OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf107
https://doi.org/10.1093/bib/bbaf107
Case Study

REDInet: a temporal convolutional network-based classifier for A-to-I RNA editing detection harnessing million known events

Adriano Fonzino[1,‡], Pietro Luca Mazzacuva[2,3,‡], Adam Handen[4], Domenico Alessandro Silvestris[1], Annette Arnold[5], Riccardo Pecori[5,6], Graziano Pesole [1,2], Ernesto Picardi [1,2,*]

[1]Department of Biosciences, Biotechnology and Environment, University of Bari Aldo Moro, Via Orabona 4, 70125, Bari, Italy
[2]Institute of Biomembranes, Bioenergetics and Molecular Biotechnology, National Research Council, Via Amendola 122/O, 70126, Bari, Italy
[3]Department of Engineering, University Campus Bio-Medico of Rome, Via Álvaro del Portillo 21, 00128, Rome, Italy
[4]Biological Sciences Division, University of Chicago, 5841 S Maryland Avenue, 60637, Chicago, USA
[5]Division of Immune Diversity, German Cancer Research Center, Im Neuenheimer Feld 28069120, Heidelberg, Germany
[6]Helmholtz Institute for Translational Oncology (HI-TRON), Obere Zahlbacherstr., 55131, Mainz, Germany

[*]Corresponding author. Department of Biosciences, Biotechnology and Environment, University of Bari Aldo Moro, Via Orabona 4, 70125, Bari, Italy.
E-mail: ernesto.picardi@uniba.it
[‡]Adriano Fonzino and Pietro Luca Mazzacuva contributed equally to this work.

Abstract

A-to-I ribonucleic acid (RNA) editing detection is still a challenging task. Current bioinformatics tools rely on empirical filters and whole genome sequencing or whole exome sequencing data to remove background noise, sequencing errors, and artifacts. Sometimes they make use of cumbersome and time-consuming computational procedures. Here, we present REDInet, a temporal convolutional network-based deep learning algorithm, to profile RNA editing in human RNA sequencing (RNAseq) data. It has been trained on REDIportal RNA editing sites, the largest collection of human A-to-I changes from >8000 RNAseq data of the genotype-tissue expression project. REDInet can classify editing events with high accuracy harnessing RNAseq nucleotide frequencies of 101-base windows without the need for coupled genomic data.

Keywords: temporal convolutional network; A-to-I RNA editing; RNAseq; REDItools

Introduction

Adenosine (A) to inosine (I) RNA editing is pervasive in human transcriptomes [1, 2]. It is carried out by members of the ADAR (adenosine deaminase acting on RNA) family of enzymes acting on double-stranded RNAs [3]. Depending on its location in the target RNA, i.e. coding deoxyribonucleic acid (DNA) Sequences (CDS), intron, or untranslated regions (UTRs), A-to-I editing plays relevant biological roles and initiates a plethora of downstream effects [4]. It can alter the physiology of neuroreceptors, tune alternative splicing, influence the biogenesis of circular RNAs, and modulate gene expression through edited miRNAs or edited miRNA binding sites of target RNAs [4, 5]. ADAR-mediated changes in repetitive regions, especially those in opposite orientations formed by Alu elements [6], are involved in cytosolic innate immunity by suppressing type I interferon signaling triggered through the MDA5-MAVS axis [7–9]. A-to-I events in protein-coding regions, though limited in number, have been extensively investigated because their deregulation has been associated with several human disorders [10, 11]. Advancements in high-throughput sequencing technologies have fostered the deciphering of the human inosinome, but accurate detection of RNA editing remains challenging [12, 13]. Empirical methods developed so far require high-quality RNAseq data coupled with whole genome sequencing (WGS) or whole exome sequencing (WES) data to mitigate genomic sources of variation as well as sequencing errors and artifacts [12–14]. Multiple filtering steps using prior knowledge and public genome annotations are usually needed, as well as cumbersome computational procedures [12–15]. Recent approaches based on machine- and deep-learning algorithms hold great promise in overcoming empirical methods' limitations and do not require time-consuming and computationally intensive filtering steps [16]. However, their generalization is strongly dependent on the selected input features as well as the size and quality of the training set [17]. Harnessing millions of known A-to-I events stored in our REDIportal database [1], based on >8000 genotype-tissue expression (GTEx) [18] RNAseq experiments through a rigorous computational protocol, we have trained REDInet, a temporal convolutional network (TCN) algorithm in which sequencing data are processed and treated like audio data. REDInet does not require WGS/WES data and employs RNAseq nucleotide frequencies with 101-base windows, centered at the investigated putative editing position, calculated through a brand-new and computationally efficient REDItools [19] implementation. In particular, REDItools is a suite of Python scripts to detect RNA editing changes in deep transcriptome sequencing datasets harnessing a variety of empirical filters [19]. REDItools traverse the entire reference genome position by

Received: November 29, 2024. **Revised:** February 19, 2025. **Accepted:** February 24, 2025

position and provide a list of variant and invariant sites supported by RNAseq reads, facilitating the extraction of 101-base windows centered at the putative target site.

