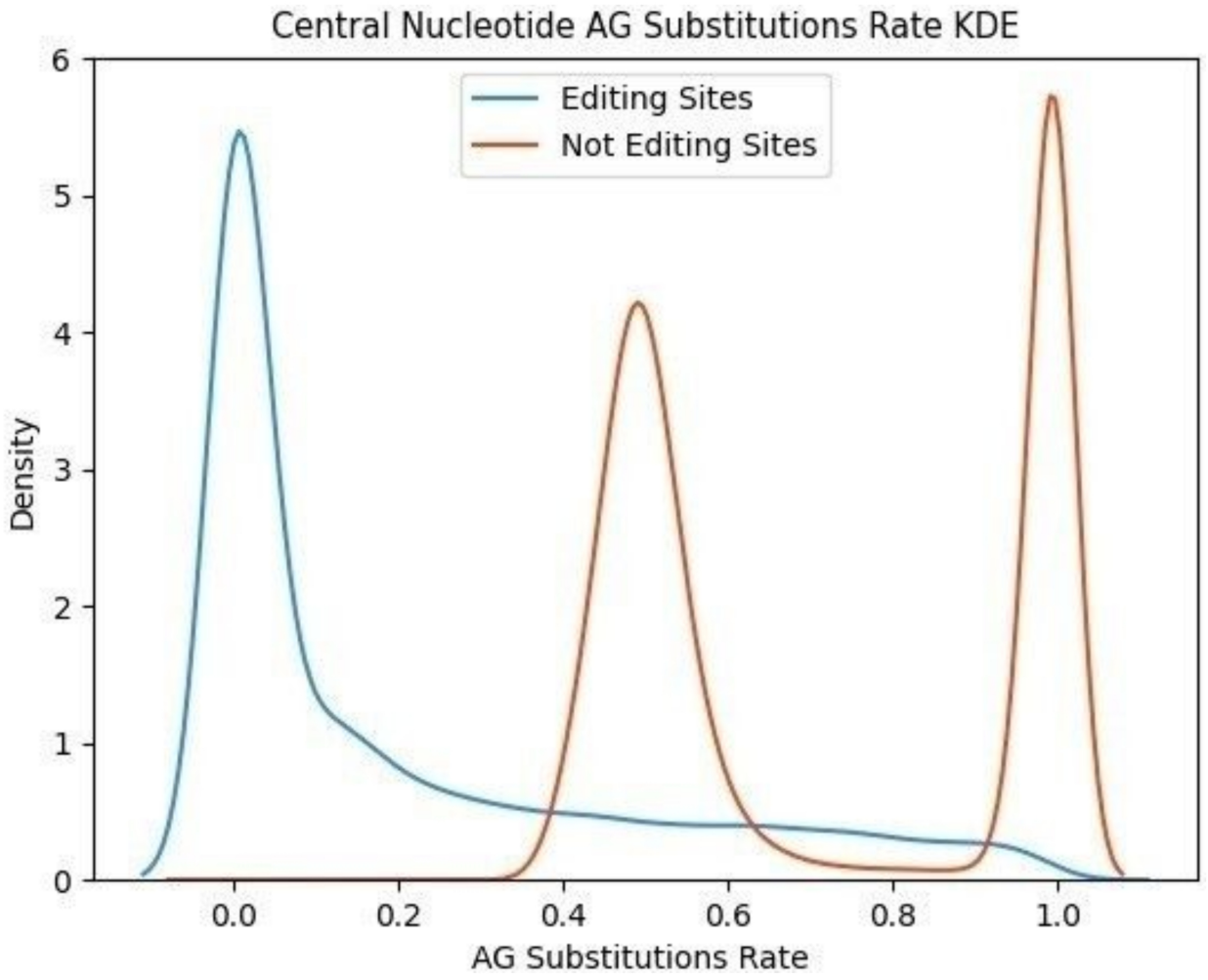

Figure 2.2: A-to-G substitution rates for true A-to-I RNA editing events and SNPs.

Negative sites, including mainly SNPs, were selected using coupled RNAseq and WGS/WES data. Their A-to-G substitution rates were, as expected, bimodally distributed with picks at 0.5 and 1.0 (Figure 2.2, red line).

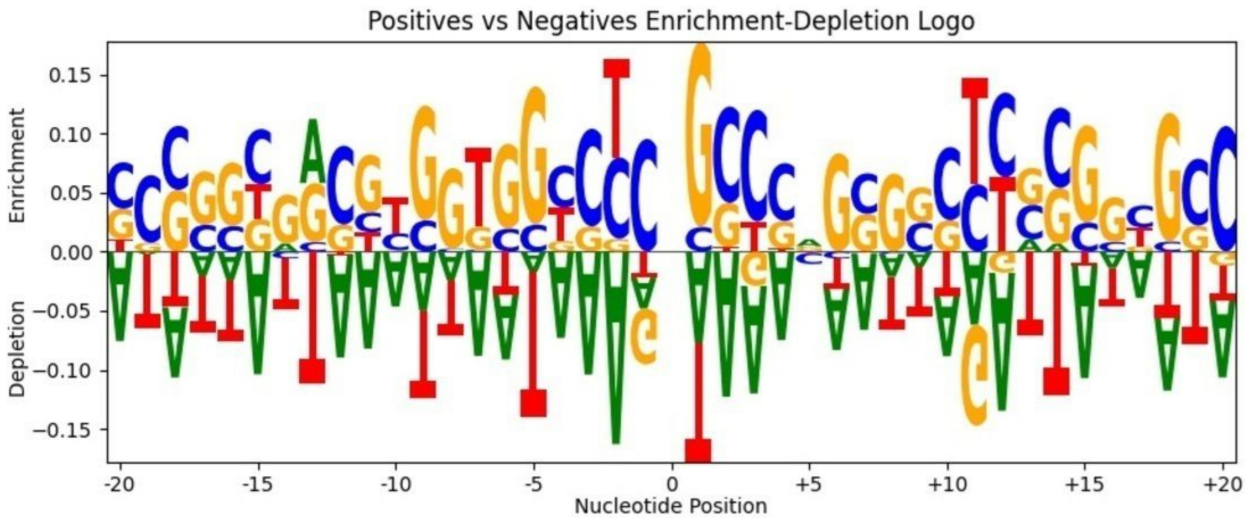

Figure 2.3: Two Sample Logos of Putative positive A-to-I RNA editing sites against putative negative ones.

Randomly selected sites with no evidence of RNA editing at the RNAseq level and concurrent no evidence of SNP at the genomic level were also included in the group of negatives. To better characterize the sequence context of collected examples, a Two Sample Logo of positive versus negative sites was created. As expected, positive A-to-I RNA editing sites (Figure 2.3) matched the expected ADAR motif (Figure 1.4).

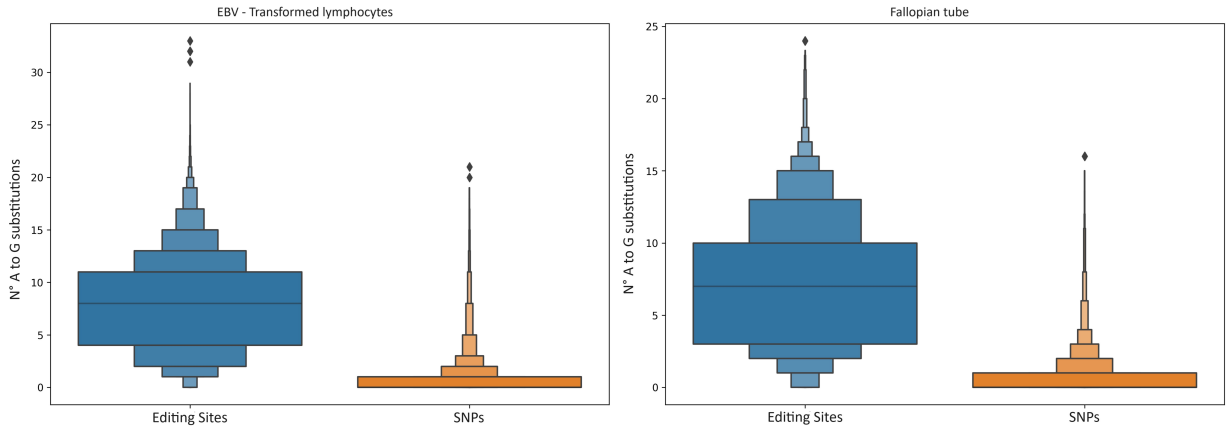

Figure 2.4: Distributions of in A-to-G substitutions in identified A-to-I RNA editing and SNPs sites flanking regions. For the sake of brevity, only the distributions of two body sites are reported.

In particular, positive sites showed a guanosine depletion one base upstream and a guanosine enrichment one base downstream, as previously reported in the literature [[178], [179]]. In addition, since A-to-I RNA editing events tend to occur in clusters [[77]], especially in Alu-rich regions, the bases surrounding extracted sites were further explored by counting the

numbers of A-to-G substitutions. As expected, bases surrounding positive A-to-I RNA editing sites were enriched in A-to-G substitutions (Figure 2.4), while bases surrounding SNPs were depleted (Figure 2.4). This further finding proves the accuracy and reliability of our datasets for training, validation, and testing, comprising high-quality A-to-I RNA editing sites and SNPs. Using the pre-computed REDItools output tables, for each putative or not putative A-to-I RNA editing site, gap-free flanking regions of 50 nucleotides (upstream and downstream) were extracted recovering RNAseq base counts and the corresponding nucleotide sequences of the reference genome. Each example was encoded by a matrix of 8x101 features (Figure 2.1) used to feed a TCN-based neural architecture, after removing all duplicates.

	N° Positives	N° Negatives	Total
Training-Set	4431461	4341336	8772797
Validation-Set	948362	931524	1879886
Test-Set	948946	930939	1879885
A549	6287	6856	13143
HEK293T WT-KO	20806	9589	30395
HEK293T OE-KO	35436	9752	45188
HEK293T (Public Data)	6919	6541	13460
U87	1622	1254	2876

Table 2.1: Dataset compositions detail.

This model (Figure 2.5) employs a sequence of ten Residual Units [[167], [180]], each one composed of a series of Causal-Dilated Conv1D layers with a maximum dilation rate of 32.This setting provides a wide receptive field to cover the entire temporal dimension, according to the equation previously reported.

	Accuracy	Precision	Recall	F1-Score	Specificity	G-Mean
A549	0.962	0.932	0.990	0.961	0.934	0.962
HEK293T (Public Data)	0.952	0.931	0.980	0.955	0.923	0.951
HEK293T WT-KO	0.956	0.969	0.981	0.975	0.931	0.956
HEK293T OE-KO	0.956	0.981	0.980	0.981	0.932	0.956
U87	0.975	0.974	0.985	0.979	0.966	0.975
Training-Set	1000	1000	1000	1000	1000	1000
Validation-Set	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
Test-Set	0.9998	0.9999	0.9998	0.9998	0.9999	0.9998

Table 2.2: Proposed TCN performance on each dataset.

It provides also enough depth to the network to effectively learn complex patterns.

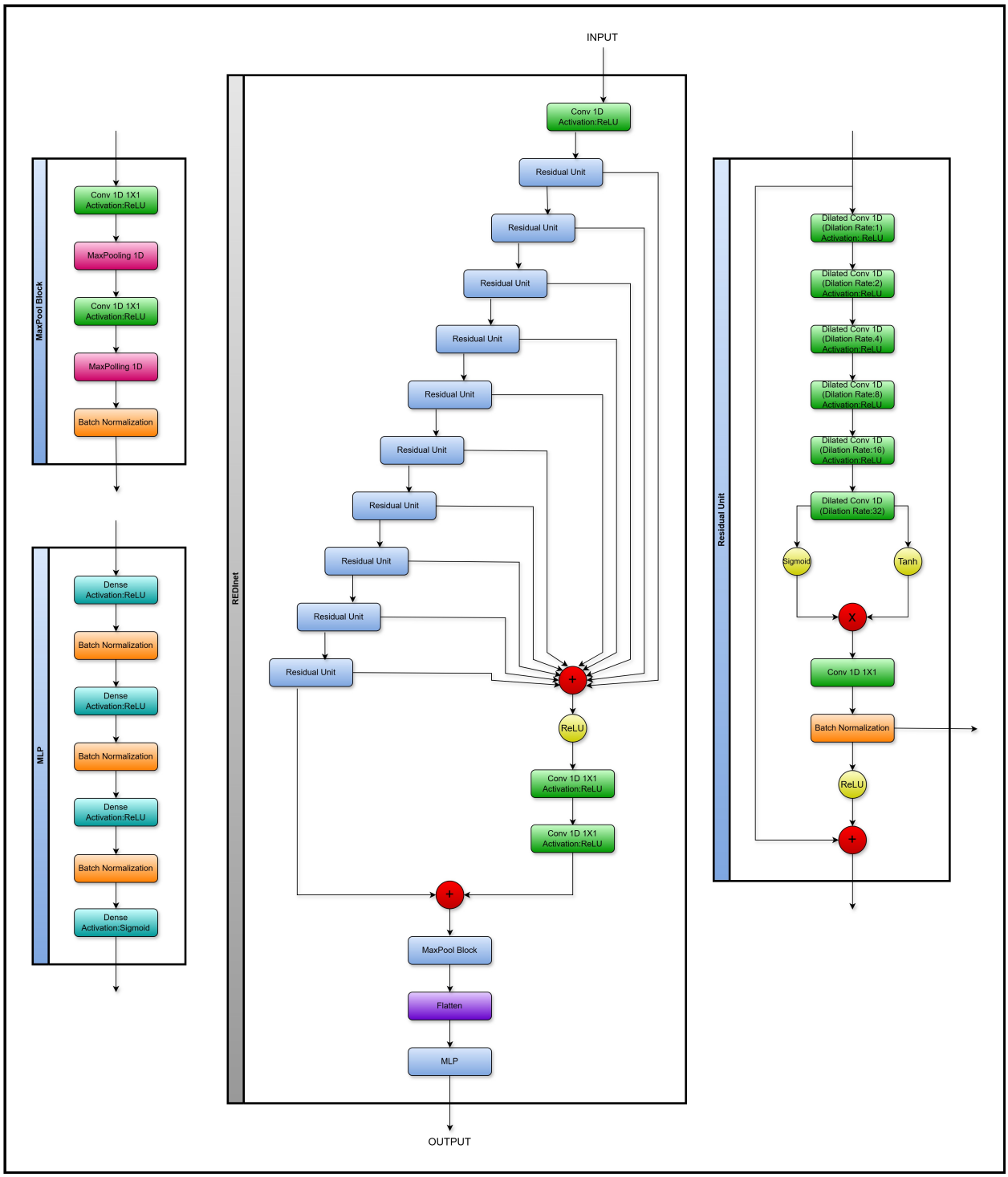

Figure 2.5: Proposed TCN architecture schematics.

As required by the Residual Units, general architectural schematics residual connections were adopted and according to the WaveNet architecture schematics skip connections were

also adopted as an effective way to maximize the retention of produced information along the entire network. The model was fine-tuned via grid search and 10-fold cross-validation to find the better set of Batch Normalization parameters, max dilation rate, and Conv1D filters. Next, a final model was trained by randomly choosing 70% of the collected examples for the training set (Table 2.1) and 15% for the validation set (Table 2.1). The remaining 15% of examples were used as a test set (Table 2.1) to evaluate the model performance. The selected model achieved on all 3 datasets balanced accuracy values of $> 99.9\%$, precision values of $> 99.9\%$, recall values of $> 99.9\%$, F1-score values of $> 99.9\%$, specificity values of $> 99.9\%$, and G-mean values of $> 99.9\%$ (Figure 2.7 and Table 2.2). These last results prove both the correctness of the fine-tuning step and the absence of overfitting and underfitting.

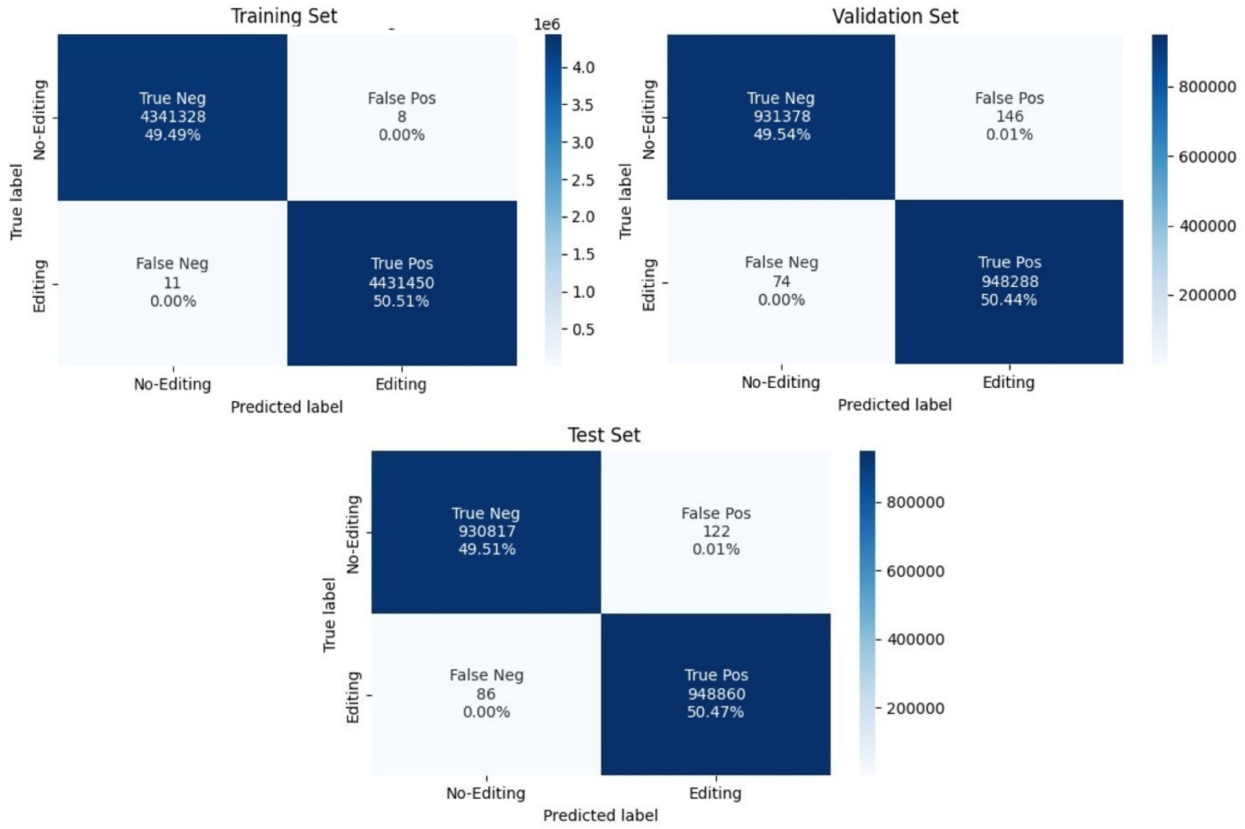

Figure 2.6: Proposed TCN confusion matrices of training set, validation set, and test set.

2.2 Performance Assessment On Independent Datasets

To better assess the model performance on real data, it was tested on four datasets comprising thirty-four RNA-seq experiments from three immortalized cancer cell lines, A549, HEK293T, and U87, downloaded from the equence Read Archive (SRA) database (not included in the GTEx portal) and never seen by the model (Figure 2.1, Table 2.2).

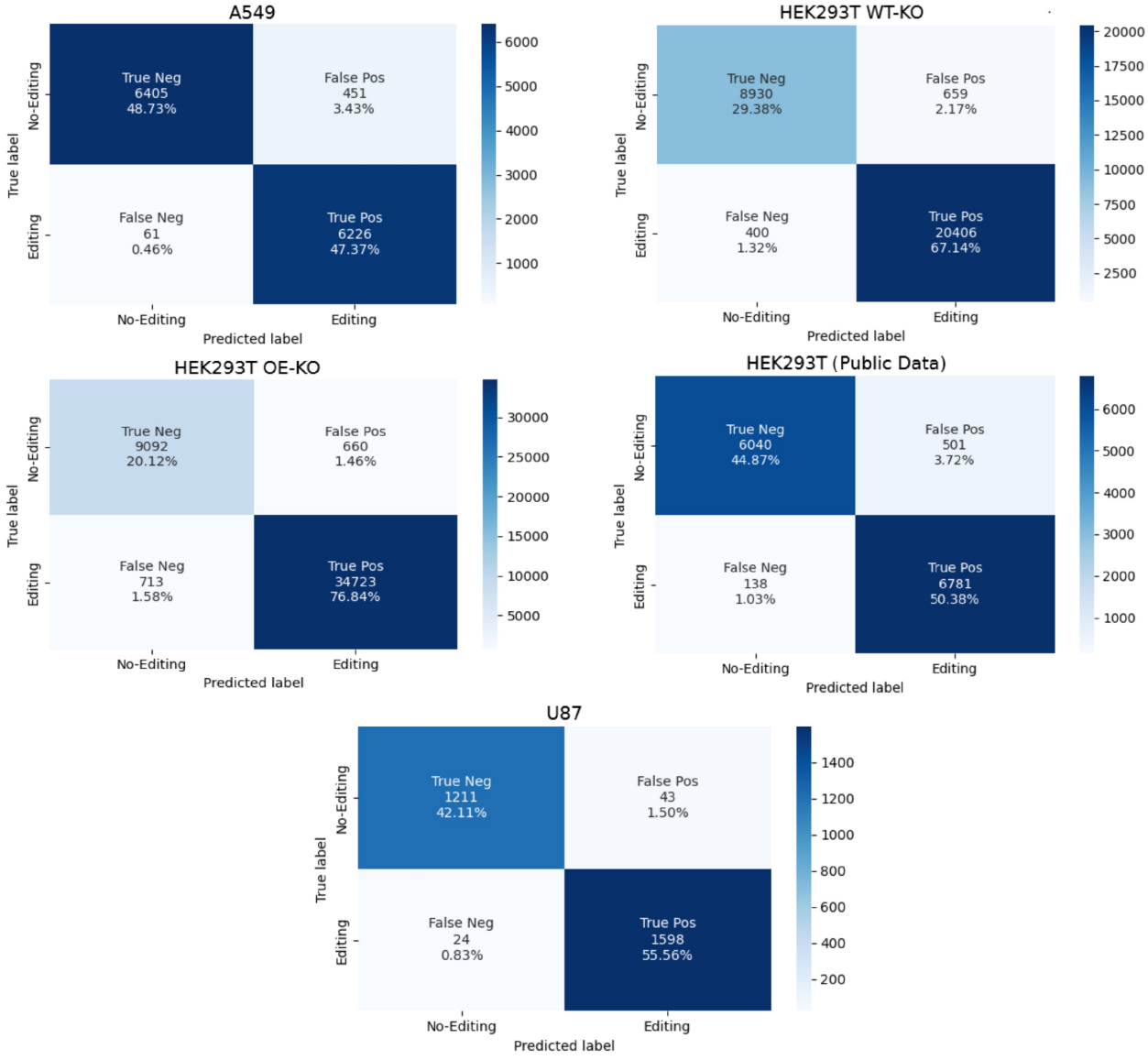

Figure 2.7: Proposed TCN confusion matrices of the model on the independent datasets.

For each cell line, the corresponding WGS data were also downloaded from the SRA database and used as sources of known SNPs to filter out erroneous A-to-I RNA editing calls. Each dataset included both bona fide positive sites and bona fide negative sites in different ra-

tios, which dictated the necessity to adopt both balanced accuracy and G-mean as well as precision, recall, F1-score, and specificity as reliable metrics for the performance evaluation. To quickly preprocess aligned RNA-seq samples and provide a list of RNA variants, a brand-new version of REDItools (REDItools v3) was used. In particular, REDItools v3 was able to speed up the pileup step, the bottleneck of the entire pipeline, by an order of magnitude allowing the traversing and analysis of input BAM files in minutes (depending on the sequencing throughput).The A549 dataset included three wild-types (WT) and three ADAR-silenced (SI) samples (used as the negative control). We extracted only sites with A-to-G evidence in WT samples and no A-to-G evidence in SI samples. In particular, sites with A-to-G substitutions detected in WT samples only were used as putative positive sites. In contrast, sites with A-to-G substitutions detected in both WT and SI as well as WGS were used as putative negative sites.On the whole, the A549 dataset comprised more than 13,000 sites (Table 2.1), 6,286 putative positive sites, and 6,856 putative negative sites (Table 2.1). When the model was applied to this dataset, it was able to correctly classify the majority of sites with a low percentage of false positives and a much lower percentage of false negatives, 3.43% and 0.46%, respectively (Figure 19). It achieved an accuracy of $>$ 96%, a precision of $>$ 93%, a recall of 99%, an F1-score of $>$ 96%, a specificity of $>$ 93%, and a G-mean of $>$ 96% (Table 2.2). For the HEK293T cell line, two different RNA-seq datasets were analyzed. The first one included eighteen in-house high throughput stranded paired-end RNAseq experiments containing on average 437.4 million reads per sample. Of these, three samples were WT, three ADAR KO, and three overexpressed ADAR (OE). Two distinct sets of grand truth RNA editing sites were extracted (Table 2.1). The first set, KO-WT, was obtained by comparing KO and WT samples yielding more than 30000 bona fide sites (Table 2.1), with 20,806 putative positive sites, and 9589 putative negative sites (Table 2.1). The second set, KO-OE, was created by contrasting KO and OE samples returning more than 45188 bona fide sites (Table S1) with 35,436 putative positive sites and 9752 putative negative sites (Table 2.1). When the model was applied to these HEK293T datasets, it was able to correctly classify the vast majority of sites with a low percentage of both false positives (2.17% and 1.45% for WT-KO and OE-KO respectively, as shown in Figure 2.7) and false negatives (1.32% and 1.58% for WT-KO and OE-KO respectively, as shown in Figure 2.7). Regarding the overall performance on the WT-KO dataset, the model achieved an accuracy of 95.6%, a precision of 96.9%, a recall of 98.1%, an F1-score of 97.5%, a specificity of 93.1% of specificity and a G-mean of 95.6% (Table 2.2). On the OE-KO dataset, instead, the model reached an accuracy of 95.6%, a precision of 98.1%, a recall of 98%, an F1-score of 98.1%, a specificity of 93.2% and a G-mean of 95.6% (Table 2.2).
The second HEK293T dataset contained public stranded paired-end RNA-seq experiments

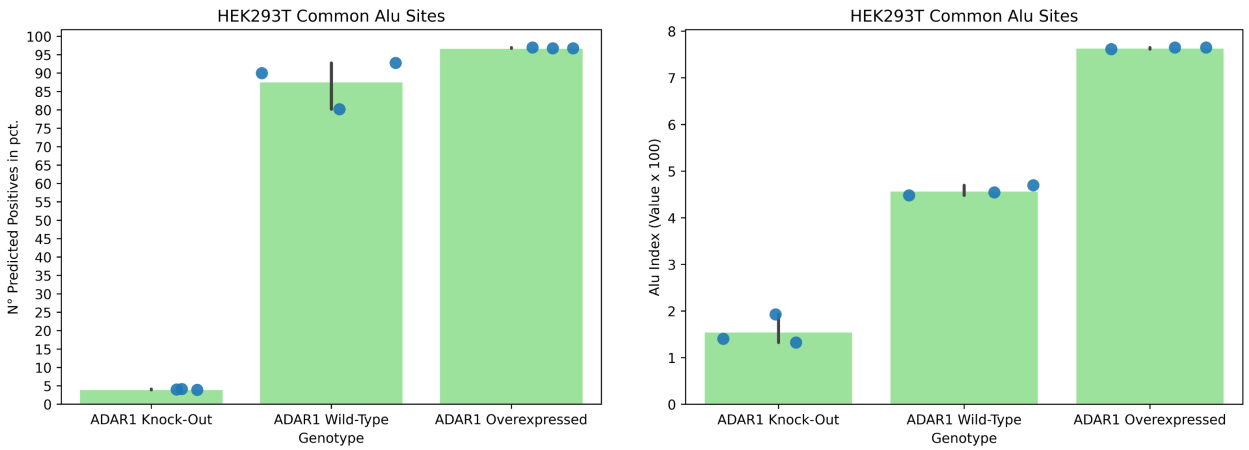

Figure 2.8: In-house HEK293T dataset' fraction of common Alu repeats adenosines predicted as putative A-to-I RNA editing sites and corresponding AEI-like score.

downloaded from the SRA database and included three WT and three ADAR KO samples sequenced at a lower throughput (about 69.5 million reads on average per sample). In this case, only 13,000 bona fide sites, with 6,919 putative positive sites and 6,541 putative negative sites, were extracted (Table 2.1). The model classified correctly the plurality of sites with low percentages of false positives and false negatives, 3.72% and 1.03% respectively (Figure 2.7). Regarding the computed metrics, the model achieved an accuracy of $> 95\%$, a precision of $> 93\%$, a recall of 98%, an F1-score of $> 95\%$, a specificity of $> 92\%$, and a G-mean of $> 95\%$ (Table 2.2). The last independent dataset comprised public U87 unstranded paired-end RNA-seq experiments downloaded from the SRA database and employed also as an independent dataset for evaluating the performance of DeepRed [[138]]. The U87 dataset included 2 WT and 2 ADAR KO low coverage samples, which yielded less than 3,000 bona fide sites, with 1,622 putative positive sites, and 1,254 putative negative sites (Table 2.1). Also in this case the model classified correctly the majority of sites, with a low percentage of false positives and a very low percentage of false negatives, 1.50% and 0.83% respectively (Figure 2.7). It also achieved an accuracy of $> 97\%$, a precision of $> 97\%$, a recall of $> 98\%$, an F1-score $>> 97\%$, a specificity of $> 96\%$, and a G-mean of $> 97\%$. Since the vast majority of A-to-I editing occurs in Alu repeats, and it is commonly used to quantify the activity of endogenous ADARs [[118]], the model behavior was tested on about 4,000 adenines falling in Alu sequences supported by all RNAseq reads from our in house HEK293T dataset (three WT samples, three KO samples, and three OE samples). The model predicted about 5% of them as putative A-to-I RNA editing events in the KO samples (Figure 2.8), about 87% of them as putative A-to-I RNA editing events in the WT samples (Figure 2.8), and about 95% of them as putative A-to-I RNA editing events in the OE samples (Figure 2.8). A-to-I RNA

editing events predicted in the KO samples can be primarily ascribed to residual ADAR activity. On the whole, our results prove the model's efficacy in recognizing information embedded in the RNA-seq data. Indeed, the percentage of predicted positives increased with the ADAR expression. Additionally, an AEI-like (Eq. 1.1) score [[128]] was calculated using the same adenosines (about 4,000) previously mentioned (Figure 2.8). As expected, the AEI-like scores increased with the ADAR expression.

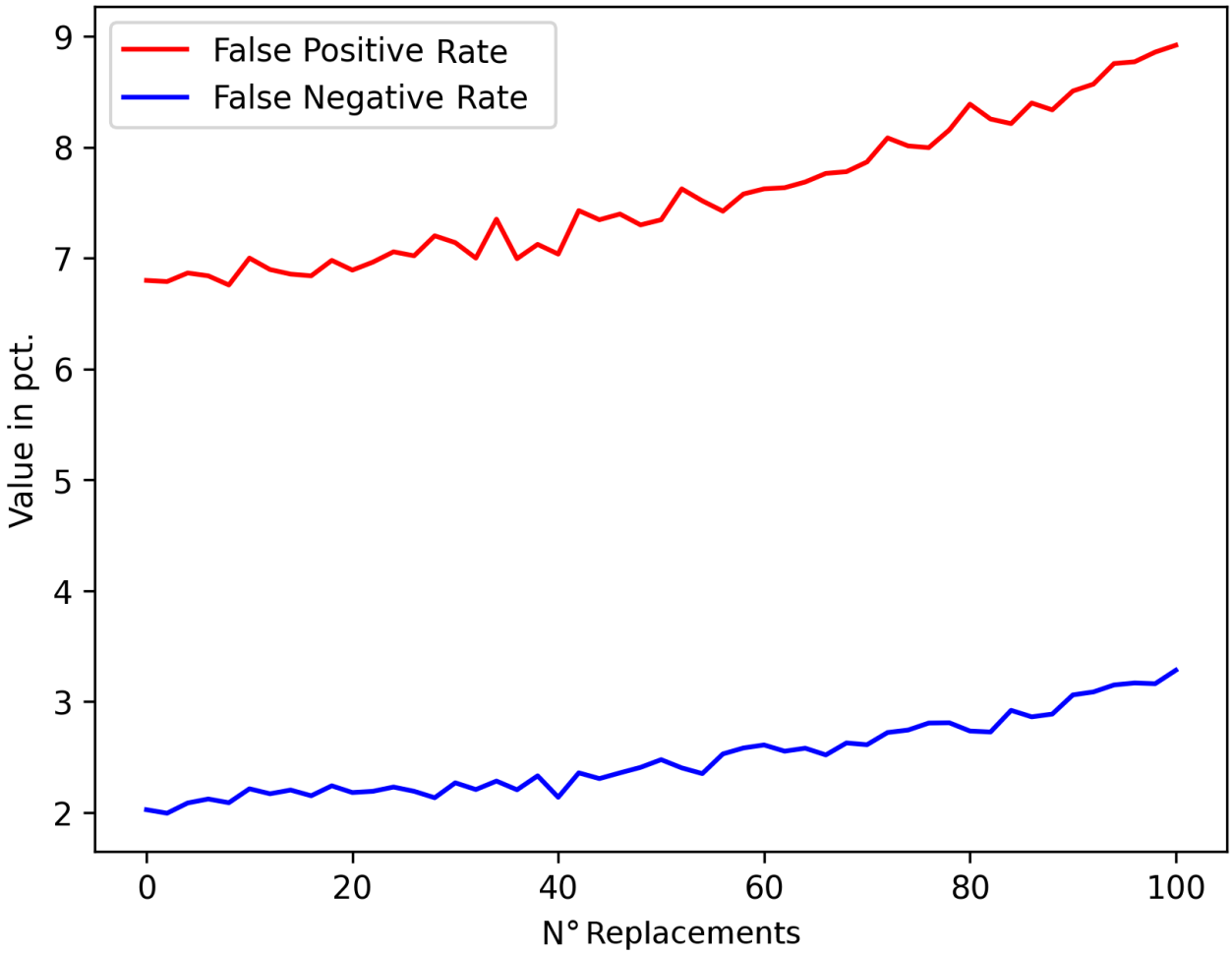

Figure 2.9: Proposed TCN false positive and false negative rate calculated in missing data simulations

The model predicts A-to-I events using windows of 101 bases centered at the investigated site. Such windows could not be fully covered, challenging the RNA editing detection. To check the model robustness in case of windows not fully covered, we used our in-house HEK293T KO-WT dataset and simulated missing data in the 101 nucleotides windows by a random replacement of different percentages of the RNAseq data with the corresponding genomic data. The model was able to fully tolerate missing data up to 30% of the nucleotide window, being practically unaffected in its capacity to identify A-to-I RNA editing events (

Figure 2.9). For missing data of $> 30\%$, the model was still effective in correctly predicting A-to-I RNA editing events with a maximum value of false positive rate $< 9\%$ and a maximum value of false negative rate $< 3.5\%$, when 100% of bases were replaced (Figure 2.9). It is worth mentioning that the nucleotide window was chosen based on several previous and independent studies, demonstrating that a length of 101 bases, centered on the investigated site, was the optimal one for capturing the RNA editing signal [[134], [141]].