Results and discussion

Unlike current machine (ML) and deep learning (DL) tools for detecting A-to-I changes [20–22], our approach uses RNAseq frequencies of genomic regions flanking putative events to capture the dynamic nature of RNA editing, including its tissue and cell type specificity [23, 24]. Harnessing 8404 pre-processed human RNAseq data from 55 GTEx body sites, already used in combination with WGS/WES data to populate our REDIportal database [1], we collected about 13 million unique examples for training, validation, and testing purposes (Table S1). Positive sites, comprising bona fide A-to-I changes, displayed a unimodal distribution of editing levels with values <0.2 for the majority of events (Fig. 1a), as expected for bulk human tissues. Indeed, most RNA editing sites reside in Alu elements and exhibit low editing frequencies (sometimes <1%) [1, 24]. Negative sites, including mainly Single Nucleotide Polymorphisms (SNPs), were instead selected using coupled RNAseq and WGS/WES data, and their A-to-G substitution rates were, as expected, bimodally distributed with picks at 0.5 (if both alleles were expressed) and 1.0 (if only the allele carrying the G was expressed; Fig. 1a) [25]. 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 depicted in Fig. 1b, Gs were depleted one base upstream and enriched one base downstream of the positive sites, as previously reported for the ADAR sequence motif [26, 27]. In addition, bases surrounding positive sites were enriched in A-to-I editing events that are known to occur in clusters (Fig. S2) [4].

Through pre-computed REDItools tables, we extracted gap-free intervals of 50 nucleotides flanking the putative or not putative editing sites supported by RNAseq reads. Each interval, encoded by a matrix of 8×101 features (Fig. 1c), was used to feed a TCN model. This model (Fig. 1d), employs a sequence of Residual Units [28, 29], each one composed of a series of Dilated 1D Causal Convolution layers with residual and skip connections. While 70% of collected examples were used for training, half of the remaining sites were used in validation to select the optimal model (Table S1). Our trained algorithm, implemented in the REDInet tool, was initially applied on training, validation, and test sets, achieving accuracy values >99% without under- or over-fitting (Fig. 2a; Fig. S3a and b, and Table S1).

To better assess the performance of REDInet, it was tested on three ground truth RNA editing datasets from A549 and HEK293T cell lines, not included in the GTEx atlas and never seen by the model (Fig. 1e and Table S1). The A549 dataset, downloaded from SRA, comprised stranded paired-end reads from three wild-type (WT) and three ADAR-silenced (SI) samples. A brand-new version of REDItools was used to quickly pre-process genome-aligned reads and extract only sites with A-to-I editing evidence in WT samples and no evidence in SI samples, leading to a balanced dataset of >13 000 sites (Table S1). REDInet was able to correctly classify them with a very low number of false predictions, obtaining a mean accuracy of >96% (Fig. 2b and Table S2). The HEK293T dataset, instead, encompassed in-house high throughput stranded RNAseq reads from three WT, three ADAR knockout (KO), and three ADAR overexpressing samples (OE; 437.4 million reads on average per sample), enabling the creation of two grand truth sets of RNA editing sites (Table S1). The first set (KO-WT) was obtained by comparing KO and WT samples, yielding >20,000 bona fide A-to-I events. The second set (KO-OE) was created by contrasting KO and OE samples, returning >35,000 RNA editing sites (Table S1). REDInet achieved an accuracy of >95% on both KO-WT (Fig. 2c) and KO-OE datasets, with a specificity of >93% and precision of >96% (Fig. S3c and Table S2). Similar results (accuracy >95%, precision >93%, and specificity >92%) were also obtained by applying REDInet to a further ground truth set of sites from public RNAseq reads of HEK293T, including WT and ADAR KO samples (Fig. S3d and Table S1) sequenced at a lower throughput (69.5 million reads on average per sample) (Table S2).

Since the vast majority of A-to-I editing occurs in repetitive regions and is harnessed to quantify the activity of endogenous ADARs [30], REDInet was applied to about 4000 adenosines falling in Alu elements supported by RNAseq reads from KO, WT, and OE samples of the HEK293T cell line. The number of positives (predicted as putative RNA editing events) increased from KO (5%) to OE (95%), proving the REDInet capability to catch RNA editing signals embedded in RNAseq data (Fig. 2d). A small fraction of editing sites in KO samples was expected because at least one of the ADAR enzymes was fully functional. In addition, using predicted A-to-I sites we calculated an Alu editing index (AEI) like score [17] showing that RNA editing activity increased from KO to OE samples, as expected (Fig. 2e).

REDInet was also compared to available tools implementing ML or DL algorithms for detecting A-to-I editing in the bona fide HEK293T KO-WT dataset (Fig. 2f). Although each tested tool is based on different paradigms and algorithms, and requires specific preprocessing steps of input data and distinct software dependencies, to conduct a comparison as fair as possible, the analyses were carried out on the same set of bona fide sites from HEK293T KO and WT samples, according the guidelines of each specific program. Such bona fide sites were not included in the training dataset of our REDInet model. The highest area under the curve (AUC) score was achieved by REDInet (0.98), followed by EditPredict (0.78) [20] based on a convolutional neural network architecture trained on sequence contexts (101-bases long centered at the putative editing site) flanking REDIportal v1 sites [31] and tested on public HEK293T datasets. RED-ML [21], based on a logistic regression classifier trained on three broad classes of features from experimentally validated events, achieved only an AUC score of 0.72. RDDpred, specialized in distinguishing sequencing artifacts from genuine AG substitutions [22] and implementing a Random Forest RDD classifier (based on 15 features reflecting the read-alignment patterns) trained on DARNED [32] and RADAR [33] sites, obtained a False Negative Rate of ~7%. In contrast, REDInet achieved 2% only. Unfortunately, the assessment of DeepRed [34], a further tool implementing a DL algorithm similar to EditPredict, was hampered by installation issues and software dependencies. Nonetheless, using public U87 RNAseq datasets already employed by DeepRed in combination with an available U87 WGS run, a bona fine set of >2800 sites was generated. On this U87 dataset, REDInet obtained an AUC score of 0.99 (accuracy >97%, specificity >96%, and precision >97%) while EditPredict 0.51 (Table S2, Fig. S3e and f).