2.3 Performance Comparison With Other ML and DL Tools

To further assess the model efficacy in detecting RNA editing events, its performance was compared with those of publicly available ML or DL, using ROC-AUC curves (Figure 2.10) and the set of bona fide sites extracted from our in-house HEK293T WT-KO dataset. In particular, the model was compared to EditPredict [[134]] and RED-ML [[141]].

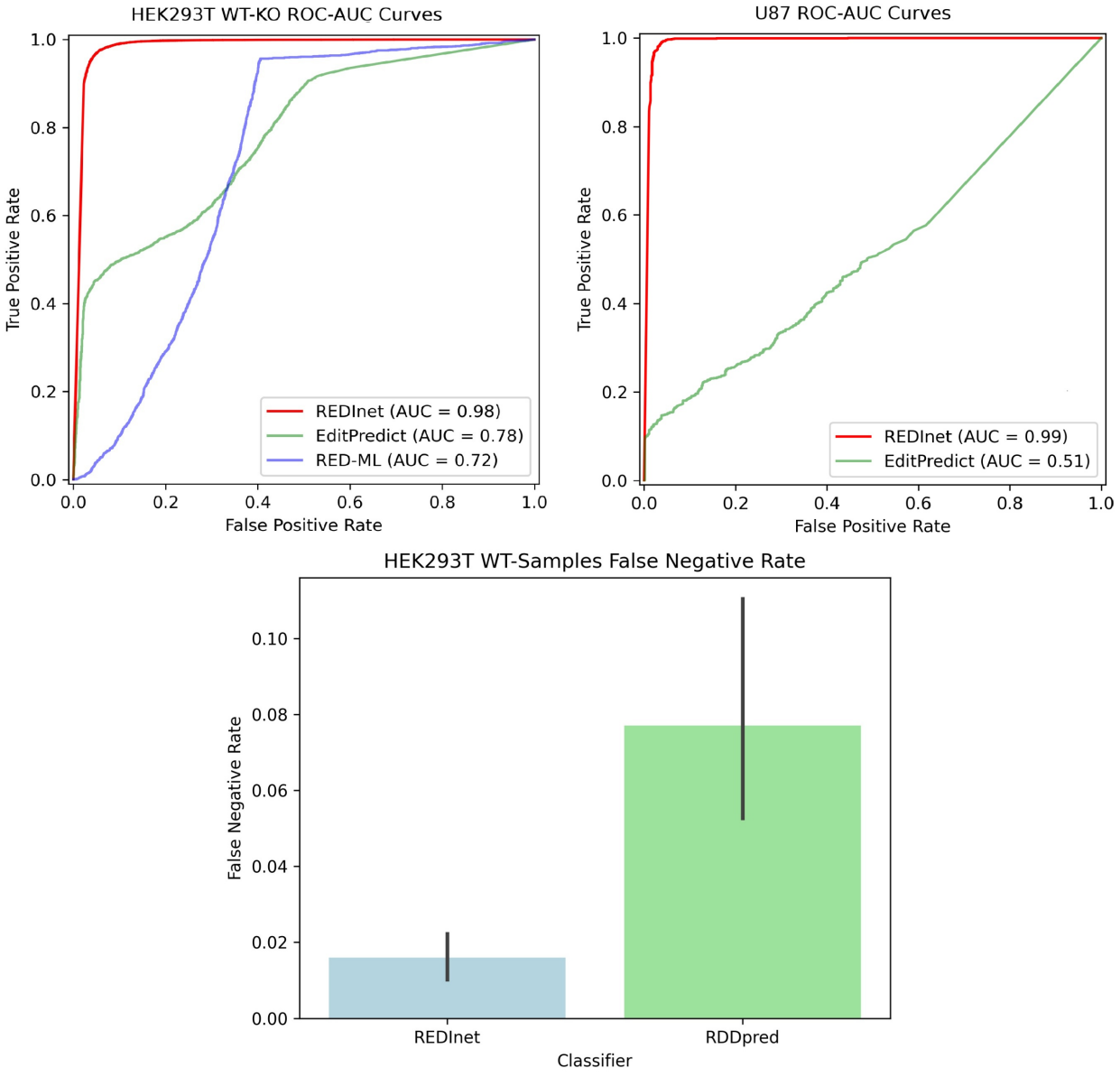

Figure 2.10: Proposed TCN performance comparison with other publicly available tools, by means of ROC-AUC curves and false negative rates.

EditPredict employs a CNN-based neural architecture trained on sequence context of 101 bases centered at REDIportal v1 sites [[86]] and tested on Jurkat cell line RNA-seq datasets. RED-ML, instead, is a logistic regression-based classifier trained on three broad classes of features derived from experimentally validated events. In the comparison (Figure 2.10), our model achieved the highest AUC score (0.98), followed by EditPredict (0.78) and RED-ML (0.72). A further comparison between our model and EditPredict was carried out using the public U87 dataset consisting of over 2800 bona fide and well-covered sites. In this case, our model achieved an AUC score of 0.99 while EditPredict achieved an AUC score of 0.51 (Figure 2.10). Additionally, our model was compared with RDDpred [[139]], an RF-based classifier employing 15 features reflecting the read-alignment patterns, specialized in distinguishing sequencing artifacts from genuine AG substitutions [[139]], trained on sites from the DARNED [[181]] and RADAR [[85]] databases. We compared our predictions with those from RDDpred on our HE293T WT-KO bona fide sites. In order not to misuse RDDpred, considering that this tool is expressly designed to distinguish editing events from sequencing errors, and not from genomic sources of variation, it was compared to the proposed TCN only on bona fide positive sites, evaluating them in terms of false negative rate. RDDpred achieved a false negative rate of about 7%, while the proposed TCN obtained a false negative rate of less than 2% (Figure 2.10). Unfortunately, a further comparison with DeepRed [[138]], an MLP stacked ensemble-based tool, was not possible due to installation issues and software dependencies. It should be noted that we did not compare the performance of the proposed TCN and DeepRed [[138]], on the U87 independent dataset, as it was not possible to retrieve the genomic coordinates of the 48 sites (47 positive and 1 negative), used in the evaluation of DeepRed [[138]].

2.4 Limitations And Future Development

Human A-to-I RNA editing events are generally classified into three groups using RepeatMasker annotation [[168], [124], [175], [86]]:

- ALU:

  If A-to-I RNA editing events occur inside Alu regions.

- REP:

  If A-to-I RNA editing events occur inside repetitive non-Alu regions.

- NON REP:

  If A-to-I RNA editing events occur inside non-repetitive regions.

Our model does not take into account these categories to avoid drawbacks due to unbalanced training sets largely rich in ALU sites, which represent >95% of the known A-to-I RNA editing events. Nonetheless, our model is quite flexible and could be easily extended to incorporate these features. Additionally, our model requires pre-computed pileup tables from RNA-seq-aligned reads. This step is computationally intensive and time-consuming [[182]], and could hinder model usage in RNA-seq experiments with high sequencing depth. This potential disadvantage was avoided using REDItools v3, a novel re-implementation of the REDItools algorithm that is an order of magnitude faster than the previous version. For example, on the HEK293T KO BAM files comprising on average 277 million reads, REDItools v3 profiled variant and invariant sites in less than 60 minutes per file. In contrast, REDItools v1 took about 9 hours per sample to complete the same jobs. On the whole, our model showed a great generalization capability, achieving high performance on all independent datasets from immortalized cancer cell lines and outperforming other ML/DL tools. Nonetheless, model performances could change depending on different factors such as the sequencing depth, the library used to produce RNA-seq reads (stranded or unstranded, single or paired-end), or the sequencing technology. Our model is currently optimized to predict A-to-I events in Illumina-derived RNA-seq data and it has been trained on human data only. We plan to extend the model to other organisms and sequencing technologies such as PacBio or Oxford Nanopore able to produce long reads. Additionally, we plan to train the model using RNA editing data from The Cancer Genome Atlas Program (TCGA) already included in the REDIportal database, enabling the identification of cancer-specific A-to-I events and to better investigate RNA editing dynamics in cancer cells. Our model can also provide highly

valuable insights to understand sequence editability by ADARs and to design antisense oligonucleotides for emerging site-directed A-to-I RNA editing technologies for correcting disease-causing point mutations [[77], [183], [184], [185]].

Chapter 3

Conclusions

In this thesis, it has been successfully demonstrated that A-to-I RNA editing events can be accurately identified using only RNA-seq base counts, encoded and normalized like audio frequencies, without any prior knowledge of genomic variants. It has also been successfully demonstrated that genomic regions spanning 101 nucleotides, centered in the investigated sites, carry enough information to effectively catch the A-to-I RNA editing dynamicity taking into account RNA motifs specific for edited and not edited transcripts. Harnessing 8404 RNA-seq data (Figure 2.1) from the GTEx project, a great number of unique examples (about 12.5 million) has been extracted, including putative positive A-to-I RNA editing sites and putative negative sites (Table 2.1). These examples have been collected using the REDItools protocol and a new implementation of REDItools. The unique size of the dataset, which to date is the biggest by at least an order of magnitude among the ones used to train A-to-I RNA editing classifiers, is further proved by collected from the same REDItools pileup tables used to populate the REDIportal database in its latest version [[174]], guarantees, by means of its size, a sufficient coverage of the A-to-I RNA editing portion of the human epitranscriptome, in terms of site specific A-to-I RNA editing (ALU, REP NON ALU and NON REP sites), even though the Poly(A) RNA-seq is on purpose biased toward mRNAs, so that the trained classifier can achieve an high generalizing capability without overlooking REP NON ALU and NON REP sites. As main objects of this thesis it has been demonstrated on one hand that the apparent similarity between RNA-seq base counts and raw audio data, as well as the one between ADARs specificity for particular motifs on dsRNAs and temporal sequences, can be effectively exploited to analyze RNA-seq data as if it were temporal sequences, and on the other hand that TCN-based architectures are efficient, low demanding, relatively simple and powerful tools to analyze temporal sequences; as it has been clearly observed in the other works mentioned in this study. In this regard it has been shown that the use of TCN-based architectures guarantees large receptive fields, in the proposed archi-

tecture more than six times the width of the temporal dimension, increasing the depth of the neural network, 149 ones in the proposed architecture, but maintaining a relatively limited number of parameters, about 1.6 million trainable and non-trainable ones in the proposed architecture. In particular it has been proved that the specific TCN architecture proposed here (Figure 17) accurately predicted putative positive and negative A-to-I RNA editing events in both the validation and test set, as clearly shown in the respective confusion matrices (Figure 2.6) and computed metrics (Table 2.2). It has also been demonstrated that the adopted regularization and network stabilization methods (L2 regularization and Batch Normalization), contributed to achieve an high generalization capability, as shown by the proposed TCN performance on the test test and on the independent datasets, by means of avoiding overfitting and underfitting. In this regard the proposed TCN-based binary classifier achieved also high performance on the four independent datasets as shown by the respective confusion matrices (Figure 2.7) and computed metrics (Table 2.2), even though they were collected from RNA-seq sample obtained on immortalized cancer cell lines, instead of healthy human tissues, with different Illumina sequencing protocols; single-end Poly(A) unstranded RNA-seq in the case of training validation and test sets, unstranded paired-end RNA-seq for the U87 dataset, stranded paired-end RNA-seq for the HEK293T datasets (both in house and public) and A549 dataset. This proved the efficacy of the proposed feature selection and feature encoding methods that allow the proposed TCN architecture to extrapolate the relevant characteristic of edited and non edited transcripts, invariant with respect to both the RNA-seq protocols and the biological samples origins. The analyses conducted on the about 2000 common ALU sites among all the in house HEK293T samples (Figure 2.8) proved the trained TCN operates accordingly to the specific biological context intrinsic to the RNA-seq data, indeed the proposed TCN yielded prediction that reflect both the increasing ADAR1 expression levels and the corresponding increase in ADAR1 enzymatic activity.The analysis on potential gaps in the 101 nucleotides windows proved the overall robustness of the trained TNC and its clear potential in real case application (Figure 2.9), given a maximum false negative rate of about 3% and a maximum false positive rate of about 9% when all but the central nucleotide RNA-seq base count are replaced by means of imputations.using genomic information. The TCN performance comparison with the ones yielded by existing tools on the in-house HEK293T KO-WT set of ground truth sites proved the proposed classifier outperformed the others, indeed the proposed TCN achieved an AUC score of 98%; in contrast to EditPredict AUC score of 78% and RED-ML AUC score of 72% (Figure 2.10). The TCN better performance was further proved by its AUC score of 99% on the U87 dataset, in contrast to EditPredict AUC score of 51% (Figure 2.10), and its false negative rate $< 2\%$ on the in house HEK293T KO-WT dataset, in contrast to RDDpred false negative rate of

about 8% (Figure 2.10). Given the proposed TCN-based classifier overall performances it has been integrated in the REDIportal ecosystem as a tool to calculate aggregated P-values (Eq. 4.23) for A-to-I RNA editing sites over each one of the 55 GTEx body sites, opening the way to use it on large scale application in conjunction with the new ultrafast version of REDItools proposed in this study. Given the TCN architecture flexibility several further improvements of the entire computational pipeline could be made to address its current limitation, for instance preliminary results, not reported in this study, suggest the 101 nucleotides windows could be reduced without affecting the classifier performance, drastically extending its applicability given we found eventual gaps are most likely located at the ends of the 101 nucleotides windows. The proposed TCN could also be extended to use specific sites annotation or transcripts secondary structures so that additional fundamental information is taken into consideration. In the end it can be said that in this study it has been successfully implemented a novel and effective A-to-I RNA editing TCN-based classifier, specific for human NGS RNA-seq data, trained, validated and tested on the largest available collection of A-to-I RNA editing events from the REDIportal database, which can discriminate with high accuracy genuine A-to-I RNA editing events.

Chapter 4

Materials and Methods

4.1 HEK293T cell lines, library preparation, and Illumina sequencing

HEK293T (Human Embryonic Kidney) cells (a kind gift of Dr. Marco Binder, DKFZ, Heidelberg, ATCC, Cat# CRL-3216) were maintained in high-glucose DMEM (Sigma-Aldrich, Cat# D6429) supplemented with 10% FBS (PAN Biotech, Cat# P40-37100) and 1% penicillin/streptomycin (Sigma-Aldrich, Cat# P4333) in 5% CO2 at 37°C. The ADAR1 knockout cell line was generated according to the protocol by Pecori et al. [[186]]. In order to generate the ADAR1 overexpressing cell line, the coding sequence of ADAR1 p110 isoform was amplified by RT-PCR (Qiagen, Cat# 210212) using the following primers:

- F primer:

  attcaagcttgggccaccggctagcATGGCCGAGATCAAGGAG.

- R primer:

  gggggggagggagaggggcagatctCTATACTGGGCAGAGATAAAAG.