Since REDInet predicts A-to-I events using 101-base windows centered at the investigated putative editing site, its predictive strength was tested in the HEK293 KO-WT dataset simulating missing data in flanking regions by randomly replacing the RNAseq data with the corresponding genomic data. REDInet was still effective at detecting RNA editing events when >30%

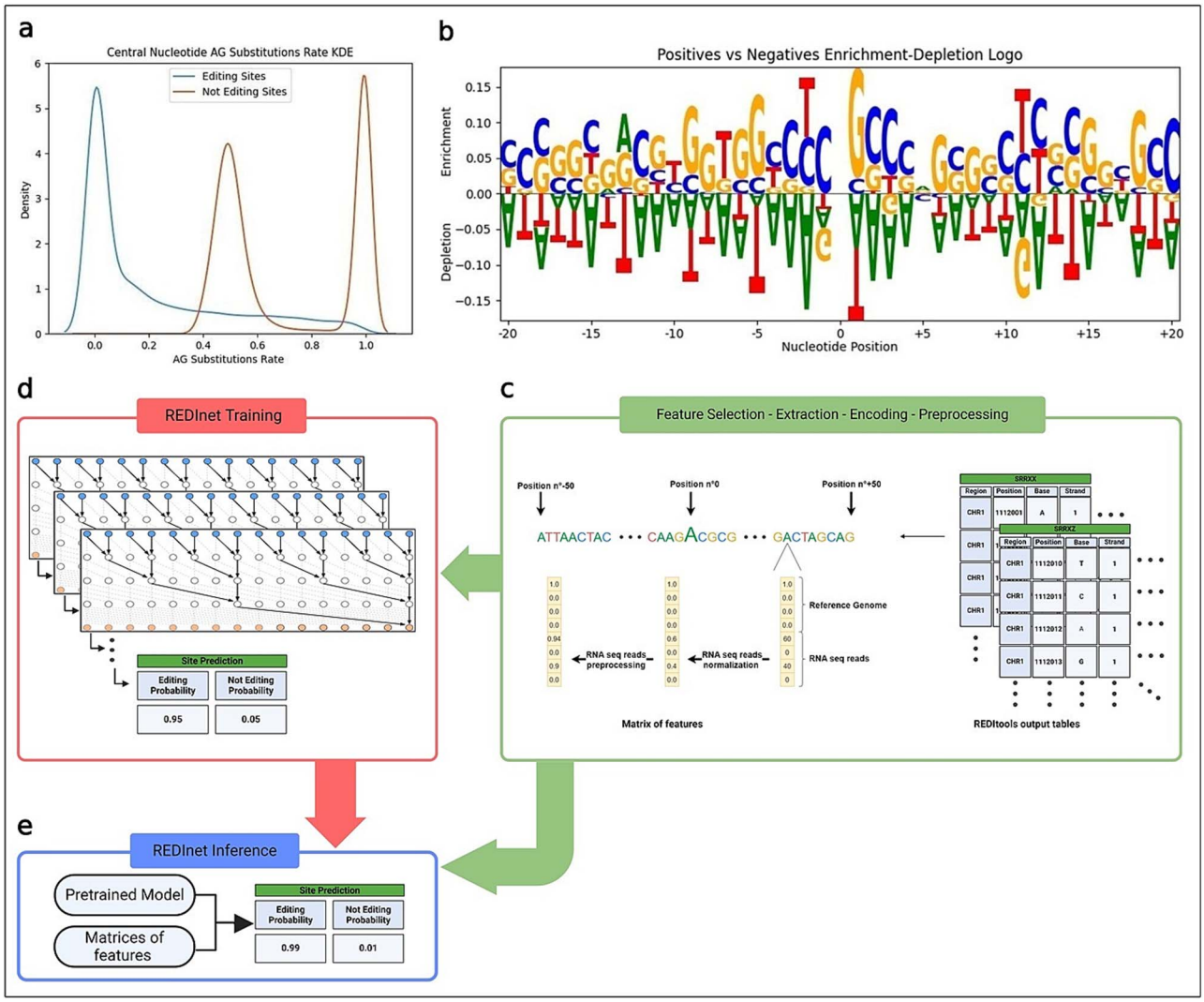

Figure 1. Training dataset and REDInet architecture. (a) Density plot of A-to-G substitution levels for positive and negative examples. Editing candidates are unimodally distributed while negatives display two picks at 0.5 and 1. (b) Two sample logo of collected examples with positives up and negatives down (for graphical constraints only 20 nucleotides up and down the investigated sites are reported here. An extended version is shown in Fig. S5). (c) REDItools tables are parsed and individual candidate sites are extracted as well as flanking regions. For each position an array of 8 features is extracted, including one-hot-encoded genomic bases and RNAseq counts per base. (d) REDInet leverages a TCN architecture and was trained on millions of positive or negative (SNPs) editing examples. e. Schematics of REDInet pre-trained model used on RNAseq data in inference mode. REDInet works as a binary classifier, providing two probability scores per site, corresponding to the editing and non-editing classes.

of flanking regions were missing, with false positive and false negative rates <10% (Fig. S4).

It is worth mentioning that the sequence length surrounding the putative editing site was not arbitrarily chosen. Indeed, several previous and independent studies demonstrated that a 101-base window centered on the investigated site was optimal for capturing sequence properties of edited and unedited regions [20–22]. As previously reported, 101-base windows are optimal in balancing missing values (if too long) and good predictions (if too short) [20–22].

Human A-to-I editing events are usually annotated by RepeatMasker into three groups: (i) ALU for sites in Alu elements, (ii) REP for sites in repetitive non-Alu elements, and (iii) NONREP for sites in non-repetitive regions [1, 12, 19, 31]. REDInet does not take into account these categories to avoid a largely unbalanced training set. Indeed, NONREP sites represent a tiny fraction of known events, while >95% of editing sites reside in Alu elements. Nonetheless, the REDInet architecture is quite flexible and ready to handle multiple categories of editing sites, if needed, or easy to extend for incorporating additional features.

To profile A-to-I editing transcriptome-wide, REDInet requires RNAseq alignment data tables, i.e. pileup data. Getting such a table is computationally intensive and time-consuming [35], hampering the use of REDInet on large datasets or experiments with high sequencing depth. To avoid potential bottlenecks, RNAseq pileup data were generated through a re-implementation of the REDItools algorithm. Its third generation, combined here with REDInet, allowed the traversing of a BAM file of aligned RNAseq reads in minutes rather than hours. For example, in HEK293T KO samples, comprising on average 277 million aligned reads, REDItools v3 took <60 minutes to provide a complete pileup profile of variant and invariant sites without any specific active filter. In contrast, REDItools v1 was 10x slower, completing the same jobs in ~9 hours.

REDInet can be applied to human RNAseq data only. It is not a limitation because high accuracy values by DL algorithms

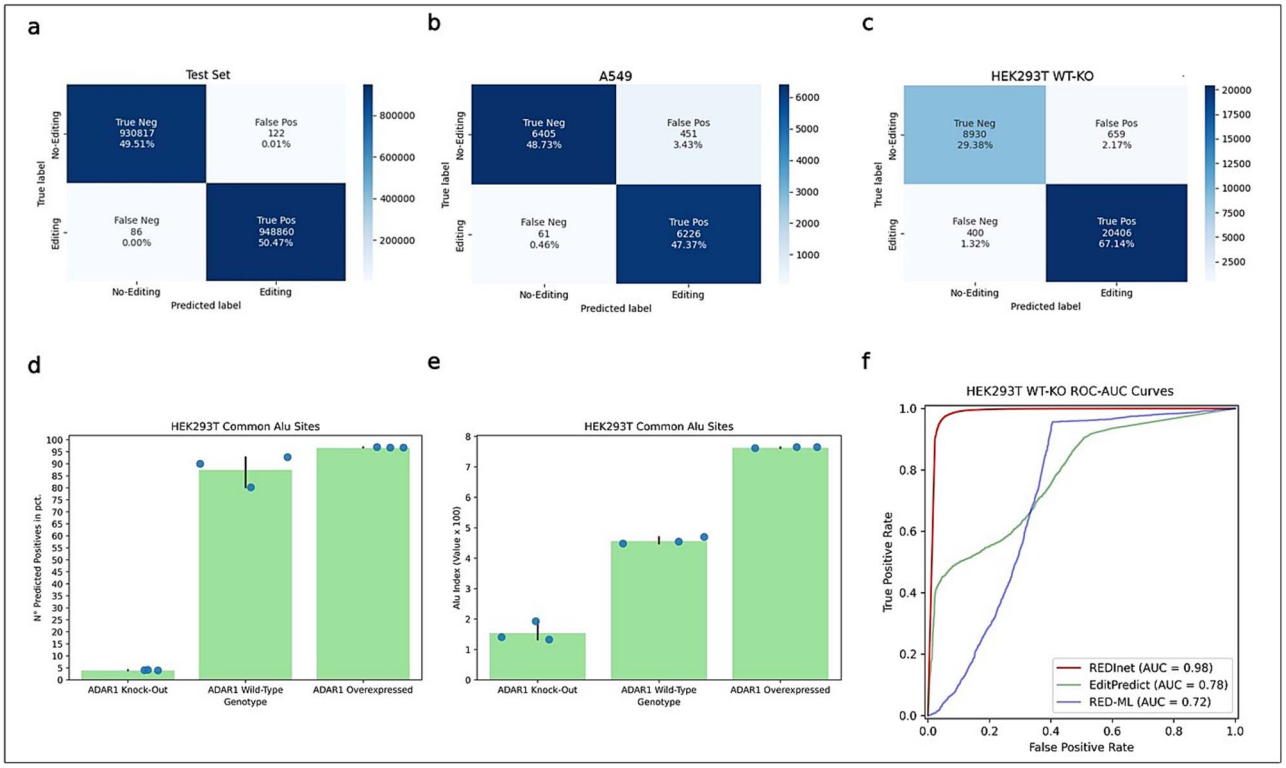

Figure 2. REDInet performances. (a–c) Confusion matrices on the REDIportal-based test set, on bona fide ground truth for A549 and HEK293T (WT-KO) independent datasets, respectively. (d) Positives predicted in Alu sites of HEK293T cells. (e) AEI-like score on positives predicted in Alus of HEK293T cells. (f) ROC-AUC curves of available tools on the bonafide ground truth WT-KO dataset from HEK293T.