The ADAR1 p110 isoform coding sequence was cloned inside the AID-express-puro2 plasmid [[187]] digested with NheI and BglII. Thus replacing the coding sequence of Activation Induced Deaminase (AID) and having a GFP as a reporter for ADAR1 p110 exogenous expression. The plasmid was later transfected using Lipofectamine 2000 (ThermoFisher, Cat# 11668019), according to the manufacturer's instructions, into HEK293T cells; transfections were carried out as three independent biological replicates. Cells were later selected, after

24 h from the transfection, using puromycin (1.5 µg/ml). During selection cells were sorted in bulk by the highest level of GFP (and thus ADAR1 p110). ADAR1 p110 overexpression was validated by western blot (Cell Signaling Technology, Cat# 14175) using GAPDH as a loading control (Cell Signaling Technology, Cat# 2118). Total RNA was firstly extracted and purified using a Qiagen RNeasy Plus kit (Qiagen, Cat# 74134) and secondly treated with DNase (Invitrogen, Cat# AM1907). The Illumina libraries were prepared, according to the manufacturer's protocol, starting from 500 ng of total RNA extracted, using Illumina's TruSeq Stranded Total RNA Sample Preparation Kit (Illumina, San Diego, CA, USA). Later the produced cDNA libraries were checked on the Bioanalyzer 2100 and quantified by fluorimetry using the Quant-iTTM PicoGreen® dsDNA Assay Kit (Thermo Fisher Scientific) on NanoDrop™ 3300 Fluorospectrometer (Thermo Fisher Scientific). Illumina sequencing was performed on a NovaSeq 6000 platform using the paired-end approach (2×100 bp) with 262–638 million reads per sample.

4.2 Feature Selection and Extraction -Training, Validation and Test Sets

The training, validation and test sets were built using 8404 human RNA-seq healthy tissue samples and their corresponding WGS/WES data (if any) collected from the GTEx [[78]] project database. Positive A-to-I RNA editing events were collected from transcriptomic and genomic reads processed according to the REDItools protocol [[124]] (Figure 4.1), summarized in the following steps:

1. Download of RNA-seq experiments and WGS/WES experiments in fastq format from the SRA database.

2. RNA-seq reads quality checks by FastQC (Table 4.1) software to spot potential problems or biases in the data.

3. RNA-seq read trimming by fastp (Table 4.1)software to remove their low quality regions, setting a minimum base phred quality (Eq. 1.2) of 25, a maximum unqualified bases fraction of 10% and minimum read length of 50 bases as well as ruling out also low complexity and polyX (region of multiple repetitions of the same base) in 3' ends.

4. Creation of reference genome index for STAR aligner software (Table 4.1) setting the length of the genomic sequence surrounding the annotated junction to 75 bases.

5. RNA-seq reads alignment to the reference genome by STAR aligner software (Table 4.1) setting a maximum loci number a read could be mapped to of 1 (only unique reads are retained), a minimum splicing junction overhang of 8 bases, a maximum number of mismatches per reads pair of 999 (such high number practically switch off this filter), a maximum mismatches fraction of 0.04 and a minimum intron length of 20 bases as well as ruling out reads that do not contain splicing junction present in the star SJ.out.tab file.

6. RNA-seq reads strand orientation inference by infer_experiment.py script from the RSeQC package (Table 4.1), which compares the reads orientations with known annotated transcripts ones.

7. Creation of reference genome index for BWA mem aligner software (Table 4.1).

8. WGS/WES reads alignment to the reference genome by BWA mem aligner software (Table 4.1).

9. Pileup, by means of REDItoolDnaRNA.py (Table 4.1) script from the REDItools package, of DNA-RNA alignments' output BAM files setting a minimum mapping quality score (measure of the estimated probability a read has been misaligned) of 30 for RNA-seq reads and 255 for WGS/WES reads, a minimum number of RNA-seq reads supporting the variation of 1 and minimum base quality score of 30 for both RNA-seq and WGS/WES reads as well as ruling out homopolymeric regions, ruling out non paired and concordant RNA-seq reads, using confidence values to calculate the strand orientation and operating reads strand correction.

10. Selection of potential DNA-RNA variants and exclusion of position supported by < 10 WGS/WES reads.

11. Annotation of potential DNA-RNA variants by means of RepeatMasker and dbSNP annotation files using AnnotateTable.py (Table 4.1) script from REDItools package.

12. Selection of ALU sites supported by ≥ 5 RNA-seq reads excluding sites with evidence of genomic source of variation from REDItoolDnaRNA.py (Table 4.1) output tables.

13. Selection of non REP NON ALU sites supported by ≥ 10 RNA-seq reads, excluding sites in Simple repeats, in Low complexity regions and those with evidence of genomic source of variation from REDItoolDnaRNA.py (Table 4.1) output tables.

14. Select NON REP sites supported by ≥ 10 RNA-seq reads, excluding those with evidence of genomic source of variation from REDItoolDnaRNA.py (Table 4.1) output tables.

15. Annotation of ALU, REP NON ALU and NON REP sites using known A-to-I RNA editing events.

16. Extraction of known A-to-I RNA editing events from REDItoolDnaRNA.py (Table 4.1) output tables.

17. Extraction of candidates ALU sites from from REDItoolDnaRNA.py (Table 4.1) output tables in GFF format.

18. Extraction of candidates REP NON ALU and NON REP sites from from REDItoolDnaRNA.py (Table 4.1) output tables in GFF format.

19. Candidates ALU sites filtering by REDItoolDnaRna.py script (Table 4.1) to identify novel potential A-to-I RNA editing events using a much more stringent setting, i.e increasing the minimum RNA-seq mapping quality to 255, minimum homopolymeric region length to 5 bases for both RNA-seq reads and WGS/WES reads, as well as trimming 11 bases in RNA-seq reads 5' end and 6 bases in RNA-seq reads 3' ends, removing substitutions in RNA-seq reads homopolymeric regions and excluding multi hits in RNA-seq.

20. Candidates REP NON ALU and NON REP sites filtering by REDItoolDnaRna.py (Table 4.1) script to identify novel potential A-to-I RNA editing events using the same stringent setting adopted for candidates ALU sites adding a minimum number of RNA-seq reads supporting the variation of 3 and a minimum A to G substitution frequency of 0.1.

21. Alignment of RNA-seq reads harboring candidates REP NON ALU and NON REP sites to a reference genome by PBLAT (Table 4.1)software to detect and exclude multi-mapping reads as well as PCR duplicates.

RNA-seq reads were aligned to the reference genome using the splice-aware (alignment considering the loss of intron sequences due to alternative splicing) STAR aligner given its accuracy and high speed in performing alignments. WGS/WES reads were aligned to the reference genome, using BWA-MEM which is a fast and accurate aligner, based on the general Burrows-Wheeler Aligner algorithm, well suited for low-divergent sequence mapping to a reference genome. The positive A-to-I RNA editing events collected, as well as the pileup table produced by REDItoolDnaRNA.py script, were previously used to populate the REDIportal database [[168]] in its current version [[174]].

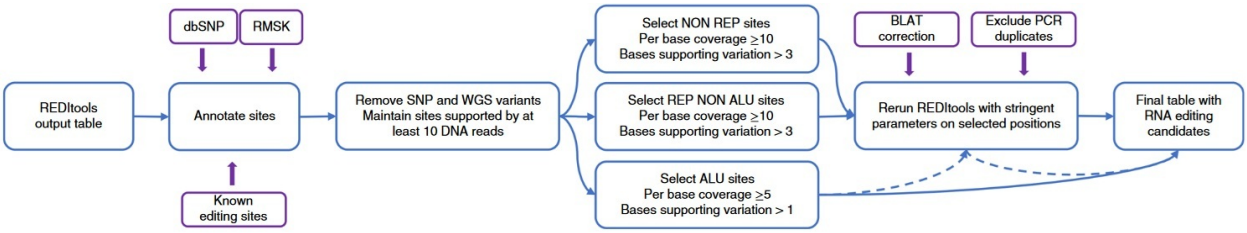

Figure 4.1: REDItools protocol workflow. Talen and modified from Lo Giudice C, Tangaro MA, Pesole G, Picardi E. Nat Protoc 2020, vol 15, pp.1098–1131.

Each REDItools output table was indexed by means of samtools tabix, for quick access lines corresponding to genomic positions of interest in the following phases. For each REDItools output table, all sites with A-to-G evidence in RNAseq, no base change in DNAseq (WGS or WES data), and present in the REDIportal [[168]] database version were selected as positive A-to-I RNA editing sites. Negative A-to-I RNA editing sites were collected from the same pileup tables used for the positive ones, including all positions with A-to-G evidence in both RNAseq and DNAseq data to capture RNA variants not due to RNA editing; unlike the positive ones that have A-to-G evidence in RNA-seq read only and no observed base substitution in the DNA-seq (WGS or WES) data. Negative sites were further filtered, excluding sites supported by < 50 RNA-seq reads and < 10 DNA-seq reads. To reduce the noise caused by genomic sources of variation in the training data, low-frequency SNPs were ruled out by retained only negative sites with an RNA-seq A-to-G rate ≥ 0.4 and a comparable DNAseq A-to-G rate; diverging from the RNAseq A-to-G rate less than 5%. To provide examples as close as possible to eventual Illumina sequencing base substitution errors, which tends to overlap to genuine A-to-I RNA editing events in ALU sites, a pool of random sites, belonging to NON REP regions and having A-to-G substitution rates < 0.01, was collected from the pileup tables derived from GTEx RNA-seq experiment on skeletal muscle; which is known to be a tissue little affected by A-to-I RNA editing. Sites with only A-to-G evidence in $< 30\%$ of the skeletal muscle samples were selected as a second class of negative examples approximating potential Illumina sequencing base substitution errors. For each positive and negative site, flanking regions of 50 bases upstream and 50 bases downstream were extracted both from the reference genome and the pileup tables base counts.

All sites with flanking regions fully covered by RNAseq reads were maintained in the final dataset of positives and negatives, instead regions with gaps were discarded. In the end each positive or negative site is represented by two sets of features:

- Base sequence of a gap free human genomic region of 101 nucleotides with center in the selected site.

- Sequence of 101 RNA-seq base counts, corresponding to the selected genomic region.

	Package	Version	Dependencies
FastQC	FastQC	0.11.8	OpenJDK v11.0.1
fastp	fastp	0.20.0	libdeflate v1.0
STAR	STAR	2.7.0f	gcc v11.5.0, glibc v2.34
RSeQC	RSeQC	2.6.4	bx-python v0.8.2, pysam v0.15.2
BWA-mem	BWA	0.7.17	zlib v1.2.12
REDItoolDnaRNA.py	REDItools v1	1.3	htslib v1.9, pysam v0.15.2
AnnotateTable.py	REDItools v1	1.0	htslib v1.9, pysam v0.15.2
PBLAT	PBLAT	2.1	htslib v1.9, zlib v1.2.12

Table 4.1: REDItools protocol required software.

4.3 Independent Datasets Pileup With REDItools v3

The independent datasets RNA-seq reads pileup (point n° 7 in the summarized REDItools protocol listed above) was carried out using a new version of REDItools [[188]] available at https://github.com/BioinfoUNIBA/REDItools3. This new version consists in an improved implementation of the original algorithm by means of new C libraries made available through PySAM [[189], [190]], dynamic implementation of REDItools' features and migration from python2 to python3; speeding up the BAM files analysis by an order of magnitude compared to REDItools v1. The dynamic implementation allows only necessary parts of code to be loaded and implemented according to the input provided. Additional improvements in the coding logic and loop structures represent further sources of speed up in REDItools v3. Like the previous versions, REDItools v3 has numerous options and features to customize each editing analysis according to specific needs of the users and it's now streamlined for parallel processing.

4.4 Feature Selection and Extraction - Independent Datasets

Three WT and three ADAR silenced (SI) stranded and paired-end Illumina RNA-seq experiment were downloaded from the SRA database in fastq format, (accessions: SRX8983709, SRX8983710, SRX8983711, SRX8983718, SRX8983719, SRX8983720) as well as the A549 WGS dataset (accession SRX5437560) also in fastq format. Transcriptome reads were quality-checked by means of FastQC software, (point n°2 of the described REDItools protocol), trimmed with fastp software (point n°3 of the described REDItools protocol) and then aligned onto the reference human genome assembly hg38 (point n°4 of the described REDItools protocol) by STAR (version 2.7.0f using GENCODE v.46 during the genome indexing). WGS reads, instead, were aligned onto the same reference human genome assembly hg38 (point n°6 of the described REDItools protocol) by BWA mem (version 0.7.17). Reads pileup and RNA variant calling were carried out using REDItools v3 with no stringent parameters (point n°7 of the described REDItools protocol). Candidates positive and negative A-to-I RNA editing events were selected from the pileup tables similarly to previously done on the GTEx pileup tables. In detail all sites showing A-to-G evidence in WT data but not in SI and WGS data were marked as positives (thus using SI samples and WGS as negative controls for putative positive A-to-I RNA editing events and as positive controls for genomic sources of variation), instead, sites with A-to-G evidence in WT, SI, and WGS data, were marked as negatives, ruling out sites having depth $\leq$ of 50 for RNA-seq, depth $\leq$ 10 for WGS, not supported by at least 3 A-to-G substitutions and with a A-to-G substitution rate $\leq$ 0.01. In addition, negative sites with no A-to-G evidence in all samples were retrieved. Both positive and negative candidates were later annotated using a python3-compliant version of the AnnotateTable.py script from the REDItools [[124], [175]] package by means of the hg38 specific version of RepeatMasker annotation. All sites falling in repetitive regions were ruled in, instead sites in non-repetitive regions were ruled in only if present in the RefSeq annotations. The same procedure was also applied to extract positive and negative candidates from the in house RNA-seq data of WT, KO, and OE HEK293T cell lines, using WGS data from HEK293T downloaded from SRA (accession SRX5437560), and finally generating two ground truth sets of sites; one using WT (KO-WT) samples and the other using OE (KO-OE) ones in place of WT. In detail the KO-WT dataset includes putative positives and negatives A-to-I RNA editing events contrasting pairs of KO (used like the SI samples in the A549 dataset as negative control for A-to-I RNA editing events and as positive control for genomic sources of variation) and WT samples, and the KO-OE dataset obtained comparing pairs of

KO and OE samples; as well as contrasting the RNA-seq samples with the WGS sample used also in this case as a negative control for A-to-I RNA editing events and as a positive control for genomic sources of variation. The same procedure followed for the A549 samples and the in house HE293T samples was carried out to establish the ground truth set of sites for the public HEK293T and U87 samples, with the exception of the adoption of less stringent filtering step regarding the RNA-seq depth, ruling out sites with depth leq 30 instead of 50. Public HEK293T and U87 data, were also downloaded from the SRA database (HEK293T accessions: SRX2825941, SRX2825942, SRX2825943, SRX2825944, SRX2825945, SRX2825949; U87 accessions: SRX110671, SRX110672, SRX110672, SRX110674), as well as the WGS for U87 (accession: SRX5466662). Finally for each one of the four independent datasets the feature extraction of the ground truth sites was carried out following the same procedure adopted for training, validation and test sets.

4.5 Features Encoding

Each one of the extracted genomic regions of 101 bases, centered in the putative positive or negative A-to-I RNA editing site, was encoded as a 4x101 matrix. This matrix was obtained by concatenating one-hot encoding column vectors, each vector encoding one of the 101 bases. In detail the bases one-hot encoding is always described by the following set of equations:

$$A_{OHE} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, C_{OHE} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix},$$
$$G_{OHE} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, T_{OHE} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad (4.1, 4.2, 4.3, 4.4)$$

For instance, according to the above mentioned method (Eq. 4.1, Eq. 4.2, Eq. 4.3, Eq. 4.4), an hypothetical sequence of 4 bases TCGA is encoded by the following 4x4 square matrix:

$$TCAG_{OHE} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad (4.5)$$

Each one of the extracted 101 sequences of 101 RNA-seq base counts, centered in the investigated site RNA-seq base counts, are also encoded as 4x101 matrices. This matrix is obtained by concatenating the RNA-seq base count column vectors, each vector reporting the corresponding RNA-seq base count observed for one of the bases extracted previously.