can be obtained using only huge amounts of training examples. Once large datasets from non-human organisms are collected in REDIportal, they will be employed to feed new REDInet models and improve their generalization. The availability of RNA editing sites in diseased human samples (such as cancer data from TCGA) will lead to updated REDInet models to better investigate the A-to-I editing dynamics in human disorders.

Predicting tools like REDInet may play a key role in understanding the properties of sequence editability by ADARs and help in the design of antisense oligonucleotides for emerging site-directed RNA editing applications for correcting disease-causing point mutations [4, 36–38], whose applicability is largely site-, sequence context- and cell type-dependent.

Conclusions

REDInet is a novel and effective RNA editing classifier for human RNAseq data based on a Deep Neural Network model harnessing the largest available collection of A-to-I editing events from the REDIportal database. In combination with a brand-new REDItools implementation, REDInet can profile A-to-I editing with high accuracy in aligned RNAseq data without any prior knowledge of genomic variants in minutes.

REDInet is written in Python. Its code and usage details are freely available at https://github.com/BioinfoUNIBA/REDInet.

Methods

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% fetal bovine serum (PAN Biotech, Cat# P40–37100) and 1% penicillin/streptomycin (Sigma-Aldrich, Cat# P4333) in 5% CO2 at 37 °C. ADAR1 KO cell line was generated as previously described [39]. To generate the ADAR1 overexpressing cell line, we first RT-PCR amplified (Qiagen, Cat# 210212) the coding sequence of ADAR1 p110 (primer F attcaagcttgggccaccggctagcATGGCCGAGATCAAGGAG, primer R: gggggggagggagaggggcagatctCTATACTGGGCAGAGATAAAAG) and clone it inside the AID-express-puro2 plasmid [40] 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. This plasmid was then transfected into HEK293T cells using Lipofectamine 2000 (ThermoFisher, Cat# 11668019) according to the manufacturer's instructions. 24 h after transfection, cells were selected with puromycin (1.5 μg/ml). Upon selection, cells were sorted for the highest level of GFP (and thus ADAR1 p110) in bulk. Finally, ADAR1 p110 overexpression was validated by western blot (Cell Signaling Technology, Cat# 14175) using GAPDH as a loading control (Cell Signaling Technology, Cat# 2118). Transfections were performed as three independent biological replicates.

Total RNA was extracted and purified using a Qiagen RNeasy Plus kit (Qiagen, Cat# 74134) and further treated with DNase (Invitrogen, Cat# AM1907). The Illumina libraries were prepared from 500 ng of total RNA, using Illumina's TruSeq Stranded Total RNA Sample Preparation Kit (Illumina, San Diego, CA, USA), according to the manufacturer's protocol. The cDNA libraries were then 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). Sequencing of 3

WT, 3 KO, and 3 OE samples was performed with a NovaSeq 6000 platform using the paired-end approach (2 × 100 bp) yielding 262–638 million reads per sample.

Building of positive and negative datasets

The training dataset was built using 8404 human RNAseq samples and corresponding WGS/WES data (if any) from the GTEx [18] project. Transcriptomic and genomic reads were processed according to our REDItools protocol to detect RNA editing sites, already used to populate the REDIportal database [1] and well described elsewhere [12]. For each REDItools output table, we selected as positives all sites with A-to-G evidence in RNAseq, no base change in DNAseq (WGS or WES data), and present in the latest REDIportal [1] database version. Negative sites included all positions with A-to-G evidence in both RNAseq and DNAseq data to capture RNA variants not due to RNA editing. Only sites supported by 50 RNAseq reads and 10 DNAseq reads were considered. To rule out low-frequency SNPs in the training dataset, we retained as negatives all sites with an RNAseq A-to-G rate $\geq$ 0.4 and a comparable DNAseq A-to-G rate diverging from the RNAseq A-to-G rate <5%.

For each positive and negative site, windows of 50 bases upstream and 50 bases downstream of the site position, for a total of 101 bases, were extracted from the reference genome. All sites with the 101-base window fully covered by RNAseq reads were maintained in the final dataset of positives and negatives.

Data pre-processing and REDInet architecture

Each nucleotide of the 101-base window, centered at the putative and not putative editing site, was encoded via a feature vector of eight variables, the first four corresponding to reference genome sequences encoded by a one-hot numeric array (i.e. a binary vector of four variables representing the four canonical bases in which the values of 1 or 0 are assigned if a specific base is present or absent in the reference genome), the second four representing alignment profiles of the region of interest, including the counts of supporting RNAseq reads. RNAseq base counts were normalized by calculating their frequency on each reference position within the 101-base window and were then further preprocessed using log2 function. To avoid 0 values, an arbitrary $\alpha = 0.0001$ value was added to computed frequencies, simulating a single base over 10 000 aligned reads. Ultimately, each 101-base window was encoded by an 8x101 matrix of features with rescaled RNAseq counts (as detailed in the Supplementary Material).

The TCN architecture implemented in REDInet was inspired by the WaveNet architecture [41] and based on a sequence of Residual Units [28, 29], each one composed of a series of Dilated 1D Causal Convolutions layers with dilation rates that vary according to powers of 2, followed by the Gated Activation Unit [42]. Each Residual Unit carried a Residual Connection, a Residual Unit, and a skip connection. Several Batch Normalization [43] layers were used to stabilize REDInet and reach convergence faster. The complete description of REDInet architecture is provided in the Supplementary Methods and shown schematically in Fig. S1.

Training and evaluation of REDInet performances

The original training dataset was split into training, validation, and test sets with a 70/15/15 ratio. REDInet training was set on 1000 epochs with an early break after 150 epochs if no increase in the model performance was recorded. After each epoch, the trained model was assessed on the validation set. The model with the highest accuracy value was retained for inference on test and independent sets.

For performance evaluation metrics, Recall, Specificity, Precision, F1-Score, G-Mean, and Balanced Accuracy were calculated according to formulas detailed in the Supplementary Methods. For each dataset, results were summarized using Confusion Matrices. The performances of existing tools were evaluated by ROC-AUC curves. Custom Python scripts were used to create Confusion Matrices and ROC-AUC curves.

Ground truth ribonucleic acid editing datasets for validation

Stranded and paired-end Illumina RNAseq reads from three WT and three ADAR silenced (SI) A549 cell lines were downloaded from the SRA database (accessions: SRX8983709, SRX8983710, SRX8983711, SRX8983718, SRX8983719, SRX8983720) as well as the A549 WGS dataset (accession SRX5437560). Transcriptome reads were quality-checked by FASTQC and aligned onto the reference human genome (assembly hg38) by STAR (version 2.7.0f using GENCODE v.46 during the genome indexing). WGS reads, instead, were aligned onto the reference human genome (assembly hg38) by BWA (version 0.7.17). RNA variants per sample were called by REDItools v3 with no stringent parameters. Custom Python scripts were used to select positives and negatives. All sites showing A-to-G evidence in WT data but not in SI and WGS data were marked as positives. On the contrary, sites with A-to-G evidence in WT, SI, and WGS data or no A-to-G evidence in all samples, were marked as negatives. A minimal depth of 50 for RNAseq and 10 for WGS was used. Each A-to-G change was considered if supported by at least 3 Gs or showed a substitution rate $\geq$1%. Positive and negative candidates were finally annotated using a python3-compliant version of the AnnotateTable.py script from the REDItools [12, 19] package. Sites in repetitive regions were selected using RepeatMask annotations. Sites in non-repetitive regions were further filtered taking only those in RefSeq annotations.

The above-described procedure was also applied to extract positive and negative candidates from RNAseq data of WT, KO, and OE HEK293T cell lines generated in this study. WGS data from HEK293T were downloaded from SRA (accession SRX5437560). Two ground truth sets were created. The KO-WT dataset in which positives and negatives were selected contrasting pairs of KO and WT samples, and the KO-OE dataset obtained comparing pairs of KO and OE samples. All ground truth datasets are available in our GitHub repository.

Less stringent filtering steps were adopted for public HEK293T and U87 data, downloaded from SRA (HEK293T accessions: SRX2825941, SRX2825942, SRX2825943, SRX2825944, SRX2825945, SRX2825949; U87 accessions: SRX110671, SRX110672, SRX110672, SRX110674), in which the minimal depth for RNAseq was fixed to 30. WGS for U87 was downloaded from SRA under the accession SRX5466662.

Performance comparison of REDInet with existing tools

EditPredict, RDDpred, and RED-ML were compared to REDInet 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) and providing KO-WT sites annotated in Alu and non-Alu categories using a python3-compliant version of the AnnotateTable.py script from the REDItools [12, 19] package and RepeatMasker annotations downloaded from UCSC. 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-WT bam file aligned onto the GRCh38 reference genome. The same bam file was provided in input to RED-ML along with dbSNP141 (further filtered using also dbSNP147 to take into account a wider set of known SNPs) and RepeatMasker annotations. 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 REDInet ones, to draw the ROC-AUC curves or compute False-Negative-Rate scores (fig. 2F and Fig. S3F and G). In the case of RDDpred, since this tool works on groups of samples to make predictions for individual positions, only bona fide positive sites were chosen to calculate the False Negative Rate and enable the comparison with REDInet (Fig. S3G).

Speeding up the binary alignment map traversing by REDItools3

To speed up the calling of RNA variants for REDInet, a new version of REDItools was used. The original algorithm was maintained, but the implementation improved. The primary sources of speedup in the latest version are new C libraries made available through PySAM (https://github.com/pysam-developers/pysam) [44, 45], updating to Python3, and a dynamic implementation of REDItools' features. REDItools has numerous options and features for customizing each editing analysis. With the dynamic implementation, only the necessary pieces of code will be loaded and implemented based on the user's input.

Additional speedups were achieved by improvements in coding logic and loop structures. REDItools is now also streamlined for parallel processing.