For instance assuming for the above mentioned TCAG base sequence the observed RNA-seq base counts are the following:

$$T_{RBC} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 40 \end{bmatrix}, C_{RBC} = \begin{bmatrix} 0 \\ 35 \\ 0 \\ 15 \end{bmatrix},$$
$$A_{RBC} = \begin{bmatrix} 30 \\ 0 \\ 10 \\ 0 \end{bmatrix}, G_{RBC} = \begin{bmatrix} 0 \\ 0 \\ 40 \\ 0 \end{bmatrix} \quad (4.6, 4.7, 4.8, 4.9)$$

then, according to the previously described approach (Eq. 4.6, Eq. 4.7, Eq. 4.8, Eq. 4.9), the TCAG sequence corresponding RNA-seq base count matrix is the following 4x4 one:

$$TCAG_{RBC} = \begin{bmatrix} 0 & 0 & 30 & 0 \\ 0 & 35 & 0 & 0 \\ 0 & 0 & 10 & 40 \\ 40 & 15 & 0 & 0 \end{bmatrix} \quad (4.10)$$

Finally each putative positive and negative A-to-I RNA editing site is encoded by an 8x101 matrix obtained by concatenating the two 4x101 matrices previously constructed (Eq. 4.5, Eq. 4.10), so that the first four rows encode the base sequence in one-hot encoding format and the last four the RNA-seq data. For instance the final encoding matrix for the TCAG sequence is the following 8x4 one (Figure 2.1):

$$TCAG_{OHE-RBC} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 30 & 0 \\ 0 & 35 & 0 & 0 \\ 0 & 0 & 10 & 40 \\ 40 & 15 & 0 & 0 \end{bmatrix} \quad (4.11)$$

4.6 Features Preprocessing

As previously elucidated, each putative positive or negative A-to-I RNA editing site is encoded as an 8x101 a matrix of features M such as:

$$M = \left[r_{i,j} \right]_{0 \leq i \leq 7,\ 0 \leq j \leq 100} \tag{4.12}$$

where ri, j is the value of the feature i relative to base j. In order to feed the neural network with manageable feature values, the features derived from RNA-seq were firstly normalized between zero and one. In order to retain the specific RNA-seq information embedded in each one of the extracted base counts, without it being affected by the other ones in the 101 nucleotides window, the normalization was carried out for each base count separately according to the following equation:

$$r'_{i,j} = \frac{r_{i,j}}{\sum_{k=4}^{7} r_{k,j}} \quad \forall\, i,\ j \in [4,\ 7]\ [0,\ 100] \tag{4.13}$$

where r'i, j is the normalized value of ri, j. In order to avoid potential loss of information corresponding to low frequency substitutions, in particular the ones in ALU sites that often have a very small number of observed A-to-G substitutions, the normalized values have been later further preprocessed, first by transforming each one of them into its correspondent pseudocount binary logarithm according to the following equations:

$$\begin{cases} \alpha = 0.0001 \\ r''_{i,j} = \log_2(r'_{i,j} + \alpha) \end{cases} \quad \forall\, i,\ j \in [4,\ 7]\ [0,\ 100] \tag{4.14, 4.15}$$

where r''i, j it the pseudocounts binary logarithm associated with the normalized value r'i, j and α is the pseudocount constant, and second by rescaling the pseudocount binary logarithm according to the following equation:

$$r'''_{i,j} = \frac{r''_{i,j} - \log_2(\alpha)}{\log_2(1+\alpha) - \log_2(\alpha)} \quad \forall\, i,\ j \in [4,\ 7]\ [0,\ 100] \tag{4.16}$$

where r'''i, j is the rescaled value of r''i, j and α is the previously introduced pseudocount constant. This preprocessing method was adopted to reduce the difference between the scaled values of the high- and low-frequency substitutions, given the behavior of the logarithm function with arguments in the interval [0, 1] and α was conveniently chosen to be about one order of magnitude smaller than the minimum A-to-G substitution rate observed in putative A-to-I RNA editing events.

4.7 TCN Architecture

The TCN architecture is inspired by the WaveNet architecture [[144]], and it is, like WaveNet, based on the use of a sequence of ten Residual Units [[162], [167]] (Figure 2.5), preceded by a Conv1D layer (Figure 2.5). Each Residual Unit i processed by series of six Causal-Dilated Conv1D layers, with dilation rates varying according to powers of 2 from 1 to 32 (Figure 2.5); so to provide the TCN, along the Residual Units stack, a receptive field capable of covering the full 101 time steps temporal dimension (101 encoded nucleotides) more than six times. Each series of Causal-Dilated Conv1D layers is followed by a simplified version of WaveNet Gated Activation Unit [[144]] (Figure 2.5), which reproduces LSTM' Input Gate-like mechanism, i.e. the pairwise product between the product of a tanh activation layer output tensor and a sigmoid activation layer output tensors; which has shown to have higher capacity to capture nonlinearity in audio and audio-like signals with respect to ReLU activation function [[191]]. In The simplified Gated Activation Unit the input tensor of both the tanh activation layer and the sigmoid activation layer is the Causal-Dilated Conv1D layers output tensor; instead of one different input tensor for each different activation layer, obtained by means of one different additional Causal Conv1D layer before for each one of the Gated Activation Unit activation layers. Like WaveNet architecture the simplified Gated Activation Unit is followed by a 1x1 Causal Conv1D layer to perform a downsampling, according to TCN Residual Blocks general schematics [[149]], over the simplified Gated Activation Unit output (Figure 2.5), and unlike WaveNet Architecture the 1x1 Causal Conv1D layer if followed by Batch Normalization Layer [[192]] (Figure 2.5), to stabilize and further regularize each Residual Unit, and a ReLU activation layer, to insert nonlinearity before the Residual Connection (Figure 2.5). In each Residual Unit the Batch Normalization, put before the activation layer as it is usually done [[192]], output is directed outside the former via a Skip Connection (Figure 2.5), similarly to what is done in the WaveNet Architecture. Like WaveNet each Residual Unit has a Residual Connection (Figure 2.5) that sums the Residual Connection input tensor to the last ReLU activation layer output tensor to further stabilize the TCN like the standard TCNs Residual Units [[149]]. The ten Skip Connections output tensors, like WaveNet ones, are summed together, and the resulting tensor directed to a ReLU activation followed by two 1X1 Causal Conv1D layers in sequence, with an additional ReLU activation layer inserted between the latter (Figure 2.5). Unlike the WaveNet though, the second Causal Conv1D layer is not followed by a Softmax activation layer but by a ReLU activation layer (Figure 2.5), and its output tensor summed to the last Residual Unit output tensor (Figure 2.5). This resulting tensor is directed to a block of two additional Causal Conv1D Layers in sequence, each one of them

followed by a Max Pooling layer; to carry out a final downsampling. The last Max Pooling layer output tensor is finally fed to a sequence of four fully connected layers (Figure 2.5), which serve as a bottleneck to find the minimal data representation for the classification task. Each fully connected layer but the last is followed by a sequence of a ReLu activation Layer and a Batch Normalization layer (Figure 2.5), similarly to the what is done in the SE-Res2Block of the ECAPA-TDNN architecture [[193]]; since we found this provide additional stability to our TCN. Each fully connected layer but the last is regularized by means of L2 regularization to provide additional stability to the fully connected layers. Finally the last fully connected layer is followed by a Sigmoid activation layer to assign data output probabilities (Figure 2.5). Each one of the Dilated-Causal Conv1D layer of the proposed TCN architecture, is followed by a ReLU activation layer, instead of Linear activation layer, since we found this gives better performance in the specific task addressed in this thesis.

4.8 Fine Tuning and Training

The model hyperparameters tuning was carried out by means of ten-folds cross validation and grid search, to find the best combination of batch normalization parameters, max dilation rate, kernel width, kernel strides, pooling method, pooling size, number of output channels, optimization method, learning rate, batch size, weight regularization method and weight initialization method. At the end of the hyperparameters tuning phase the emerged optimized setup, later adopted for the final model training, is the following:

1. First Conv1D layer output channels: 64.
2. Dilated-Causal Conv1D output channels: 64.
3. Residual Units 1X1 Conv1D output channels: 64.
4. Last 1X1 Conv1D layers output channels: 192
5. Max dilation rate: 32.
6. Kernel width: 2.
7. Kernel strides: 1
8. Pooling method: Max Polling 1D.
9. Pooling size: 2.
10. Kernel initialization method: Orthogonal Initializer [[194]].
11. Kernel Initialization multiplicative factor: 1.0.
12. Kernel regularization method: L2 Regularization.
13. Regularization factor: 0.01.

14. Batch Normalization momentum: 0.2.

15. Batch Normalization epsilon: 2e-5.

16. Optimizer: Adam [[195]].

17. Learning rate: 0.001.

18. Batch size: 512.

The tuned model is characterized by 149 layers and about 1.6 millions of parameters (trainable and non-trainable).The final model training was carried out by splitting the original dataset into training, validation, and test sets with a 70/15/15 ratio. The model was trained on 1000 epochs, setting an early break after 150 epochs if no increase in the model performance evaluation metric was recorded. On each epoch end, the trained model was assessed on the validation set and the computed accuracy used as the aforementioned model performance evaluation metric. The model with the highest accuracy value on the validation set, along all the training epochs, was retained for inference on test and independent sets. Given the model operates as a binary classifier Binary Cross-Entropy was adopted as a loss function L. Like all DL classifiers the model training is mathematically formulated as a loss function minimization problem, which is solved by updating the neural network weights, via backpropagation algorithm, on the basis of the loss function gradient with respect to neural network last layer weights.

In detail the adopted last layer activation function σ and the loss function L for a binary classification, are described by the following equations [[196]]:

$$\sigma(Z_i) = \frac{1}{1+e^{-Z_i}} \tag{4.17}$$

$$Z_i = \sum_{j=1}^{m} w_j x_j + b_j \tag{4.18}$$

$$L(Z) = -\frac{1}{n}\sum_{i=1}^{n}[\, q(Z_i) log(\sigma(Z_i)) + (1 - q(Z_i)\, log(1 - \sigma(Z_i)) \tag{4.19}$$

where i denotes a generic input tensor fed to the last layer, m is the number of features in the input tensor, wj is the generic last layer weight, xj is the generic input feature, bj is the generic bias value associated to xj, Zi is the weighted sum of the generic input tensor i, n is the number of input tensors, q(Zi) is the is the observed class for input tensor i and σ(Zi) is the output of the last layer given the generic input i. The gradient of the loss function with respect to the last layer weights W, is calculated by means of the chain rule in the following way [[196]]:

$$\begin{aligned}
\frac{\partial L}{\partial W} &= \left[-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial L}{\partial \sigma}\frac{\partial \sigma}{\partial W_j} \right]_{1\le j\le m} \\
&= \left[-\frac{1}{n}\sum_{i=1}^{n} \left(\frac{q(Z_i)}{\sigma(Z_i)} - \frac{1-q(Z_i)}{1-\sigma(Z_i)} \right)\frac{\partial \sigma}{\partial W_j} \right]_{1\le j\le m} \\
&= \left[-\frac{1}{n}\sum_{i=1}^{n} \frac{\sigma(Z_i)-q(Z_i)}{\sigma(Z_i)(1-\sigma(Z_i))}\frac{\partial \sigma}{\partial W_j} \right]_{1\le j\le m}
\end{aligned} \tag{4.20}$$

Where ∂ is derivative operator.

The $\partial\sigma/\partial$Wj differentiation can be solved in the following way:

$$\frac{\partial\sigma}{\partial w_j} = \frac{\partial}{\partial w_j}\left(1 + e^{-\sum_{j=1}^{m} w_j x_j + b_j}\right)^{-1}$$

$$= -x_j e^{--\sum_{j=1}^{m} w_j x_j + b_j}\left(1 + e^{-\sum_{j=1}^{m} w_j x_j + b_j}\right)^{-2}$$

$$= -\frac{x_j e^{--\sum_{j=1}^{m} w_j x_j + b_j}}{\left(1+e^{-\sum_{j=1}^{m} w_j x_j + b_j}\right)\left(1+e^{-\sum_{j=1}^{m} w_j x_j + b_j}\right)}$$

$$= -x_j \frac{\left(1+e^{--\sum_{j=1}^{m} w_j x_j + b_j}\right)-1}{\left(1+e^{-\sum_{j=1}^{m} w_j x_j + b_j}\right)\left(1+e^{-\sum_{j=1}^{m} w_j x_j + b_j}\right)}$$

$$= x_j\left[\frac{1}{\left(1+e^{-\sum_{j=1}^{m} w_j x_j + b_j}\right)} - \left(\frac{1}{\left(1+e^{-\sum_{j=1}^{m} w_j x_j + b_j}\right)}\right)^2\right]$$

$$= x_j\left(\sigma(Z_i) - \sigma(Z_i)^2\right) = x_j\sigma(Z_i)(1 - \sigma(Z_i)) \tag{4.21}$$

so the gradient of the loss function with respect to the last layer weights $\partial L \partial / W$ (Eq. 4.20)is finally obtained by replacing $\partial\sigma/\partial$Wj (Eq. 4.21) in the following way [[196]]:

$$\frac{\partial L}{\partial W} = \left[-\frac{1}{n}\sum_{i=1}^{n} x_j \frac{\sigma(Z_i) - q(Z_i)}{\sigma(Z_i)(1-\sigma(Z_i))}\sigma(Z_i)(1 - \sigma(Z_i)) \right]_{1\leq j\leq m}$$

$$= \left[-\frac{1}{n}\sum_{i=1}^{n} x_j \sigma(Z_i) - q(Z_i) \right]_{1\leq j\leq m} \tag{4.22}$$

Given the training is performed over batches, the number of input tensors n corresponds to the batch size and the weights update is calculated via the Adam optimizer algorithm using the gradient calculated with the equation above.

4.9 Performance Evaluation and Aggregated P-Value

The overall performance evaluation was carried out using metrics well suited for the evaluation of models performance on unbalanced dataset, in detail precision, recall, specificity, G-Mean, F1-score, balanced accuracy. The computation of adopted metrics (Table 2.2) is described by the following set of equations:

$$\mathrm{Precision} = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$\begin{aligned} G - Mean &= \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}} \\ &= \sqrt{Recall \cdot Specificity} \end{aligned}$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN}$$

$$\begin{aligned} Balanced\ Accuracy &= \frac{1}{2}\Big(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\Big) \\ &= \frac{1}{2}(Recall + Specificity) \end{aligned}$$

$$\begin{aligned} TP &= True\ Positives \\ FP &= False\ Positives \\ TN &= True\ Negatives \\ FN &= False\ Negatives \end{aligned}$$

and the model results were summarized by means of confusion matrices (Figure 2.6, Figure 2.7). The presence of gaps in the 101 nucleotides window was simulated in the in house HEK293T KO-WT dataset, by means of random replacements of the RNAseq data with the corresponding genomic data from one up to one hundred replacements, and its effect assessed

using false negative rate and false positive rate as evaluation metrics, which are calculated with the following equations:

$$False\ Positive\ Rate = 1 - Specificity = \frac{FP}{TN + FP}$$
$$False\ Negative\ Rate = 1 - Recall = \frac{FN}{TP + FN}$$

The model output A-to-I RNA editing probability was later used to calculate a set aggregate P-values PA incorporated in the last REDIportal version [[174]], with respect to the A-to-I RNA editing positives sets of events retrieved from the GTEx samples, by means of the following equation:

$$P_A = \left[\frac{1}{N} \sum_{i=1}^{N} \sigma\left(X_{i,j,k}\right) \right]_{1 \le j \le M(k),\ 1 \le k \le 55} \tag{4.23}$$

where k indicate the generic human body site' GTEx collection of RNA-seq samples, M(k) is the total number of positive A-to-I RNA editing events extracted in the k body site, j is generic extracted A-to-I RNA editing event, N is the number of samples were event j was observed and extracted in the k body site, i indicates the generic sample, Xi, j, k is the matrix of feature for event j in sample i and body site k and σ(Xi, j, k) is the model output probability calculated for the matrix of feature Xi, j, k.

4.10 Performance Comparison With Existing Tools

The model performance was compared to the EditPredict, RDDpred, and RED-ML ones using the HEK293T KO-WT grand truth set of sites. EditPredict was launched using its online version (http://innovebioinfo.com/Sequencing_Analysis/RNAediting/RNA1.php) on the HEK293T KO-WT ground truth set of sites, previously sorted by Alu and non Alu sites by mean of annotation carried out with a python3-compliant version of the AnnotateTable.py script from the REDItools [12,19] package using the RepeatMasker annotations compatible with the specific genome assembly used for the alignment. The same procedure was carried out launching EditPredict on the U87 ground truth set of sites. RDDpred and RED-ML were downloaded from their GitHub repositories, https://github.com/vibbits/RDDpred and https://github.com/BGIRED/RED-ML, respectively. RDDpred was launched on the HEK293T KO and WT BAM files, obtained by means of alignment onto the GRCh38 reference genome with STAR. The same BAM files were used as input to RED-ML along with dbSNP141 and RepeatMasker annotations RED-ML output was further filtered using also dbSNP147 to take into account a wider set of known SNPs. EditPredict, RDDpred, and RED-ML predicted A-to-I RNA editing probabilities on the ground truth set of sites were collected from the corresponding output files and used, along with our model ones, to draw the corresponding ROC-AUC curves or compute False-Negative-Rate scores (Figure 2.10). In details ROC curves were calculated using the method elucidated by Fawcett [[197]] and the AUC score defined using the Somers' D according to the following equations [[198]]:

$$AUC = \frac{D_{XY}+1}{2} \qquad (4.24)$$

$$D_{XY} = \frac{N_C - N_D}{N_C + N_D + N_T}$$

where DXY is the Somers' D, Nc is the number of concordant pairs, i.e. couples of computed probabilities (pi, pj) and corresponding classes (yi, yj) where pi < pj (or pi > pj) and yi ¡ yj (or yi > yj), Nd is the number of discordant pairs, i.e. couples of computed probabilities (pi, pj) and corresponding classes (yi, yj) where pi < pj (or pi > pj) and yi > yj (or yi < yj), and NT is the number of non concordant and non discordant pairs, i.e. couples of computed probabilities (pi, pj) and corresponding classes (yi, yj) where p i= pj and yi = yj, was calculated according to the method elucidated by Fawcett [[197]]. The false negative rate used to assess the comparison with RDDpred was calculated with the equation previously reported.

4.11 Software and Hardware

All the analyses were carried out on the HPC cluster provided by the ReCaS Bari - Data Center (https://www.recas-bari.it/index.php/it/) with HTCondor Software Suite v9.013-3 for CentOS Linux v7 machines. The TCN training was carried out with a NVIDIA A100-PCIE-40GB GPU. GTEx RNA-seq samples, and corresponding WGS/WES samples, were processed in a dedicated python v2.7.15 Anaconda2 v5.2.0 virtual environment to REDItools v1-3 and pblat use, with the following python libraries and software from Bioconda channel (https://bioconda.github.io/conda-package_index.html) :

- bcftools v1.9 (required by the REDItools protocol).
- bedtools v2.28.0 (required by the REDItools protocol).
- bwa v0.7.17 (required by the REDItools protocol).
- bx-python v0.8.2 (required by the REDItools protocol).
- fastp v0.20.0 (required by the REDItools protocol).
- fastqc v0.11.8 (required by the REDItools protocol).
- fisher v0.1.4 (optional) (required by the REDItools protocol).
- gmap v2018.07.04 (required by the REDItools protocol).
- htslib v1.9 (required by the REDItools protocol).
- libdeflate v1.0 (required by the REDItools protocol).
- pysam v0.15.2 (critical for REDItools v1.3 functioning).
- rseqc v2.6.4 (required by the REDItools protocol).

- samtools v1.9 (required by the REDItools protocol).
- star v2.7.0f (required by the REDItools protocol).
- bzip2 v1.0.6 (required by the REDItools protocol).
- git v2.21.0 (required by the REDItools protocol).
- numpy v1.16.2 (required by the REDItools protocol).
- scipy v1.2.1 (required by the REDItools protocol).
- wget v1.20.1 (required by the REDItools protocol).
- REDItools v1.3.
- pblat (required by the REDItools protocol).

REDItools v1.3 is available at https://github.com/BioinfoUNIBA/REDItools and pblat is available at http://icebert.github.io/pblat/. Independent RNA-seq samples, and corresponding WGS/WES samples, were processed in a dedicated python v3.11.5 Anaconda3 v4.16.14 virtual environment for REDItools v3 use, with the following python libraries and software:

- pysam v0.22.0 (critical for REDItools v3 functioning).
- sortedcontainers v2.4.0 (critical for REDItools v3 functioning).
- REDItools v3.

Feature extraction, data preprocessing, TCN training, TCN performance evaluation and its comparison with existing tools were carried out in a dedicated python v3.9.0 Anaconda3 v4.16.14 virtual environment with the following python libraries:

- htcondor v23.0.2 (HTCondor python binding).
- jupyter_client v7.4.9 (used for testing the code).
- jupyter_core v4.12.0 (used for testing the code).
- jupyter_server v1.24.0 (used for testing the code).
- logomaker v0.8 (used for creating the sequence logo in Figure 2.3).
- keras v2.14.0 (used for the TCN implementation training and evaluation).
- matplotlib v3.5.3 (used for creating plots, confusion matrices and the sequence logo, Figure 2.2, Figure 2.3, Figure 2.4, Figure 2.6, Figure 2.7, Figure 2.8, Figure 2.9, Figure 2.10).
- numpy v1.21.6 (used for the data preprocessing).
- pandas v1.3.5 (used for feature extraction and general dataframes management).
- pysam v0.15.4 (used for fast genomic coordinates checking in the tabix indexed pileup tables for feature extraction).
- scikit-learn v1.0.2 (used for computation of ROC curves, AUC scores, confusion matrices, false positive rates and false negative rates, Figure 2.6, Figure 2.7, Figure 2.9, Figure 2.10).
- seaborn v0.12.2 (used for creating plots and confusion matrices, Figure 2.3, Figure 2.4, Figure 2.6, Figure 2.7, Figure 2.8, Figure 2.9, Figure 2.10).

- tqdm v0.0.1 (used to monitor the progress of processes launched in the various phases).

- tensorflow v2.14.0 (used for the TCN implementation training and evaluation).

python v3.9.0 releases of os, sys, builtin libraries were used to handle files in the various phases, python v3.9.0 release of random builtin library was used in the gaps related analysis and python v3.9.0 release of multiprocessing builtin library was used to parallelize the feature extraction along the 8404 pileup tables.

Bibliography

[1] Susumu Ohno. So much" junk" dna in our genome. in" evolution of genetic systems". In *Brookhaven symposium in biology*, volume 23, pages 366–370, 1972.

[2] Jianjun Chen, Miao Sun, Sanggyu Lee, Guolin Zhou, Janet D Rowley, and San Ming Wang. Identifying novel transcripts and novel genes in the human genome by using novel sage tags. *Proceedings of the National Academy of Sciences*, 99(19):12257–12262, 2002.

[3] Saurabh Saha, Andrew B Sparks, Carlo Rago, Viatcheslav Akmaev, Clarence J Wang, Bert Vogelstein, Kenneth W Kinzler, and Victor E Velculescu. Using the transcriptome to annotate the genome. *Nature biotechnology*, 20(5):508–512, 2002.

[4] John S Mattick. The central role of rna in human development and cognition. *FEBS letters*, 585(11):1600–1616, 2011.

[5] Andrew Fire, SiQun Xu, Mary K Montgomery, Steven A Kostas, Samuel E Driver, and Craig C Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *nature*, 391(6669):806–811, 1998.

[6] Rosalind C Lee, Rhonda L Feinbaum, and Victor Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *cell*, 75(5):843–854, 1993.

[7] Alain Jacquier. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small rnas. *Nature Reviews Genetics*, 10(12):833–844, 2009.

[8] Ryan J Taft, Ken C Pang, Timothy R Mercer, Marcel Dinger, and John S Mattick. Non-coding rnas: regulators of disease. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 220(2):126–139, 2010.

[9] Thomas Derrien, Roderic Guigó, and Rory Johnson. The long non-coding rnas: a new (p) layer in the "dark matter". *Frontiers in genetics*, 2:107, 2012.

[10] Benjamin J Blencowe. Alternative splicing: new insights from global analyses. *Cell*, 126(1):37–47, 2006.

[11] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.

[12] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–1415, 2008.

[13] Dione Kampa, Jill Cheng, Philipp Kapranov, Mark Yamanaka, Shane Brubaker, Simon Cawley, Jorg Drenkow, Antonio Piccolboni, Stefan Bekiranov, Gregg Helt, et al. Novel rnas identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome research*, 14(3):331–342, 2004.

[14] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[15] Timothy Ravasi, Harukazu Suzuki, Ken C Pang, Shintaro Katayama, Masaaki Furuno, Rie Okunishi, Shiro Fukuda, Kelin Ru, Martin C Frith, M Milena Gongora, et al. Experimental validation of the regulated expression of large numbers of non-coding rnas from the mouse genome. *Genome research*, 16(1):11–19, 2006.

[16] Moran N Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L Rinn. Integrative annotation of human large intergenic non-coding rnas reveals global properties and specific subclasses. *Genes & development*, 25(18):1915–1927, 2011.

[17] Paul Bertone, Viktor Stolc, Thomas E Royce, Joel S Rozowsky, Alexander E Urban, Xiaowei Zhu, John L Rinn, Waraporn Tongprasit, Manoj Samanta, Sherman Weissman, et al. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306(5705):2242–2246, 2004.

[18] Jill Cheng, Philipp Kapranov, Jorg Drenkow, Sujit Dike, Shane Brubaker, Sandeep Patel, Jeffrey Long, David Stern, Hari Tammana, Gregg Helt, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *science*, 308(5725):1149–1154, 2005.

[19] Philipp Kapranov, Jill Cheng, Sujit Dike, David A Nix, Radharani Duttagupta, Aarron T Willingham, Peter F Stadler, Jana Hertel, Jorg Hackermuller, Ivo L Hofacker, et al. Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science*, 316(5830):1484–1488, 2007.

[20] E Birney, JA Stamatoyannopoulos, A Dutta, R Guigó, TR Gingeras, EH Margulies, Z Weng, M Snyder, ET Dermitzakis, and RE Thurman. Baylor college of medicine human genome sequencing center; washington university genome sequencing center; broad institute; children's hospital oakland research institute.(2007). identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

[21] Philipp Kapranov, Simon E Cawley, Jorg Drenkow, Stefan Bekiranov, Robert L Strausberg, Stephen PA Fodor, and Thomas R Gingeras. Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, 296(5569):916–919, 2002.

[22] Pea Carninci, T Kasukawa, S Katayama, J Gough, MC Frith, Norihiro Maeda, Rieko Oyama, T Ravasi, B Lenhard, C Wells, et al. The transcriptional landscape of the mammalian genome. *science*, 309(5740):1559–1563, 2005.

[23] John L Rinn, Ghia Euskirchen, Paul Bertone, Rebecca Martone, Nicholas M Luscombe, Stephen Hartman, Paul M Harrison, F Kenneth Nelson, Perry Miller, Mark Gerstein, et al. The transcriptional activity of human chromosome 22. *Genes & development*, 17(4):529–540, 2003.

[24] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[25] Harm Van Bakel, Corey Nislow, Benjamin J Blencowe, and Timothy R Hughes. Most "dark matter" transcripts are associated with known genes. *PLoS biology*, 8(5):e1000371, 2010.

[26] Tim R Mercer, Daniel J Gerhardt, Marcel E Dinger, Joanna Crawford, Cole Trapnell, Jeffrey A Jeddeloh, John S Mattick, and John L Rinn. Targeted rna sequencing reveals the deep complexity of the human transcriptome. *Nature biotechnology*, 30(1):99–104, 2012.

[27] Barbara Uszczynska-Ratajczak, Julien Lagarde, Adam Frankish, Roderic Guigó, and Rory Johnson. Towards a complete map of the human long non-coding rna transcriptome. *Nature Reviews Genetics*, 19(9):535–548, 2018.

[28] Viktoriia A Arzumanian, Georgii V Dolgalev, Ilya Y Kurbatov, Olga I Kiseleva, and Ekaterina V Poverennaya. Epitranscriptome: review of top 25 most-studied rna modifications. *International Journal of Molecular Sciences*, 23(22):13851, 2022.

[29] Frank F Davis and Frank Worthington Allen. Ribonucleic acids from yeast which contain a fifth nucleotide. *J Biol Chem*, 227(2):907–15, 1957.

[30] Natalia Pinello, Renhua Song, Quintin Lee, Emilie Calonne, Mark Larance, François Fuks, and Justin J-L Wong. A multiomics dataset for the study of rna modifications in human macrophage differentiation and polarisation. *Scientific Data*, 11(1):252, 2024.

[31] Ian A Roundtree, Molly E Evans, Tao Pan, and Chuan He. Dynamic rna modifications in gene expression regulation. *Cell*, 169(7):1187–1200, 2017.

[32] Shuang Deng, Jialiang Zhang, Jiachun Su, Zhixiang Zuo, Lingxing Zeng, Kaijing Liu, Yanfen Zheng, Xudong Huang, Ruihong Bai, Lisha Zhuang, et al. Rna m6a regulates transcription via dna demethylation and chromatin accessibility. *Nature genetics*, 54(9):1427–1437, 2022.

[33] Jane E Jackman and Juan D Alfonzo. Transfer rna modifications: nature's combinatorial chemistry playground. *Wiley Interdisciplinary Reviews: RNA*, 4(1):35–48, 2013.

[34] Tao Wang, Xiaojun Li, Xiaojing Zhang, Qing Wang, Wenqian Liu, Xiaohong Lu, Shunli Gao, Zixi Liu, Mengshuang Liu, Lihong Gao, et al. Rna motifs and modification involve in rna long-distance transport in plants. *Frontiers in Cell and Developmental Biology*, 9:651278, 2021.

[35] Katherine I Zhou, Hailing Shi, Ruitu Lyu, Adam C Wylder, Żaneta Matuszek, Jessica N Pan, Chuan He, Marc Parisien, and Tao Pan. Regulation of co-transcriptional pre-mrna splicing by m6a through the low-complexity protein hnrnpg. *Molecular cell*, 76(1):70–81, 2019.

[36] James Anderson, Lon Phan, Rafael Cuesta, Bradley A Carlson, Marie Pak, Katsura Asano, Glenn R Björk, Mercedes Tamame, and Alan G Hinnebusch. The essential gcd10p–gcd14p nuclear complex is required for 1-methyladenosine modification and maturation of initiator methionyl-trna. *Genes & development*, 12(23):3650–3662, 1998.

[37] Jie Tong, Wuchao Zhang, Yuran Chen, Qiaoling Yuan, Ning-Ning Qin, and Guosheng Qu. The emerging role of rna modifications in the regulation of antiviral innate immunity. *Frontiers in Microbiology*, 13:845625, 2022.

[38] Nicky Jonkhout, Julia Tran, Martin A Smith, Nicole Schonrock, John S Mattick, and Eva Maria Novoa. The rna modification landscape in human disease. *Rna*, 23(12):1754–1769, 2017.

[39] Sylvain Delaunay and Michaela Frye. Rna modifications regulating cell fate in cancer. *Nature cell biology*, 21(5):552–559, 2019.

[40] Qiang Feng, Dongxu Wang, Tianyi Xue, Chao Lin, Yongjian Gao, Liqun Sun, Ye Jin, and Dianfeng Liu. The role of rna modification in hepatocellular carcinoma. *Frontiers in pharmacology*, 13:984453, 2022.

[41] Rosaura Esteve-Puig, Fina Climent, David Piñeyro, Eva Domingo-Domènech, Veronica Davalos, Maite Encuentra, Anna Rea, Nadia Espejo-Herrera, Marta Soler, Miguel Lopez, et al. Epigenetic loss of m1a rna demethylase alkbh3 in hodgkin lymphoma targets collagen, conferring poor clinical outcome. *Blood, The Journal of the American Society of Hematology*, 137(7):994–999, 2021.

[42] Kellie D Nance and Jordan L Meier. Modifications in an emergency: the role of n1-methylpseudouridine in covid-19 vaccines. *ACS central science*, 7(5):748–756, 2021.

[43] Xiulin Jiang, Baiyang Liu, Zhi Nie, Lincan Duan, Qiuxia Xiong, Zhixian Jin, Cuiping Yang, and Yongbin Chen. The role of m6a modification in the biological functions and diseases. *Signal transduction and targeted therapy*, 6(1):74, 2021.

[44] Wendan Ren, Jiuwei Lu, Mengjiang Huang, Linfeng Gao, Dongxu Li, Gang Greg Wang, and Jikui Song. Structure and regulation of zcchc4 in m6a-methylation of 28s rrna. *Nature communications*, 10(1):5042, 2019.

[45] Ying Yang, Phillip J Hsu, Yu-Sheng Chen, and Yun-Gui Yang. Dynamic transcriptomic m6a decoration: writers, erasers, readers and functions in rna metabolism. *Cell research*, 28(6):616–624, 2018.

[46] Hailing Shi, Jiangbo Wei, and Chuan He. Where, when, and how: context-dependent functions of rna methylation writers, readers, and erasers. *Molecular cell*, 74(4):640–650, 2019.

[47] Huaqing Yan, Liqi Zhang, Xiaobo Cui, Sinian Zheng, and Rubing Li. Roles and mechanisms of the m6a reader ythdc1 in biological processes and diseases. *Cell Death Discovery*, 8(1):237, 2022.

[48] Li-Juan Deng, Wei-Qing Deng, Shu-Ran Fan, Min-Feng Chen, Ming Qi, Wen-Yu Lyu, Qi Qi, Amit K Tiwari, Jia-Xu Chen, Dong-Mei Zhang, et al. m6a modification: recent advances, anticancer targeted drug discovery and beyond. *Molecular cancer*, 21(1):52, 2022.

[49] Marianna Penzo, Ania N Guerrieri, Federico Zacchini, Davide Treré, and Lorenzo Montanaro. Rna pseudouridylation in physiology and medicine: for better and for worse. *Genes*, 8(11):301, 2017.

[50] Anne C Rintala-Dempsey and Ute Kothe. Eukaryotic stand-alone pseudouridine synthases–rna modifying enzymes and emerging regulators of gene expression? *RNA biology*, 14(9):1185–1196, 2017.

[51] Darrell R Davis. Stabilization of rna stacking by pseudouridine. *Nucleic acids research*, 23(24):5020–5026, 1995.

[52] Rosaura Esteve-Puig, Alberto Bueno-Costa, and Manel Esteller. Writers, readers and erasers of rna modifications in cancer. *Cancer letters*, 474:127–137, 2020.

[53] Alexandre Garus and Chantal Autexier. Dyskerin: an essential pseudouridine synthase with multifaceted roles in ribosome biogenesis, splicing, and telomere maintenance. *Rna*, 27(12):1441–1458, 2021.