Key Points

- Adenosine to inosine (A-to-I) ribonucleic acid (RNA) editing detection in RNA sequencing (RNAseq) data is still challenging due to background noise, sequencing errors, and artifacts.
- We developed REDInet, a novel tool based on a deep learning (DL) algorithm trained on millions of A-to-I RNA editing events from the REDIportal database.
- REDInet allows the A-to-I RNA editing profiling in RNAseq data without the need for whole genome sequencing or whole exome sequencing experiments and employs RNAseq nucleotide frequencies with 101-base windows, centered at the investigated putative editing position.
- REDInet works in combination with a brand-new and computationally efficient REDItools implementation allowing the processing of a BAM file in minutes.
- REDInet performance has been assessed using real datasets and compared to state-of-the-art software implementing machine and DL methods.

Acknowledgements

Authors are grateful to the following 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). The 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.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

None declared.

Data availability

All sequencing data generated in this study have been deposited in the Sequence Read Archive (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. REDInet is publicly available (under the MIT license) at the GitHub repository https://github.com/BioinfoUNIBA/REDInet and https://doi.org/10.5281/zenodo.14236576, while REDItools v3 is available at https://github.com/BioinfoUNIBA/REDItools3 and https://doi.org/10.5281/zenodo.13981271.

References

1. Mansi L, Tangaro MA, Lo Giudice C. *et al.* REDIportal: Millions of novel A-to-I RNA editing events from thousands of RNAseq experiments. *Nucleic Acids Res* 2021;**49**:D1012–9. https://doi.org/10.1093/nar/gkaa916
2. GTEx Consortium, Tan MH, Li Q. *et al.* Dynamic landscape and regulation of RNA editing in mammals. *Nature* 2017;**550**:249–54. https://doi.org/10.1038/nature24041
3. Savva YA, Rieder LE, Reenan RA. The ADAR protein family. *Genome Biol* 2012;**13**:252. https://doi.org/10.1186/gb-2012-13-12-252
4. Eisenberg E, Levanon EY. A-to-I RNA editing—Immune protector and transcriptome diversifier. *Nat Rev Genet* 2018;**19**:473–90. https://doi.org/10.1038/s41576-018-0006-1
5. Lundin E, Wu C, Widmark A. *et al.* Spatiotemporal mapping of RNA editing in the developing mouse brain using in situ sequencing reveals regional and cell-type-specific regulation. *BMC Biol* 2020;**18**:6. https://doi.org/10.1186/s12915-019-0736-3