[54] Nicola Guzzi, Sowndarya Muthukumar, Maciej Cieśla, Gabriele Todisco, Phuong Cao Thi Ngoc, Magdalena Madej, Roberto Munita, Serena Fazio, Simon Ekström, Teresa Mortera-Blanco, et al. Pseudouridine-modified trna fragments repress aberrant protein synthesis and predict leukaemic progression in myelodysplastic syndrome. *Nature Cell Biology*, 24(3):299–306, 2022.

[55] Yelena Bykhovskaya, Kari Casas, Emebet Mengesha, Aida Inbal, and Nathan Fischel-Ghodsian. Missense mutation in pseudouridine synthase 1 (pus1) causes mitochondrial myopathy and sideroblastic anemia (mlasa). *The American Journal of Human Genetics*, 74(6):1303–1308, 2004.

[56] Chie Tomikawa. 7-methylguanosine modifications in transfer rna (trna). *International journal of molecular sciences*, 19(12):4080, 2018.

[57] Anand Ramanathan, G Brett Robb, and Siu-Hong Chan. mrna capping: biological functions and applications. *Nucleic acids research*, 44(16):7511–7526, 2016.

[58] Yuejun Luo, Yuxin Yao, Peng Wu, Xiaohui Zi, Nan Sun, and Jie He. The potential role of n 7-methylguanosine (m7g) in cancer. *Journal of hematology & oncology*, 15(1):63, 2022.

[59] Lionel Malbec, Ting Zhang, Yu-Sheng Chen, Ying Zhang, Bao-Fa Sun, Bo-Yang Shi, Yong-Liang Zhao, Ying Yang, and Yun-Gui Yang. Dynamic methylome of internal mrna n 7-methylguanosine and its regulatory role in translation. *Cell research*, 29(11):927–941, 2019.

[60] Mariam Jaafar, Hermes Paraqindes, Mathieu Gabut, Jean-Jacques Diaz, Virginie Marcel, and Sébastien Durand. 2 o-ribose methylation of ribosomal rnas: natural diversity in living organisms, biological processes, and diseases. *Cells*, 10(8):1948, 2021.

[61] Sunny Sharma, Virginie Marchand, Yuri Motorin, and Denis LJ Lafontaine. Identification of sites of 2-o-methylation vulnerability in human ribosomal rnas by systematic mapping. *Scientific reports*, 7(1):11490, 2017.

[62] Lilia Ayadi, Adeline Galvanin, Florian Pichot, Virginie Marchand, and Yuri Motorin. Rna ribose methylation (2-o-methylation): Occurrence, biosynthesis and biological functions. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1862(3):253–269, 2019.

[63] Qing Dai, Sharon Moshitch-Moshkovitz, Dali Han, Nitzan Kol, Ninette Amariglio, Gideon Rechavi, Dan Dominissini, and Chuan He. Nm-seq maps 2-o-methylation sites in human mrna with base precision. *Nature methods*, 14(7):695–698, 2017.

[64] Dilyana G Dimitrova, Laure Teysset, and Clément Carré. Rna 2-o-methylation (nm) modification in human diseases. *Genes*, 10(2):117, 2019.

[65] Virginie Marcel, Sandra E Ghayad, Stéphane Belin, Gabriel Therizols, Anne-Pierre Morel, Eduardo Solano-Gonzàlez, Julie A Vendrell, Sabine Hacot, Hichem C Mertani, Marie Alexandra Albaret, et al. p53 acts as a safeguard of translational control by regulating fibrillarin and rrna methylation in cancer. *Cancer cell*, 24(3):318–330, 2013.

[66] Jonatha M Gott and Ronald B Emeson. Functions and mechanisms of rna editing. *Annual review of genetics*, 34(1):499–531, 2000.

[67] Sundaramoorthy Srinivasan, Adrian Gabriel Torres, and Lluís Ribas de Pouplana. Inosine in biology and disease. *Genes*, 12(4):600, 2021.

[68] Brenda L Bass. Rna editing by adenosine deaminases that act on rna. *Annual review of biochemistry*, 71(1):817–846, 2002.

[69] Yiannis A Savva, Leila E Rieder, and Robert A Reenan. The adar protein family. *Genome biology*, 13:1–10, 2012.

[70] Yongfeng Jin, Wenjing Zhang, and Qi Li. Origins and evolution of adar-mediated rna editing. *IUBMB life*, 61(6):572–578, 2009.

[71] Kazuko Nishikura. A-to-i editing of coding and non-coding rnas by adars. *Nature reviews Molecular cell biology*, 17(2):83–96, 2016.

[72] Adrian Gabriel Torres, David Piñeyro, Liudmila Filonava, Travis H Stracker, Eduard Batlle, and Lluis Ribas de Pouplana. A-to-i editing on trnas: biochemical, biological and evolutionary implications. *FEBS letters*, 588(23):4279–4286, 2014.

[73] Erez Y Levanon, Eli Eisenberg, Rodrigo Yelin, Sergey Nemzer, Martina Hallegger, Ronen Shemesh, Zipora Y Fligelman, Avi Shoshan, Sarah R Pollock, Dan Sztybel, et al. Systematic identification of abundant a-to-i editing sites in the human transcriptome. *Nature biotechnology*, 22(8):1001–1005, 2004.

[74] Heather A Hundley and Brenda L Bass. Adar editing in double-stranded utrs and other noncoding rna sequences. *Trends in biochemical sciences*, 35(7):377–383, 2010.

[75] Amanda Wright and Bryce Vissel. The essential role of ampa receptor glur2 subunit rna editing in the normal and diseased brain. *Frontiers in molecular neuroscience*, 5:34, 2012.

[76] William Slotkin and Kazuko Nishikura. Adenosine-to-inosine rna editing and human disease. *Genome medicine*, 5:1–13, 2013.

[77] Eli Eisenberg and Erez Y Levanon. A-to-i rna editing—immune protector and transcriptome diversifier. *Nature Reviews Genetics*, 19(8):473–490, 2018.

[78] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[79] Alekos Athanasiadis, Alexander Rich, and Stefan Maas. Widespread a-to-i rna editing of alu-containing mrnas in the human transcriptome. *PLoS biology*, 2(12):e391, 2004.

[80] Matthew Blow, P Andrew Futreal, Richard Wooster, and Michael R Stratton. A survey of rna editing in human brain. *Genome research*, 14(12):2379–2387, 2004.

[81] Dennis DY Kim, Thomas TY Kim, Thomas Walsh, Yoshifumi Kobayashi, Tara C Matise, Steven Buyske, and Abram Gabriel. Widespread rna editing of embedded alu elements in the human transcriptome. *Genome research*, 14(9):1719–1725, 2004.

[82] Bernd Sommer, Martin Köhler, Rolf Sprengel, and Peter H Seeburg. Rna editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell*, 67(1):11–19, 1991.

[83] Barry Hoopengardner, Tarun Bhalla, Cynthia Staber, and Robert Reenan. Nervous system targets of rna editing identified by comparative genomics. *Science*, 301(5634):832–836, 2003.

[84] Colleen M Burns, Hsin Chu, Susan M Rueter, Linda K Hutchinson, Hervé Canton, Elaine Sanders-Bush, and Ronald B Emeson. Regulation of serotonin-2c receptor g-protein coupling by rna editing. *Nature*, 387(6630):303–308, 1997.

[85] Gokul Ramaswami and Jin Billy Li. Radar: a rigorously annotated database of a-to-i rna editing. *Nucleic acids research*, 42(D1):D109–D113, 2014.

[86] Ernesto Picardi, Anna Maria D'Erchia, Claudio Lo Giudice, and Graziano Pesole. Rediportal: a comprehensive database of a-to-i rna editing events in humans. *Nucleic acids research*, 45(D1):D750–D757, 2017.

[87] Meng How Tan, Qin Li, Raghuvaran Shanmugam, Robert Piskol, Jennefer Kohler, Amy N Young, Kaiwen Ivy Liu, Rui Zhang, Gokul Ramaswami, Kentaro Ariyoshi, et al. Dynamic landscape and regulation of rna editing in mammals. *Nature*, 550(7675):249–254, 2017.

[88] Miyoko Higuchi, Stefan Maas, Frank N Single, Jochen Hartner, Andrei Rozov, Nail Burnashev, Dirk Feldmeyer, Rolf Sprengel, and Peter H Seeburg. Point mutation in an ampa receptor gene rescues lethality in mice deficient in the rna-editing enzyme adar2. *Nature*, 406(6791):78–81, 2000.

[89] Marion Horsch, Peter H Seeburg, Thure Adler, Juan Antonio Aguilar-Pimentel, Lore Becker, Julia Calzada-Wack, Lilian Garrett, Alexander Götz, Wolfgang Hans, Miyoko Higuchi, et al. Requirement of the rna-editing enzyme adar2 for normal physiology in mice. *Journal of Biological Chemistry*, 286(21):18614–18622, 2011.

[90] Erez Y Levanon, Martina Hallegger, Yaron Kinar, Ronen Shemesh, Kristina Djinovic-Carugo, Gideon Rechavi, Michael F Jantsch, and Eli Eisenberg. Evolutionarily conserved human targets of adenosine to inosine rna editing. *Nucleic acids research*, 33(4):1162–1168, 2005.

[91] Maja Stulić and Michael F Jantsch. Spatio-temporal profiling of filamin a rna-editing reveals adar preferences and high editing levels outside neuronal tissues. *RNA biology*, 10(10):1611–1617, 2013.

[92] Jin Billy Li, Erez Y Levanon, Jung-Ki Yoon, John Aach, Bin Xie, Emily LeProust, Kun Zhang, Yuan Gao, and George M Church. Genome-wide identification of human rna editing sites by parallel dna capturing and sequencing. *Science*, 324(5931):1210–1213, 2009.

[93] Leilei Chen, Yan Li, Chi Ho Lin, Tim Hon Man Chan, Raymond Kwok Kei Chow, Yangyang Song, Ming Liu, Yun-Fei Yuan, Li Fu, Kar Lok Kong, et al. Recoding rna editing of azin1 predisposes to hepatocellular carcinoma. *Nature medicine*, 19(2):209–216, 2013.

[94] Jongchan Yeo, Rena A Goodman, Nicole T Schirle, Sheila S David, and Peter A Beal. Rna editing changes the lesion specificity for the dna repair enzyme neil1. *Proceedings of the National Academy of Sciences*, 107(48):20715–20719, 2010.

[95] Yishay Pinto, Haim Y Cohen, and Erez Y Levanon. Mammalian conserved adar targets comprise only a small fragment of the human editosome. *Genome biology*, 15:1–15, 2014.

[96] Takuto Hideyama, Takenari Yamashita, Takeshi Suzuki, Shoji Tsuji, Miyoko Higuchi, Peter H Seeburg, Ryosuke Takahashi, Hidemi Misawa, and Shin Kwak. Induced loss of adar2 engenders slow death of motor neurons from q/r site-unedited glur2. *Journal of Neuroscience*, 30(36):11917–11925, 2010.

[97] ShuHong Liu, Lorraine Lau, JianShe Wei, DongYa Zhu, Shengwei Zou, Hong-Suo Sun, YangPing Fu, Fang Liu, and YouMing Lu. Expression of ca2+-permeable ampa receptor channels primes cell death in transient forebrain ischemia. *Neuron*, 43(1):43–55, 2004.

[98] Yukio Kawahara, Shin Kwak, Hui Sun, Kyoko Ito, Hideji Hashida, Hitoshi Aizawa, Seon-Yong Jeong, and Ichiro Kanazawa. Human spinal motoneurons express low rel-

ative abundance of glur2 mrna: an implication for excitotoxicity in als. *Journal of neurochemistry*, 85(3):680–689, 2003.

[99] Shin Kwak and Yukio Kawahara. Deficient rna editing of glur2 and neuronal death in amyotropic lateral sclerosis. *Journal of molecular medicine*, 83:110–120, 2005.

[100] Peter L Peng, Xiafen Zhong, Weihong Tu, Mangala M Soundarapandian, Peter Molner, Dongya Zhu, Lorraine Lau, Shuhong Liu, Fang Liu, and YouMing Lu. Adar2-dependent rna editing of ampa receptor subunit glur2 determines vulnerability of neurons in forebrain ischemia. *Neuron*, 49(5):719–733, 2006.

[101] Caterina Cenci, Rita Barzotti, Federica Galeano, Sandro Corbelli, Rossella Rota, Luca Massimi, Concezio Di Rocco, Mary A O'Connell, and Angela Gallo. Down-regulation of rna editing in pediatric astrocytomas: Adar2 editing activity inhibits cell migration and proliferation. *Journal of Biological Chemistry*, 283(11):7251–7260, 2008.

[102] Yukio Kawahara, Adda Grimberg, Sarah Teegarden, Cedric Mombereau, Sui Liu, Tracy L Bale, Julie A Blendy, and Kazuko Nishikura. Dysregulated editing of serotonin 2c receptor mrnas results in energy dissipation and loss of fat mass. *Journal of Neuroscience*, 28(48):12834–12844, 2008.

[103] Stefan Maas, Yukio Kawahara, Kristen M Tamburro, and Kazuko Nishikura. A-to-i rna editing and human disease. *RNA biology*, 3(1):1–9, 2006.

[104] Richard T O'Neil and Ronald B Emeson. Quantitative analysis of 5ht2c receptor rna editing patterns in psychiatric disorders. *Neurobiology of disease*, 45(1):8–13, 2012.

[105] Galit Lev-Maor, Rotem Sorek, Erez Y Levanon, Nurit Paz, Eli Eisenberg, and Gil Ast. Rna-editing-mediated exon evolution. *Genome biology*, 8:1–12, 2007.

[106] Yishay Pinto, Ilana Buchumenski, Erez Y Levanon, and Eli Eisenberg. Human cancer tissues exhibit reduced a-to-i editing of mirnas coupled with elevated editing of their targets. *Nucleic Acids Research*, 46(1):71–82, 2018.

[107] Yukio Kawahara, Boris Zinshteyn, Praveen Sethupathy, Hisashi Iizasa, Artemis G Hatzigeorgiou, and Kazuko Nishikura. Redirection of silencing targets by adenosine-to-inosine editing of mirnas. *Science*, 315(5815):1137–1140, 2007.

[108] Andranik Ivanov, Sebastian Memczak, Emanuel Wyler, Francesca Torti, Hagit T Porath, Marta R Orejuela, Michael Piechotta, Erez Y Levanon, Markus Landthaler,

Christoph Dieterich, et al. Analysis of intron sequences reveals hallmarks of circular rna biogenesis in animals. *Cell reports*, 10(2):170–177, 2015.

[109] Brian J Liddicoat, Robert Piskol, Alistair M Chalk, Gokul Ramaswami, Miyoko Higuchi, Jochen C Hartner, Jin Billy Li, Peter H Seeburg, and Carl R Walkley. Rna editing by adar1 prevents mda5 sensing of endogenous dsrna as nonself. *Science*, 349(6252):1115–1120, 2015.

[110] Kathleen Pestal, Cory C Funk, Jessica M Snyder, Nathan D Price, Piper M Treuting, and Daniel B Stetson. Isoforms of rna-editing enzyme adar1 independently control nucleic acid sensor mda5-driven autoimmunity and multi-organ development. *Immunity*, 43(5):933–944, 2015.

[111] Niamh M Mannion, Sam M Greenwood, Robert Young, Sarah Cox, James Brindle, David Read, Christoffer Nellåker, Cornelia Vesely, Chris P Ponting, Paul J McLaughlin, et al. The rna-editing enzyme adar1 controls innate immune responses to rna. *Cell reports*, 9(4):1482–1494, 2014.

[112] Friedemann Weber, Valentina Wagner, Simon B Rasmussen, Rune Hartmann, and Søren R Paludan. Double-stranded rna is produced by positive-strand rna viruses and dna viruses but not in detectable amounts by negative-strand rna viruses. *Journal of virology*, 80(10):5059–5064, 2006.

[113] Qian Feng, Stanleyson V Hato, Martijn A Langereis, Jan Zoll, Richard Virgen-Slane, Alys Peisley, Sun Hur, Bert L Semler, Ronald P van Rij, and Frank JM van Kuppeveld. Mda5 detects the double-stranded rna replicative form in picornavirus-infected cells. *Cell reports*, 2(5):1187–1196, 2012.

[114] Sadeem Ahmad, Xin Mu, Fei Yang, Emily Greenwald, Ji Woo Park, Etai Jacob, Cheng-Zhong Zhang, and Sun Hur. Breaching self-tolerance to alu duplex rna underlies mda5-mediated inflammation. *Cell*, 172(4):797–810, 2018.

[115] Jacki E Heraud-Farlow, Alistair M Chalk, Sandra E Linder, Qin Li, Scott Taylor, Joshua M White, Lokman Pang, Brian J Liddicoat, Ankita Gupte, Jin Billy Li, et al. Protein recoding by adar1-mediated rna editing is not essential for normal development and homeostasis. *Genome biology*, 18:1–18, 2017.

[116] Gillian I Rice, Paul R Kasher, Gabriella MA Forte, Niamh M Mannion, Sam M Greenwood, Marcin Szynkiewicz, Jonathan E Dickerson, Sanjeev S Bhaskar, Massimiliano

Zampini, Tracy A Briggs, et al. Mutations in adar1 cause aicardi-goutieres syndrome associated with a type i interferon signature. *Nature genetics*, 44(11):1243–1248, 2012.

[117] Lily Bazak, Ami Haviv, Michal Barak, Jasmine Jacob-Hirsch, Patricia Deng, Rui Zhang, Farren J Isaacs, Gideon Rechavi, Jin Billy Li, Eli Eisenberg, et al. A-to-i rna editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome research*, 24(3):365–376, 2014.

[118] Shalom Hillel Roth, Erez Y Levanon, and Eli Eisenberg. Genome-wide quantification of adar adenosine-to-inosine rna editing activity. *Nature methods*, 16(11):1131–1138, 2019.

[119] Jochen C Hartner, Carl R Walkley, Jun Lu, and Stuart H Orkin. Adar1 is essential for the maintenance of hematopoiesis and suppression of interferon signaling. *Nature immunology*, 10(1):109–115, 2009.

[120] Charles E Samuel. Adenosine deaminases acting on rna (adars) are both antiviral and proviral. *Virology*, 411(2):180–193, 2011.

[121] Jacqueline Wettengel, Philipp Reautschnig, Sven Geisler, Philipp J Kahle, and Thorsten Stafforst. Harnessing human adar2 for rna repair–recoding a pink1 mutation rescues mitophagy. *Nucleic acids research*, 45(5):2797–2808, 2017.

[122] Maria Fernanda Montiel-Gonzalez, Isabel Vallecillo-Viejo, Guillermo A Yudowski, and Joshua JC Rosenthal. Correction of mutations within the cystic fibrosis transmembrane conductance regulator by site-directed rna editing. *Proceedings of the National Academy of Sciences*, 110(45):18285–18290, 2013.

[123] Isabel C Vallecillo-Viejo, Noa Liscovitch-Brauer, Maria Fernanda Montiel-Gonzalez, Eli Eisenberg, and Joshua JC Rosenthal. Abundant off-target edits from site-directed rna editing can be reduced by nuclear localization of the editing enzyme. *RNA biology*, 15(1):104–114, 2018.

[124] Claudio Lo Giudice, Marco Antonio Tangaro, Graziano Pesole, and Ernesto Picardi. Investigating rna editing in deep transcriptome datasets with reditools and rediportal. *Nature protocols*, 15(3):1098–1131, 2020.

[125] Maria Angela Diroma, Loredana Ciaccia, Graziano Pesole, and Ernesto Picardi. Elucidating the editome: bioinformatics approaches for rna editing detection. *Briefings in bioinformatics*, 20(2):436–447, 2019.

[126] Claudio Lo Giudice, Domenico Alessandro Silvestris, Shalom Hillel Roth, Eli Eisenberg, Graziano Pesole, Angela Gallo, and Ernesto Picardi. Quantifying rna editing in deep transcriptome datasets. *Frontiers in genetics*, 11:194, 2020.

[127] Alfonso Monaco, Ester Pantaleo, Nicola Amoroso, Antonio Lacalamita, Claudio Lo Giudice, Adriano Fonzino, Bruno Fosso, Ernesto Picardi, Sabina Tangaro, Graziano Pesole, et al. A primer on machine learning techniques for genomic applications. *Computational and Structural Biotechnology Journal*, 19:4345–4359, 2021.

[128] Yuning Cheng, Si-Mei Xu, Kristina Santucci, Grace Lindner, and Michael Janitz. Machine learning and related approaches in transcriptomics. *Biochemical and Biophysical Research Communications*, page 150225, 2024.

[129] Wei Chen, Pengmian Feng, Hui Ding, and Hao Lin. Pai: Predicting adenosine to inosine editing sites by using pseudo nucleotide compositions. *Scientific reports*, 6(1):35123, 2016.

[130] Wei Chen, Pengmian Feng, Hui Yang, Hui Ding, Hao Lin, and Kuo-Chen Chou. irna-ai: identifying the adenosine to inosine editing sites in rna sequences. *Oncotarget*, 8(3):4208, 2016.

[131] Wei Chen, Pengmian Feng, Hui Yang, Hui Ding, Hao Lin, and Kuo-Chen Chou. irna-3typea: identifying three types of modification at rna's adenosine sites. *Molecular Therapy-Nucleic Acids*, 11:468–474, 2018.

[132] Kewei Liu and Wei Chen. imrm: a platform for simultaneously identifying multiple kinds of rna modifications. *Bioinformatics*, 36(11):3336–3342, 2020.

[133] Ahsan Ahmad and Swakkhar Shatabda. Epai-nc: Enhanced prediction of adenosine to inosine rna editing sites using nucleotide compositions. *Analytical biochemistry*, 569:16–21, 2019.

[134] Jiandong Wang, Scott Ness, Roger Brown, Hui Yu, Olufunmilola Oyebamiji, Limin Jiang, Quanhu Sheng, David C Samuels, Ying-Yong Zhao, Jijun Tang, et al. Editpredict: prediction of rna editable sites with convolutional neural network. *Genomics*, 113(6):3864–3871, 2021.

[135] Ruyi Chen, Fuyi Li, Xudong Guo, Yue Bi, Chen Li, Shirui Pan, Lachlan JM Coin, and Jiangning Song. Attic is an integrated approach for predicting a-to-i rna editing sites in three species. *Briefings in Bioinformatics*, 24(3):bbad170, 2023.

[136] Xuan Xiao, Peng Wang, Zhaochun Xu, Wangren Qiu, and Xinzhu Fang. Pai-sae: predicting adenosine to inosine editing sites based on hybrid features by using spare auto-encoder. In *IOP Conference Series: Earth and Environmental Science*, volume 170, page 052018. IOP Publishing, 2018.

[137] Huseyin Avni Tac, Mustafa Koroglu, and Ugur Sezerman. Rddsvm: accurate prediction of a-to-i rna editing sites from sequence using support vector machines. *Functional & Integrative Genomics*, 21(5):633–643, 2021.

[138] Zhangyi Ouyang, Feng Liu, Chenghui Zhao, Chao Ren, Gaole An, Chuan Mei, Xiaochen Bo, and Wenjie Shu. Accurate identification of rna editing sites from primitive sequence with deep neural networks. *Scientific Reports*, 8(1):6005, 2018.

[139] Min-su Kim, Benjamin Hur, and Sun Kim. Rddpred: a condition-specific rna-editing prediction model from rna-seq data. *BMC genomics*, 17:85–95, 2016.

[140] Michał Zawisza-Álvarez, Jesús Peñuela-Melero, Esteban Vegas, Ferran Reverter, Jordi Garcia-Fernàndez, and Carlos Herrera-Úbeda. Exploring functional conservation in silico: a new machine learning approach to rna-editing. *Briefings in Bioinformatics*, 25(4):bbae332, 2024.

[141] Heng Xiong, Dongbing Liu, Qiye Li, Mengyue Lei, Liqin Xu, Liang Wu, Zongji Wang, Shancheng Ren, Wangsheng Li, Min Xia, et al. Red-ml: a novel, effective rna editing detection method based on machine learning. *Gigascience*, 6(5):gix012, 2017.

[142] Zhi-Can Fu, Bao-Qing Gao, Fang Nan, Xu-Kai Ma, and Li Yang. Demining: A deep learning model embedded framework to distinguish rna editing from dna mutations in rna sequencing data. *Genome Biology*, 25(1):258, 2024.

[143] Pariwat Tongnueasuk and Duangdao Wichadakul. Tae-ml: a random forest model for detecting rna editing sites. In *2020 17th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 59–64. IEEE, 2020.

[144] Aaron van den Oord. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[145] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[146] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[147] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR, 2017.

[148] Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*, 2016.

[149] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[150] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[151] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. In *Backpropagation*, pages 35–61. Psychology Press, 2013.

[152] Cicero Dos Santos and Bianca Zadrozny. Learning character-level representations for part-of-speech tagging. In *International Conference on Machine Learning*, pages 1818–1826. PMLR, 2014.

[153] Yoon Kim. Convolutional neural networks for sentence classification, 2014.

[154] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*, 2016.

[155] R Pascanu. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*, 2013.

[156] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *International conference on machine learning*, pages 2342–2350. PMLR, 2015.

[157] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.

[158] Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*, 2017.

[159] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[160] James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. Quasi-recurrent neural networks, 2016.

[161] Shiyu Chang, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, Xiaodong Cui, Michael Witbrock, Mark A Hasegawa-Johnson, and Thomas S Huang. Dilated recurrent neural networks. *Advances in neural information processing systems*, 30, 2017.

[162] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[163] Jining Yan, Lin Mu, Lizhe Wang, Rajiv Ranjan, and Albert Y Zomaya. Temporal convolutional networks for the advance prediction of enso. *Scientific reports*, 10(1):8055, 2020.

[164] George Udny Yule. On a method of investigating periodicities in disturbed series with special reference to wolfer's sunspot numbers. *Statistical Papers of George Udny Yule*, pages 389–420, 1971.

[165] F Yu. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[166] André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 4(11):e21, 2019.

[167] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[168] Luigi Mansi, Marco Antonio Tangaro, Claudio Lo Giudice, Tiziano Flati, Eli Kopel, Amos Avraham Schaffer, Tiziana Castrignanò, Giovanni Chillemi, Graziano Pesole, and Ernesto Picardi. Rediportal: millions of novel a-to-i rna editing events from thousands of rnaseq experiments. *Nucleic acids research*, 49(D1):D1012–D1019, 2021.

[169] Elin Lundin, Chenglin Wu, Albin Widmark, Mikaela Behm, Jens Hjerling-Leffler, Chammiran Daniel, Marie Öhman, and Mats Nilsson. Spatiotemporal mapping of rna editing in the developing mouse brain using in situ sequencing reveals regional and cell-type-specific regulation. *BMC biology*, 18:1–15, 2020.

[170] Erez Y Levanon, Roni Cohen-Fultheim, and Eli Eisenberg. In search of critical dsrna targets of adar1. *Trends in Genetics*, 40(3):250–259, 2024.

[171] Domenico Alessandro Silvestris, Ernesto Picardi, Valeriana Cesarini, Bruno Fosso, Nicolò Mangraviti, Luca Massimi, Maurizio Martini, Graziano Pesole, Franco Locatelli, and Angela Gallo. Dynamic inosinome profiles reveal novel patient stratification and gender-specific differences in glioblastoma. *Genome biology*, 20:1–18, 2019.

[172] Khen Khermesh, Anna Maria D'Erchia, Michal Barak, Anita Annese, Chaim Wachtel, Erez Y Levanon, Ernesto Picardi, and Eli Eisenberg. Reduced levels of protein recoding by a-to-i rna editing in alzheimer's disease. *Rna*, 22(2):290–302, 2016.

[173] Gokul Ramaswami, Rui Zhang, Robert Piskol, Liam P Keegan, Patricia Deng, Mary A O'connell, and Jin Billy Li. Identifying rna editing sites using rna sequencing data alone. *Nature methods*, 10(2):128–132, 2013.

[174] Pietro D'Addabbo, Roni Cohen-Fultheim, Itamar Twersky, Adriano Fonzino, Domenico Alessandro Silvestris, Ananth Prakash, Pietro Luca Mazzacuva, Juan Antonio Vizcaino, Andrew Green, Blake Sweeney, et al. Rediportal: toward an integrated view of the a-to-i editing. *Nucleic Acids Research*, page gkae1083, 2024.

[175] Ernesto Picardi and Graziano Pesole. Reditools: high-throughput rna editing detection made easy. *Bioinformatics*, 29(14):1813–1814, 2013.

[176] Ernesto Picardi, David S Horner, and Graziano Pesole. Single-cell transcriptomics reveals specific rna editing signatures in the human brain. *Rna*, 23(6):860–865, 2017.

[177] Ernesto Picardi, Caterina Manzari, Francesca Mastropasqua, Italia Aiello, Anna Maria D'Erchia, and Graziano Pesole. Profiling rna editing in human tissues: towards the inosinome atlas. *Scientific reports*, 5(1):14941, 2015.

[178] Ashani Kuttan and Brenda L Bass. Mechanistic insights into editing-site specificity of adars. *Proceedings of the National Academy of Sciences*, 109(48):E3295–E3304, 2012.

[179] Julie M Eggington, Tom Greene, and Brenda L Bass. Predicting sites of adar editing in double-stranded rna. *Nature communications*, 2(1):319, 2011.

[180] Tapani Raiko, Harri Valpola, and Yann LeCun. Deep learning made easier by linear transformations in perceptrons. In *Artificial intelligence and statistics*, pages 924–932. PMLR, 2012.

[181] Anmol Kiran and Pavel V Baranov. Darned: a database of rna editing in humans. *Bioinformatics*, 26(14):1772–1776, 2010.

[182] Tiziano Flati, Silvia Gioiosa, Nicola Spallanzani, Ilario Tagliaferri, Maria Angela Diroma, Graziano Pesole, Giovanni Chillemi, Ernesto Picardi, and Tiziana Castrignanò. Hpc-reditools: a novel hpc-aware tool for improved large scale rna-editing analysis. *BMC bioinformatics*, 21:1–12, 2020.

[183] Brian J Booth, Sami Nourreddine, Dhruva Katrekar, Yiannis Savva, Debojit Bose, Thomas J Long, David J Huss, and Prashant Mali. Rna editing: Expanding the potential of rna therapeutics. *Molecular Therapy*, 31(6):1533–1549, 2023.

[184] Julia-Sophia Bellingrath, Michelle E McClements, M Dominik Fischer, and Robert E MacLaren. Programmable rna editing with endogenous adar enzymes–a feasible option for the treatment of inherited retinal disease? *Frontiers in Molecular Neuroscience*, 16:1092913, 2023.

[185] Jinghui Song, Yuan Zhuang, and Chengqi Yi. Programmable rna base editing via targeted modifications. *Nature Chemical Biology*, 20(3):277–290, 2024.

[186] Riccardo Pecori, Isabel Chillón, Claudio Lo Giudice, Annette Arnold, Sandra Wüst, Marco Binder, Marco Marcia, Ernesto Picardi, and Fotini Nina Papavasiliou. Adar rna editing on antisense rnas results in apparent u-to-c base changes on overlapping sense transcripts. *Frontiers in Cell and Developmental Biology*, 10:1080626, 2023.

[187] Hiroshi Arakawa, Jessica Hauschild, and Jean-Marie Buerstedde. Requirement of the activation-induced deaminase (aid) gene for immunoglobulin gene conversion. *Science*, 295(5558):1301–1306, 2002.

[188] Adriano Fonzino, Pietro Luca Mazzacuva, Adam Handen, Domenico Alessandro Silvestris, Annette Arnold, Riccardo Pecori, Graziano Pesole, and Ernesto Picardi. Redinet: a tcn-based classifier for a-to-i rna editing detection harnessing million known events, 2024.

[189] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Process-

ing Subgroup. The sequence alignment/map format and samtools. *bioinformatics*, 25(16):2078–2079, 2009.

[190] James K Bonfield, John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, and Robert M Davies. Htslib: C library for reading/writing high-throughput sequencing data. *Gigascience*, 10(2):giab007, 2021.

[191] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

[192] Sergey Ioffe. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[193] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[194] Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. *arXiv preprint arXiv:2001.05992*, 2020.

[195] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[196] Michael Nielson. *Neural Network and Deep Learning*. Determination Press, 2015.

[197] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[198] Roger Newson. Parameters behind "nonparametric" statistics: Kendall's tau, somers' d and median differences. *The Stata Journal*, 2(1):45–64, 2002.

Appendices

Appendix A

Data Availability

All sequencing data generated in this study have been deposited in the SRA database (https://www.ncbi.nlm.nih.gov/sra) and are publicly accessible under accession number: PRJNA1138417. Illumina RNAseq and WGS data of A549 cells were downloaded from SRA under accession numbers: SRX8983709, SRX8983710, SRX8983711, SRX8983718, SRX8983719, SRX8983720 and SRX5437560, respectively. Illumina WGS data of HEK293T cells were downloaded from SRA under accession number SRX6858029. Illumina RNAseq and WGS data of U87 instead were downloaded from SRA under the following accessions: SRX110671, SRX110672, SRX110673, SRX110674, and SRX5466662, respectively.

Appendix B

Code Availability

The TCN-based classifier proposed in this study is publicly available (under the MIT license) at the GitHub repository:

https://github.com/BioinfoUNIBA/REDInet

.

REDItools v1 is available at the GitHub repository:

https://github.com/BioinfoUNIBA/REDItools

.

REDItools v3 is available at the GitHub repository:

https://github.com/BioinfoUNIBA/REDItools3

.

All the python scripts and .ipynb files used in the TCN-based classifier development, performance assessment, and comparison with other available tools are also publicly available at the GitHub repository:

https://github.com/BioinfoUNIBA/REDInet

.