6. Levanon EY, Cohen-Fultheim R, Eisenberg E. In search of critical dsRNA targets of ADAR1. *Trends Genet* 2024;**40**:250–9. https://doi.org/10.1016/j.tig.2023.12.002
7. Silvestris DA, Picardi E, Cesarini V. *et al.* Dynamic inosinome profiles reveal novel patient stratification and gender-specific differences in glioblastoma. *Genome Biol* 2019;**20**:33. https://doi.org/10.1186/s13059-019-1647-x
8. Pestal K, Funk CC, Snyder JM. *et al.* Isoforms of RNA-editing enzyme ADAR1 independently control nucleic acid sensor MDA5-driven autoimmunity and multi-organ development. *Immunity* 2015;**43**:933–44. https://doi.org/10.1016/j.immuni.2015.11.001
9. Liddicoat BJ, Piskol R, Chalk AM. *et al.* RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science* 2015;**349**:1115–20. https://doi.org/10.1126/science.aac7049
10. Khermesh K, D'Erchia AM, Barak M. *et al.* Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer's disease. *RNA* 2016;**22**:290–302. https://doi.org/10.1261/rna.054627.115
11. Slotkin W, Nishikura K. Adenosine-to-inosine RNA editing and human disease. *Genome Med* 2013;**5**:105. https://doi.org/10.1186/gm508
12. Lo Giudice C, Tangaro MA, Pesole G. *et al.* Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. *Nat Protoc* 2020;**15**:1098–131. https://doi.org/10.1038/s41596-019-0279-7
13. Diroma MA, Ciaccia L, Pesole G. *et al.* Elucidating the editome: Bioinformatics approaches for RNA editing detection. *Brief Bioinform* 2019;**20**:436–47. https://doi.org/10.1093/bib/bbx129
14. Lo Giudice C, Silvestris DA, Roth SH. *et al.* Quantifying RNA editing in deep transcriptome datasets. *Front Genet* 2020;**11**:194. https://doi.org/10.3389/fgene.2020.00194
15. Ramaswami G, Zhang R, Piskol R. *et al.* Identifying RNA editing sites using RNA sequencing data alone. *Nat Methods* 2013;**10**:128–32. https://doi.org/10.1038/nmeth.2330
16. Monaco A, Pantaleo E, Amoroso N. *et al.* A primer on machine learning techniques for genomic applications. *Comput Struct Biotechnol J* 2021;**19**:4345–59. https://doi.org/10.1016/j.csbj.2021.07.021
17. Cheng Y, Xu S-M, Santucci K. *et al.* Machine learning and related approaches in transcriptomics. *Biochem Biophys Res Commun* 2024;**724**:150225. https://doi.org/10.1016/j.bbrc.2024.150225
18. Lonsdale J, Thomas J, Salvatore M. *et al.* The genotype-tissue expression (GTEx) project. *Nat Genet* 2013;**45**:580–5. https://doi.org/10.1038/ng.2653
19. Picardi E, Pesole G. REDItools: High-throughput RNA editing detection made easy. *Bioinformatics* 2013;**29**:1813–4. https://doi.org/10.1093/bioinformatics/btt287
20. Wang J, Ness S, Brown R. *et al.* EditPredict: Prediction of RNA editable sites with convolutional neural network. *Genomics* 2021;**113**:3864–71. https://doi.org/10.1016/j.ygeno.2021.09.016
21. Xiong H, Liu D, Li Q. *et al.* RED-ML: A novel, effective RNA editing detection method based on machine learning. *GigaScience* 2017;**6**:1–8. https://doi.org/10.1093/gigascience/gix012
22. Kim M, Hur B, Kim S. RDDpred: A condition-specific RNA-editing prediction model from RNA-seq data. *BMC Genomics* 2016;**17**:5. https://doi.org/10.1186/s12864-015-2301-y
23. Picardi E, Horner DS, Pesole G. Single-cell transcriptomics reveals specific RNA editing signatures in the human brain. *RNA* 2017;**23**:860–5. https://doi.org/10.1261/rna.058271.116
24. Picardi E, Manzari C, Mastropasqua F. *et al.* Profiling RNA editing in human tissues: Towards the inosinome atlas. *Sci Rep* 2015;**5**:14941. https://doi.org/10.1038/srep14941
25. Endo TA. Quality control method for RNA-seq using single nucleotide polymorphism allele frequency. *Genes Cells Devoted Mol Cell Mech* 2014;**19**:821–9. https://doi.org/10.1111/gtc.12178
26. Kuttan A, Bass BL. Mechanistic insights into editing-site specificity of ADARs. *Proc Natl Acad Sci* 2012;**109**:E3295–304. https://doi.org/10.1073/pnas.1212548109
27. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. *Nat Commun* 2011;**2**:319. https://doi.org/10.1038/ncomms1324
28. He K, Zhang X, Ren S. *et al.* Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 27-30 June 2016*, 770–78. https://doi.org/10.1109/CVPR.2016.90
29. Raiko T, Valpola H, LeCun Y. Deep learning made easier by linear transformations in perceptrons. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* 2012;**22**:924–32.
30. Roth SH, Levanon EY, Eisenberg E. Genome-wide quantification of ADAR adenosine-to-inosine RNA editing activity. *Nat Methods* 2019;**16**:1131–8. https://doi.org/10.1038/s41592-019-0610-9
31. Picardi E, D'Erchia AM, Lo Giudice C. *et al.* REDIportal: A comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res* 2017;**45**:D750–7. https://doi.org/10.1093/nar/gkw767
32. Kiran A, Baranov PV. DARNED: A DAtabase of RNa EDiting in humans. *Bioinformatics* 2010;**26**:1772–6. https://doi.org/10.1093/bioinformatics/btq285
33. Ramaswami G, Li JB. RADAR: A rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res* 2014;**42**:D109–13. https://doi.org/10.1093/nar/gkt996
34. Ouyang Z, Liu F, Zhao C. *et al.* Accurate identification of RNA editing sites from primitive sequence with deep neural networks. *Sci Rep* 2018;**8**:6005. https://doi.org/10.1038/s41598-018-24298-y
35. Flati T, Gioiosa S, Spallanzani N. *et al.* HPC-REDItools: A novel HPC-aware tool for improved large scale RNA-editing analysis. *BMC Bioinformatics* 2020;**21**:353. https://doi.org/10.1186/s12859-020-03562-x
36. Booth BJ, Nourreddine S, Katrekar D. *et al.* RNA editing: Expanding the potential of RNA therapeutics. *Mol Ther* 2023;**31**:1533–49. https://doi.org/10.1016/j.ymthe.2023.01.005
37. Bellingrath J-S, McClements ME, Fischer MD. *et al.* Programmable RNA editing with endogenous ADAR enzymes—A feasible option for the treatment of inherited retinal disease? *Front Mol Neurosci* 2023;**16**:1092913. https://doi.org/10.3389/fnmol.2023.1092913
38. Song J, Zhuang Y, Yi C. Programmable RNA base editing via targeted modifications. *Nat Chem Biol* 2024;**20**:277–90. https://doi.org/10.1038/s41589-023-01531-y
39. Pecori R, Chillón I, Lo Giudice C. *et al.* ADAR RNA editing on antisense RNAs results in apparent U-to-C base changes on overlapping sense transcripts. *Front Cell Dev Biol* 2023;**10**:1080626. https://doi.org/10.3389/fcell.2022.1080626
40. Arakawa H, Hauschild J, Buerstedde J-M. Requirement of the activation-induced deaminase (*AID*) gene for immunoglobulin gene conversion. *Science* 2002;**295**:1301–6. https://doi.org/10.1126/science.1067308
41. van den Oord A, Dieleman S, Zen H. *et al.* WaveNet: A generative model for raw audio. *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), 125 arXiv* 2016.
42. Van Den, Kalchbrenner N, Vinyals O. *et al.* Conditional image generation with PixelCNN decoders. *arXiv* 2016.
43. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *PMLR* 2015;**37**:448–56.

44. Li H, Handsaker B, Wysoker A. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9. https://doi.org/10.1093/bioinformatics/btp352

45. Bonfield JK, Marshall J, Danecek P. *et al.* HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience* 2021;**10**:giab007. https://doi.org/10.1093/gigascience/giab007

Briefings in Bioinformatics, 2025, **26(2)**, bbaf107
https://doi.org/10.1093/bib/bbaf107
Case Study